International Conference on Applied Informatics for Health and Life Sciences

in conjunction with

Turkish-German Workshop on Bioinformatics:
Recent Developments from Health to Nanotechnology

Kuşadası-TÜRKİYE
19-22 October 2014
International Conference on Applied Informatics for Health and Life Sciences
AIHLS-2014

in conjunction with

The Turkish-German Workshop on Bioinformatics: Recent Developments from Health to Nanotechnology

19-22 October 2014
Kuşadası, Türkiye

Edited by
Rainer Merkl  Ali Kılınç  Cenk Selçuki  Hüseyin Şeker
The conference proceeding is made publicly available for easy access at the conference website. If you find it useful, the editors and paper authors will be grateful if you could cite both the conference proceeding and relevant paper(s).

Rainer Merkl, Ali Kılınç, Cenk Selçuki, Hüseyin Şeker
Editors on behalf of the organisation committee and authors
This page is intentionally left blank.
The main aim of the conference is to bring together researchers and practitioners from academia and industry into an interdisciplinary environment to present their works on the latest developments and applications of informatics methods and tools in health and life sciences. Potential participants are encouraged to focus on the applications of informatics methods and tools in the following topics (but not limited to):

- Genomics
- Proteomics
- Metabolomics
- Systems Biology
- Next-generation genome sequencing
- Modelling and simulation of biological phenomena
- Collecting and managing the big biological data
- Biological and clinical decision support systems
- Electronic Health Records
- Biological and medical imaging
- Pharmacy and Pharmaceutical Technologies
- Personalized medicine
- Immunology and Toxicology
- Dentistry
- Nursery and Midwifery
- Public Health
- Social care health
- Biomedical and Environmental Health
- Assisted Living
- Aging population
- Disability
- Cognitive and Behavioural Psychology
- Computer and console games and alternative technologies for health and life sciences
- Mobile and smart phone applications in health and life sciences
- Education and training in health and life sciences

The conference is organised in conjunction with the Turkish-German Workshop on Bioinformatics and provides an ideal forum for the exchange of information among specialists in their fields. It is held in a stimulating and open scientific atmosphere to foster the development of new collaborations.

Honorary Chairs
Muzaffer Şeker (Konya, Turkey)
Azmi Telefoncu (Izmir, Turkey)

Conference General Chair
Hüseyin Şeker (Leicester, UK)

Conference Co-Chairs
Ali Kılıç (Izmir, Turkey)
Cenk Selçuki (Izmir, Turkey)

Technical Committee
Volkan Uslan (Konya, Turkey)
Taylan K. Öztürk (Izmir, Turkey)
Funda Kartal (Izmir, Turkey)

Scientific Programme Committee
Ahmed Bouridane (Newcastle, UK)
Ahmet Orun (Leicester, UK)
Azmi Telefoncu (Izmir, TR)
David Suendermann-Oeft (San Francisco, USA)
Dimitrios Fotiadis (Athens, Greece)
Esin Ozturk Isik (Istanbul, TR)
Figen Zihnioğlu (Izmir,TR)
Gokmen Altay (Istanbul, TR)
Graham Ball (Nottingham, UK)
Hanaa M. Hussain (Kuwait, Kuwait)
Hedef Dhair El-Yassin (Baghdad, Iraq)
Masahiro Sugimoto (Yamagata, Japan)
Metin Akay (Houston, USA)
Nizamettin Aydin (Istanbul, TR)
Olcay Kursun (Istanbul, TR)
Omer Sinan Sarac (Istanbul, TR)
Parvez Haris (Leicester, UK)
Tevfik Dorak (Liverpool, UK)
Turgay Isbir (Istanbul, TR)
Ugur Sezerman (Istanbul, TR)
Yucel Kocyigit (Manisa, TR)

Conference Website
http://www.aihls.ege.edu.tr
http://www.bioinformatics.ege.edu.tr

Contact e-mail
aihls.congress@gmail.com
This page is intentionally left blank.
Welcome Message

On behalf of the AIHLS2014 Organising Committee, it is our great pleasure to welcome you to the 2nd International Conference on Applied Informatics for Health and Life Sciences.

AIHLS-2014, held in conjunction with the Turkish-German Workshop on Bioinformatics at PineBay Resort Hotel Conference Centre (Kuşadası, Türkiye) on 19-22 October 2014, aims at bringing together researchers and practitioners from academia and industry into an interdisciplinary environment to focus on the development and applications of informatics methods and tools in health and life sciences. This year’s theme focuses on “Bioinformatics and Systems Biology” in line with the latest technological developments, particularly, in the post-genomic era.

This exciting event provides an ideal forum for the exchange of information among specialists in their fields. It is held in a stimulating and open scientific atmosphere to foster the development of new collaborations. The invited speakers share their expertise and views, with the participants, in order to address new developments and challenges. The submitted papers have also addressed highly significant and timely topics in the area of informatics for health and life sciences. The participants come from a wide range of backgrounds including medicine, biology, engineering and computer science, which truly suggests that the conference is highly interdisciplinary and interactive event.

The conference has been organised into fifteen scientific sessions each covering different aspects of informatics for health and life sciences. This is further supported with a pre-conference tutorial covering timely topics for post-graduate students and post-doctoral researchers.

We would like to thank you all very much for your participation and contribution of high quality research papers. The success of AIHLS-2014 relies on your efforts. We would also like to thank our sponsors, INTERLAB LTD and INTRON Sağlık Ürünleri LTD for their kind support for the organisation of this timely event.

We hope you will find AIHLS2014 a memorable and fruitful event, and you enjoy both the conference and Kuşadası being one of the most beautiful holiday destinations in the world.

Rainer Merkl, Ali Kılınç, Cenk Selçuki, Hüseyin Şeker,

Kuşadası, Aydın, Türkiye, October 2014
Programme
### Sunday, 19 October 2014

#### Session SUN_S1: Pre-Conference Tutorials
**Session Chairs: Tevfik DORAK**

<table>
<thead>
<tr>
<th>Time</th>
<th>Tutorial</th>
<th>Speaker</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:30-15:20</td>
<td>Tutorial-I</td>
<td>Tevfik DORAK, Liverpool Hope University, UK&lt;br&gt;Alfred WINTER, Universität Leipzig, FRG&lt;br&gt;Michael SCHROEDER, Technische Universität Dresden, FRG</td>
<td>“How to do research and make a presentation, and read &amp; write a paper in the area of Informatics for Health and Life Sciences”</td>
</tr>
<tr>
<td>15:20-15:50</td>
<td>Tutorial-II</td>
<td>Ilknur Melis DURASI, Sabanci University, TR</td>
<td>“Introduction to DAVID and PANOGA Bioinformatics Resources”</td>
</tr>
<tr>
<td>15:50-16:30</td>
<td>Tutorial-III</td>
<td>Wolfram GRONWALD, University of Regensburg, FRG</td>
<td>“Introduction to Metabolomics”</td>
</tr>
<tr>
<td>16:30-17:00</td>
<td>Coffee Break</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Session SUN_S2: Welcome and Official Opening
**Session Chairs: Cenk SELÇUKİ**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00-17:00</td>
<td>Registration</td>
<td>Registration Desk</td>
</tr>
<tr>
<td>17:00-17:15</td>
<td>Welcome Speech</td>
<td>Workshop/Conference Chair/Co-chairs</td>
</tr>
<tr>
<td>17:15-17:30</td>
<td>Welcome Speech</td>
<td>Honorary Chair - Professor Muzaffer SEKER (Rector of Necmettin Erbakan University, Konya TR)</td>
</tr>
<tr>
<td>17:30-18:15</td>
<td>Keynote Speaker – I</td>
<td>Prof. Dr. Franz THEURING, Charité - Universitätsmedizin Berlin, FRG&lt;br&gt;“Treating Alzheimer’s Disease by employing Tau-Aggregation-Inhibitors: the time for Tau is Now”</td>
</tr>
<tr>
<td>18:15-19:00</td>
<td>Keynote Speaker – II</td>
<td>Prof. Dr. Masahiro SUGIMOTO, Keio University, Japan&lt;br&gt;“Biomarker discoveries using mass spectrometry-based metabolomics platform”</td>
</tr>
<tr>
<td>19:00-20:00</td>
<td>Welcome Reception</td>
<td></td>
</tr>
</tbody>
</table>

### Monday, 20 October 2014

#### Session MON_S1: Invited Speeches
**Session Chairs: Wolfram GRONWALD**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Speaker</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00-09:45</td>
<td>Invited Speech-1</td>
<td>Prof. Dr. Dimitrios FOTIADIS, University of Ioannina, GREECE</td>
<td>&quot;Modelling of Long Bone Microstructure and Ultrasonic Assessment of Healing Mechanisms”</td>
</tr>
<tr>
<td>09:45-10:30</td>
<td>Invited Speech-2</td>
<td>Prof. Dr. Franz THEURING, Charité-Universitätsmedizin Berlin, FRG</td>
<td>&quot;Employing proteomic analysis to study female and male hearts in mice: insights into sex differences in cardiovascular diseases”</td>
</tr>
<tr>
<td>10:30-11:00</td>
<td>Coffee Break</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Session MON_S2: Invited Speeches
**Session Chairs: Franz THEURING**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Speaker</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00-11:45</td>
<td>Invited Speech-3</td>
<td>Assoc. Prof. Dr. Nesrin OZOREN, Bogazici University, TR</td>
<td>&quot;Whole Genome Sequencing of Turkish Genomes Reveals Functional Private Alleles and Impact of Genetic Interactions with Europe, Asia and Africa”</td>
</tr>
<tr>
<td>11:45-12:00</td>
<td>Presentation-1</td>
<td>Ulker Alkaya and Emrah Nikerel</td>
<td>&quot;Integration of transcriptome and fluxome data reveals insights in understanding the effect of weak acid stress in Saccharomyces cerevisiae”</td>
</tr>
</tbody>
</table>
### Session MON_S3:
**Session Chairs:** Nizamettin AYDIN

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker(s)</th>
</tr>
</thead>
</table>
| 12:00-12:15   | Presentation-2   | \* “Kinetic and molecular docking experiments on the inhibition of acetylcholinesterase by irgarol”  
Hakan Alyuruk and Levent Cavas |
| 12:15-12:30   | Presentation-3   | \* “Mutagenicity Assessment of Citalopram Impurities with Structure Activity Relationship Approach”  
Esra Emerce |
| 12:30-14:00   | Lunch            |                                                                             |
| 14:00-14:45   | Invited Speech-4 | **Prof. Dr. Ozlem KESKIN OZKAYA,** Koc University, TR  
“Three dimensional structures in signaling pathways towards revealing mechanisms in diseases” |
| 14:45-15:30   | Invited Speech-5 | **Prof. Dr. Michael SCHROEDER,** Technische Universitat Dresden, FRG  
“BVDU and beyond: Computational drug repositioning from herpes to cancer” |
| 15:30-16:00   | Coffee Break     |                                                                             |

### Session MON_S4:
**Session Chairs:** Figen ZIHNIOĞLU

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker(s)</th>
</tr>
</thead>
</table>
| 16:00-16:45   | Invited Speech-6 | **Prof. Dr. Alfred WINTER,** Universitat Leipzig, FRG  
“How to Assess Quality of Electronic Health Record Systems” |
| 16:45-17:00   | Presentation-4   | **Melih Gunay** and Rajarajeswari Balasubramaniyan  
“Data Management System of Taqman Array Card for Epidemiological Surveillance” |
| 17:00-17:15   | Presentation-5   | **Ferdi Sarac** and Ozgur Aktunc  
“Comparison of Web Accessibility Evaluation Tools for Visually Impaired and Blind People” |
| 17:15-17:30   | Presentation-6   | **Ali Yıldırım** and Sezer Domac  
“The missing professional in literature of the interprofessional collaborative practice: Speech and Language Therapy” |
| 17:30-17:45   | Presentation-7   | **Çiğdem Sazak,** Ali Yıldırım and Stuart Kerrigan  
“Mobile Device Application for Speech Therapy” |

### Session MON_S5:
**Session Chairs:** Graham BALL

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker(s)</th>
</tr>
</thead>
</table>
| 17:45-18:30   | Panel Discussion | **Siddik İÇLİ,** Ege University, (Alexander von Humboldt Alumni Association Izmir Branch Chair)  
“Alexander von Humboldt Foundation: Post-Doctoral Opportunities in GERMANY” |
|               |                  | **Michael SCHROEDER,** Technische Universitat Dresden, FRG  
Rainer MERKL,** University of Regensburg, Regensburg, FRG  
Graham BALL,** Nottingham Trent University, Nottingham, UK  
Masahiro SUGIMOTO,** Keio University, Japan  
Ugur SEZERMAN,** Sabanci University, TR  
“Career Development in Inter-Disciplinary Subjects; Informatics for Health and Life Sciences” |

### Session MON_S6:
**Session Chairs:** Alfred WINTER

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18:30-19:30</td>
<td>Poster Session-1</td>
<td><strong>Short Oral Presentations for Posters</strong></td>
</tr>
</tbody>
</table>

xiii
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00-09:45</td>
<td>Invited Speech-7</td>
<td>Prof. Dr. Rainer MERKL, University of Regensburg, Regensburg, FRG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Identifying in proteins functionally and evolutionarily important residues by means of computational biology”</td>
</tr>
<tr>
<td>09:45-10:00</td>
<td>Presentation-8</td>
<td>“Investigating the amigdala-thread effects on cognitive learning via graphical domains and Bayesian networks”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Serpil Ustebay, Ahmet Orun, Huseyin Seker and Ahmet Sertbas</td>
</tr>
<tr>
<td>10:00-10:15</td>
<td>Presentation-9</td>
<td>“Comparison of EEG Devices for Eye State Classification”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oliver Roesler, Lucas Bader, Jan Forster, Yoshikatsu Hayashi, Stefan Hebler and David Suendermann-Oeft</td>
</tr>
<tr>
<td>10:15-10:30</td>
<td>Presentation-10</td>
<td>“A First Step Towards Binaural Beat Classification Using Multiple EEG Devices”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jan Forster, Lucas Bader, Stefan Hebler, Oliver Roesler and David Suendermann-Oeft</td>
</tr>
<tr>
<td>10:30-11:00</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>11:00-11:45</td>
<td>Invited Speech-8</td>
<td>Prof. Dr. Graham BALL, Nottingham Trent University, UK</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“A Systems Biology Approach To Understand the Molecular Influences in the Basal-like Breast Cancer Molecular Subtype”</td>
</tr>
<tr>
<td>11:45-12:30</td>
<td>Invited Speech-9</td>
<td>Prof. Dr. Andreas BEYER, University of Cologne, FRG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“The impact of natural genetic variation on molecular traits”</td>
</tr>
<tr>
<td>12:30-14:00</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>14:00-14:45</td>
<td>Invited Speech-10</td>
<td>Prof. Dr. Tevfik DORAK, Liverpool Hope University, UK</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Nanotechnology Applications in Genetics”</td>
</tr>
<tr>
<td>14:45-15:00</td>
<td>Presentation-11</td>
<td>“Multiclass Disease Identification Employing Functional Protein Microarrays”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Patrick Kalmbach and David Suendermann-Oeft</td>
</tr>
<tr>
<td>15:00-15:15</td>
<td>Presentation-12</td>
<td>“Two-Stage Bioinformatics Approach for the Diagnosis of Hepatocellular Carcinoma and Discovery of its Bio-Network”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arinze Akutekwe and Huseyin Seker</td>
</tr>
<tr>
<td>15:15-15:30</td>
<td>Presentation-13</td>
<td>“Interpreting the Prevalence of Regulatory SNPs in Cancers and Protein-Coding SNPs among Non-Cancer Diseases Using GWAS Association Studies”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Zoya Khalid and Osman Ugur Sezerman</td>
</tr>
<tr>
<td>15:30-16:00</td>
<td>Coffee Break</td>
<td></td>
</tr>
</tbody>
</table>
Session TUE_S4:
Session Chairs: Uğur SEZERMAN

16:00-16:45 Invited Speech-11
Prof. Dr. Masahiro SUGIMOTO, Keio University, Japan
“Bioinformatics for processing, interpreting, analyzing and validating metabolomics data”

16:45-17:15 Invited Speech-12
Prof. Dr. Wolfram GRONWALD, University of Regensburg, FRG
“Recent Developments in State of the Art Analysis of Metabolomic Data”

17:15-17:30 Presentation-14
“Link Prediction for Drug-Drug Interaction Network”
Remzi Celebi, Vahab Mostafapour, Erkan Yasar, Hayriye Celikbilek, Sabrine Missaoui, Ozgur Gumus and Oguz Dikenelli

17:45-17:55 Presentation-15
“Amino acid preferences at neddylation sites”
Ahmet Sinan Yavuz, Namik Berk Sozer and Osman Ugur Sezerman

18:00-18:15 Presentation-16
“In silico determination of bioactive peptides in rubisc/o of invasive Caulerpa racemosa”
Zeynep Agirbasli and Levent Cavas

Session TUE_S5:
Session Chairs: Hedef Dhafir EL-YASSIN and Volkan USLAN

18:15-19:15 Poster Session-II
Short Oral Presentations for Posters

Wednesday, 22 October 2014
Session WED_S1:
Session Chairs: Rainer MERKL

09:00-09:45 Invited Speech-13
Prof. Dr. Ugur SEZERMAN, Sabanci University, TURKEY
“Identification of marker pathways from analysis of integrative ‘omics’ data”

09:45-10:30 Invited Speech-14
Prof. Dr. Tevfik DORAK, Liverpool Hope University, UK
“Bioinformatics Resources for Genetic Epidemiologists”

10:30-11:00 Coffee Break

Session WED_S2:
Session Chairs: Cenk SELÇUKİ

11:00-11:45 Invited Speech-15
Prof. Dr. Rainer MERKL, University of Regensburg, FRG
“Reconstructing the sequences of ancient proteins by means of computational biology”

11:45-12:00 Presentation-17
“A possible potential hormonal marker for the disorders of bone density in children with type 1 diabetes mellitus”
Hedef Dhafir EL-Yassin, Dunia Adnan and Muneeb Ahmed Al-Zubaidy

12:00-12:15 Presentation-18
“Analyzing Relations Among Measurements of Diabetic Neuropathy”
Olçay Kursun, M. Muzaffer Ilhan, Ahmet Cinar, M. Erdem Iserkul, C. Okan Sakar, A. Esra Gursoy, Ertegrul Tasan and Oleg V. Favorov

12:15-12:30 Presentation-19
“Integration of gene expression data and biological networks for predicting cancer outcomes”
Zerrin Isik, Janine Roy, Rengul Cetin-Atalay, Volkan Atalay and Michael Schroeder

12:30-13:00 Presentation of Awards for Student Paper Competition and Closing Speech
13:00-14:00 Lunch
14:00-17:00 Social Activity: Trip to Ephesus Ancient City
| Poster Oral Presentations-I  
Monday, 20 October 2014  
18:30- 19:30 |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P01</td>
<td>Chaos by Neural Networks: The Quasi-periodic Route</td>
</tr>
<tr>
<td>P02</td>
<td>Machine Learning Based Assessment of Resting Tremor for Evaluating Parkinson’s Disease</td>
</tr>
<tr>
<td>P03</td>
<td>On the comparison of validation techniques in protein classification</td>
</tr>
<tr>
<td>P04</td>
<td>Mercury Sensing Based On Genetically Encoded Fluorescent Protein and MerR</td>
</tr>
<tr>
<td>P05</td>
<td>An Implantable Meander-Line Printed Antenna for Biomedical Applications</td>
</tr>
<tr>
<td>P06</td>
<td>Understanding Brain Metabolic Network Changes in case of Glioblastoma Using Computational Systems Biology Approaches</td>
</tr>
<tr>
<td>P07</td>
<td>Genetic Distance Measurements by Using rRNA and miRNA Sequences</td>
</tr>
<tr>
<td>P08</td>
<td>Assessing the bias correction methods due to exposure measurement error in the European carbon black study</td>
</tr>
<tr>
<td>P09</td>
<td>Identification of Pathways from Proteomic Analysis with DAVID and PANOGA</td>
</tr>
<tr>
<td>P10</td>
<td>The investigation of the prenatal and postnatal alcohol exposure- induced neurodegeneration in rat brain: protection by betaine and/or omega-3</td>
</tr>
<tr>
<td>P11</td>
<td>Descriptive consideration of serum Irisin levels various factors: obesity, type 2 diabetes mellitus, pre-diabetic status, gender and athletics</td>
</tr>
<tr>
<td>P12</td>
<td>YUFES Web Server: Building Feature Vectors of Proteins</td>
</tr>
<tr>
<td>P13</td>
<td>GeneSeq: A software program for processing dna sequences</td>
</tr>
<tr>
<td>P15</td>
<td>A Comparative Study on Diabetes Disease Diagnosis Using Data Mining Classification Techniques</td>
</tr>
<tr>
<td>P16</td>
<td>Interprofessional education model for the health and life science faculties</td>
</tr>
<tr>
<td>P18</td>
<td>Extraction, characterization and evaluation of the activity of Artemisia plant as Antiseptic</td>
</tr>
<tr>
<td>P19</td>
<td>Feasibility Study of Telemedicine Implementation in West of Iran (Alimoradian Hospital: a case study)</td>
</tr>
<tr>
<td>P20</td>
<td>CNI-1493 administration improves the efficacy of cytotoxic T lymphocytes</td>
</tr>
<tr>
<td>Presentation</td>
<td>Title</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>P21</td>
<td>Molecular Modeling of Interactions of Zwitterionic Histidine Forms with Cu⁺ and Cu²⁻</td>
</tr>
<tr>
<td>P22</td>
<td>Conformation Analysis of N-Acetylglucosamine: A Density Functional Study</td>
</tr>
<tr>
<td>P23</td>
<td>M-Box Riboswitches: Modeling Perspective, Computational Study of Nucleobases and Their Interactions with Mono and Di Mg²⁺ Cations</td>
</tr>
<tr>
<td>P24</td>
<td>Amperometric biosensor for epinephrine based on laccase immobilized on polypyrrole-polyvinylsulphonate film</td>
</tr>
<tr>
<td>P25</td>
<td>Computational Investigation of Interaction of Gluconate Derivatives with Magnetite (Fe₃O₄) Surface in α-D-glucose coated iron oxide nanoparticles</td>
</tr>
<tr>
<td>P27</td>
<td>Mutations in mu phage proteins lead loss in lipase activity</td>
</tr>
<tr>
<td>P28</td>
<td>Electrochemical determination of glucose based on the immobilization of glucose oxidase onto poly(2,3-di(thiophene-2-yl) naphthalene -1,4-dione) modified glassy carbon electrode</td>
</tr>
<tr>
<td>P29</td>
<td>Production of polyvinyl alcohol/gelatin based cyrogel scaffolds for tissue engineering applications</td>
</tr>
<tr>
<td>P30</td>
<td>Dipolar Ferrocene and Ruthenocene Second-Order Nonlinear Optical Chromophores: A Time-Dependent Density Functional Theory Investigation of Their Absorption Spectra</td>
</tr>
<tr>
<td>P31</td>
<td>Investigation of DNA binding properties of the calix[4]arene amide derivatives</td>
</tr>
<tr>
<td>P32</td>
<td>Optimization of the weights and features in use of AHP for SNP prioritization</td>
</tr>
<tr>
<td>P33</td>
<td>Investigation of in vitro effect of Alkanna tinctoria root organic phase extracts on human erytrocyte 6-phosphogluconate dehydrogenose, glucose-6-phosphate dehydrogenase and glutathione reductase enzymes</td>
</tr>
<tr>
<td>P34</td>
<td>Functional Proteomics in Lipid Research</td>
</tr>
<tr>
<td>P35</td>
<td>General Approach to Bioactive Peptide Production</td>
</tr>
<tr>
<td>P36</td>
<td>Preparation of encapsulated therapeutic L-asparaginase nanoparticles</td>
</tr>
<tr>
<td>P37</td>
<td>Nanoencapsulation of L-asparaginase in Biosilica Support</td>
</tr>
<tr>
<td>P38</td>
<td>Adaptive-Network Fuzzy System for the Prediction of Binding Affinity of Peptides</td>
</tr>
<tr>
<td>P39</td>
<td>Measurement of Refractive Index using Terahertz Reflectance Mode on Fixed Quartz Sample Holder</td>
</tr>
<tr>
<td>P40</td>
<td>“Microarray Gene Expression Data-Based Bioinformatics Method for the Diagnosis of Male Hypertension”</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

Welcome Message | VIII
Conference Programme | XII
Table of Contents | XVIII

**Prof. Dr. Franz THEURING,** Charite - Universitätsmedizin Berlin, FRG  
“Treating Alzheimer’s Disease by employing Tau-Aggregation-Inhibitors: the time for Tau is Now” | 1

**Prof. Dr. Masahiro SUGIMOTO,** Keio University, Japan  
“Biomarker discoveries using mass spectrometry-based metabolomics platform” | 3

**Prof. Dr. Dimitrios FOTIADIS,** University of Ioannina, GREECE  
”Modelling of Long Bone Microstructure and Ultrasonic Assessment of Healing Mechanisms” | 6

**Prof. Dr. Franz THEURING,** Charite-Universitats medizin Berlin, FRG  
“Employing proteomic analysis to study female and male hearts in mice: insights into sex differences in cardiovascular diseases” | 10

**Assoc. Prof. Dr. Nesrin OZOREN,** Bogazici University, TR  
”Whole Genome Sequencing of Turkish Genomes Reveals Functional Private Alleles and Impact of Genetic Interactions with Europe, Asia and Africa ” | 13

**Prof.Dr.Ozlem KESKIN OZKAYA,** Koc University, TR  
”Three dimensional structures in signaling pathways towards revealing mechanisms in diseases” | 14

**Prof. Dr. Michael SCHRODER,** Technische Universitat Dresden, FRG  
”BVDU and beyond: Computational drug repositioning from herpes to cancer” | 15

**Prof. Dr. Alfred WINTER,** Universitat Leipzig, FRG  
”How to Assess Quality of Electronic Health Record Systems” | 16

**Prof. Dr. Rainer MERKL,** University of Regensburg, Regensburg, FRG  
”Identifying in proteins functionally and evolutionary important residues by means of computational biology” | 19

**Prof. Dr. Graham BALL,** Nottingham Trent University, UK  
“A Systems Biology Approach To Understand the Molecular Influences in the Basal-like Breast Cancer Molecular Subtype” | 23

**Prof. Dr. Tevfik DORAK,** Liverpool Hope University, UK  
”Nanotechnology Applications in Genetics” | 28

**Prof. Dr. Masahiro SUGIMOTO,** Keio University, Japan  
”Bioinformatics for processing, interpreting, analyzing and validating metabolomics data” | 31

**Prof. Dr. Wolfram GRONWALD,** University of Regensburg, FRG  
”Recent Developments in State of the Art Analysis of Metabolomic Data” | 36

**Prof. Dr. Ugur SEZERMAN,** Sabanci University, TURKEY  
”Identification of marker pathways from analysis of integrative ‘omics’ data” | 40

**Prof. Dr. Tevfik DORAK,** Liverpool Hope University, UK  
”Bioinformatics Resources for Genetic Epidemiologists” | 41
Prof. Dr. Rainer MERKL, University of Regensburg, FRG
"Reconstructing the sequences of ancient proteins by means of computational biology"

“Integration of transcriptome and fluxome data reveals insights in understanding the effect of weak acid stress in Saccharomyces cerevisiae”
Ulker Alkaya and Emrah Nikerel

“Kinetic and molecular docking experiments on the inhibition of acetylcholinesterase by irgarol”
Hakan Alyuruk and Levent Cavas

“Mutagenicity Assessment of Citalopram Impurities with Structure Activity Relationship Approach”
Esra Emerce

“Data Management System of Taqman Array Card for Epidemiological Surveillance”
Melih Gunay and Rajarajeswari Balasubramaniyan

“Comparison of Web Accessibility Evaluation Tools for Visually Impaired and Blind People”
Feidi Sarac and Ozgur Aktunc

“The missing professional in literature of the interprofessional collaborative practice: Speech and Language Therapy”
Ali Yıldırım and Sezer Domac

“Mobile Device Application for Speech Therapy”
Çiğdem Sazak, Ali Yıldırım and Stuart Kerrigan

“Investigating the amigdala-thread effects on cognitive learning via graphical domains and Bayesian networks”
Serpil Ustebay, Ahmet Orun, Huseyin Seker and Ahmet Sertbas

“Comparison of EEG Devices for Eye State Classification”
Oliver Roesler, Lucas Bader, Jan Forster, Yoshikatsu Hayashi, Stefan Hebler and David Suendermann-Oeft

“A First Step Towards Binaural Beat Classification Using Multiple EEG Devices”
Jan Forster, Lucas Bader, Stefan Hebler, Oliver Roesler and David Suendermann-Oeft

“Multiclass Disease Identification Employing Functional Protein Microarrays”
Patrick Kalmbach and David Suendermann-Oeft

“Two-Stage Bioinformatics Approach for the Diagnosis of Hepatocellular Carcinoma and Discovery of its Bio-Network”
Arinze Akutekwe and Huseyin Seker

“Interpreting the Prevalence of Regulatory SNPs in Cancers and Protein-Coding SNPs among Non-Cancer Diseases Using GWAS Association Studies”
Zoya Khalid and Osman Ugur Sezerman

“Link Prediction for Drug-Drug Interaction Network”
Remzi Celebi, Vahab Mostafapour, Erkan Yasar, Hayriye Celikbilek, Sabrine Missaoui, Ozgur Gumus and Oguz Dikenelli

“Amino acid preferences at neddylation sites”
Ahmet Sinan Yavuz, Namik Berk Sozer and Osman Ugur Sezerman

“In silico determination of bioactive peptides in rubisc/o of invasive Caulerpa racemosa”
Zeynep Agirbasli and Levent Cavas
“A possible potential hormonal marker for the disorders of bone density in children with type 1 diabetes mellitus”
Hedef Dhafir EL-Yassin, Dunia Adnan and Muneeb Ahmed Al-Zubaidy

“Analyzing Relations Among Measurements of Diabetic Neuropathy”
Olçay Kursun, M. Muazaffer Ilhan, Ahmet Cinar, M. Erdem Isenkul, C. Okan Sakar, A. Esra Gursoy, Ertuğrul Tasan and Oleg V. Favorov

Integration of gene expression data and biological networks for predicting cancer outcomes
Zerrin Isık, Janine Roy 1 and Michael Schroeder

“Chaos by Neural Networks: The Quasi-periodic Route”
Marat Akhmet, Mehmet Onur Fen and Ayşegül Kıvılcım

“Machine Learning Based Assessment of Resting Tremor for Evaluating Parkinson’s Disease”
Ayse Betul Oktay, Munevver Okay and Abdulkadir Kocer

“On the comparison of validation techniques in protein classification”
Çağın Kandemir-Çavaş and Selen Yıldırım

“Mercury Sensing Based On Genetically Encoded Fluorescent Protein and MerR”
Canan Özyurt, Serap Evran and Azmi Telefoncu

“An Implantable Meander-Line Printed Antenna for Biomedical Applications”
Metin Akdağ, Yunus E. Yamaç, Yalçın Albayrak and S. Cumhur Başaran

“Understanding Brain Metabolic Network Changes in case of Glioblastoma Using Computational Systems Biology Approaches”
Emrah Özcak and Tunahan Çakır

“Genetic Distance Measurements by Using rRNA and miRNA Sequences”
Nazife Çevik, M. Erdem Isenkul and Olçay Kursun

“Assessing the bias correction methods due to exposure measurement error in the European carbon black study”
Gonca Mert, Roseanne McNamee and Evridiki Batistatou

“Identification of Pathways from Proteomic Analysis with DAVID and PANOGA”
İlknur Melis Durası and Osman Ugur Sezerman

“The investigation of the prenatal and postnatal alcohol exposure- induced neurodegeneration in rat brain: protection by betaine and/or omega-3”
Kevser Kuşat Ol, Günsör Kanbak, Ayşegül Oğläkçı, Dilek Burukoğlu and Ferruh Yücel

“Descriptive consideration of serum Irisin levels various factors: obesity, type 2 diabetes mellitus, pre-diabetic status, gender and athletics”
Basil O. Saleh, Maysaa J. Majeed and Ghassan M. Oreaby

“YUFES Web Server: Building Feature Vectors of Proteins”
Uğur TURHAL and Murat GÖK

“GeneSeq: A software program for processing dna sequences”
Deniz Tanır and Urfat Nuriyev

“UGT1A1*28 and CYP3A4*1B polymorphisms in prognosis of diffuse large B-cell lymphoma patients”
Yagut Akbarova, Hilal Akalin, Sule Ketenci Ertas, Mehmet Boz, Ali Unal, Munis Dundar and Yusuf Ozkul
“Optimization of the weights and features in use of AHP for SNP prioritization”
Arif Yılmaz and Yeşim AYDIN SON

“Investigation of in vitro effect of Alkanna tinctoria root organic phase extracts on human erythrocyte 6-phosphogluconate dehydrogenase, glucose-6-phosphate dehydrogenase and glutathione reductase enzymes”
M. Çakar, V. Türkoğlu, Z. Bası, M. Güler and M.R. Kıvanc

“Functional Proteomics in Lipid Research”
Funda Kartal

“General Approach to Bioactive Peptide Production”
H. Kübra Kani and Ebru K. Kocazorbaz

“Preparation of encapsulated therapeutic L-asparaginase nanoparticles”
Didem Songurtekin, Duygu Asker, Taylan K Öztürk, Ali Kılınç

“Nanoencapsulation of L-asparaginase in Biosilica Support”
Taylan K. Öztürk, Didem Songurtekin, Duygu Asker, Ali Kılınç

“Adaptive-Network Fuzzy System for the Prediction of Binding Affinity of Peptides”
Volkan Uslan, Hanife Catman

“Measurement of Refractive Index using Terahertz Reflectance Mode on Fixed Quartz Sample Holder”
Martin Mueller-Holtz, Huseyin Seker and Geoff Smith

“Mapping the impact of natural genetic variation on molecular traits”
Andreas Beyer

“Microarray Gene Expression Data-Based Bioinformatics Method for the Diagnosis of Male Hypertension”
Nazlı Ucunoglu, Arinze Akutekwe and Turgay Isbir

Author Index

Sponsors
Treating Alzheimer's Disease by employing Tau-Aggregation-Inhibitors: the time for Tau is Now
Franz Theuring 1,2, Karima Schwab 1,2, Gernot Riedel 3, John M.D. Storey 2,4, Charles R. Harrington 2,3, Claude M. Wischik 2,3

Abstract - Given the repeated failures of amyloid-based approaches in Alzheimer's disease (AD), there is increasing interest in tau-based therapeutics. We have developed two transgenic mouse models of tauopathy which exhibit different but convergent pathways of tau pathophysiology. Line 66, expressing full-length tau carrying a double mutation, shows prominent widespread tangle-like pathology but models only the motor features of frontotemporal lobe degeneration disorders without cognitive deficits. Line 1, expressing truncated tau296-390 from the paired helical filament (PHF)-core of the AD PHFs targeted to the endoplasmic reticulum, has a much weaker histopathological phenotype, but provides a model for the cognitive phenotype of AD without confounding motor features and shows amplification and spread of pathology with age similar to Braak staging. Methylthioninium (MT) in its oxidised form as MTC (methylene blue) has limited clinical utility because of dose dependent food interference in absorption, poor tolerability and first pass metabolism. We have developed a novel stabilised reduced form of MT (leucomethylthioninium; LMTX®) which has superior clinical and pharmacokinetic properties. We now report the comparative efficacy of MTC and LMTX® with respect to the behavioural and histopathological features in the transgenic models.

The data establish that dianaminophenothiazine compounds like MT can reverse both cognitive and motor learning deficits and reduce the underlying tau pathology in two distinct tau transgenic mouse lines modelling different tauopathy phenotypes. The brain concentrations at which efficacy can be demonstrated are consistent with the concentrations of MT required for tau aggregation inhibitor activity in vitro and with the expected brain levels at the minimum effective dose determined clinically in a phase 2 trial in mild/moderate AD. These compounds therefore offer potential for the treatment of a range of tauopathies and three independent clinical phase 3 trials have been initiated worldwide to treat mild/moderate AD and Frontotemporal Dementia (FTD).

INTRODUCTION
Alzheimer’s disease (AD) is an irreversible, neurodegenerative disorder characterized by the progressive loss of memory and thinking skills. It was first presented by Alois Alzheimer, who discovered “neurofibrillary tangles” in the brain tissue of a woman who died with dementia at the age of 55 (1). The neurofibrillary tangle is composed of tau protein (2,3), and tau aggregation in the brain is directly linked to clinical dementia (4). This correlation has been formalised in the Braak staging system, which defines 6 stages according to the spread of neurofibrillary pathology (5). There are currently two main hypotheses regarding the cause of dementia in AD: the Aβ cascade hypothesis based on aggregation of extracellular Aβ, and the tau aggregation hypothesis based on intracellular tau aggregates. Contrary to the common belief that neurofibrillary pathology is a late stage process in AD, analysis of a large post-mortem series showed that Braak stage 1 begins in 4th and 5th decades of life, and that it takes some 20-30 years to develop poverty dementia (6). In another set of analyses, Duyckaerts and colleagues demonstrated that the tau pathology appears some 27 years earlier then the amyloid β pathology (7). The clinic-pathological correlations, strongly supported by the genetic evidence for a primary role for tau protein aggregation in a wide range of neurodegenerative disorders led to the inference that a drug which blocks the aberrant tau binding interaction which is necessary for aggregation could have beneficial effects in the treatment of AD.

Given the repeated failures of amyloid-based approaches in Alzheimer’s disease (AD), there is increasing interest in tau-based therapeutics. To establish the validity of the tau aggregation hypothesis we have developed two transgenic mouse models of tauopathy which exhibit different but convergent pathways of tau pathophysiology. The tau-transgenic animal models demonstrate that tau aggregation on its own is sufficient to produce cognitive and other behavioural defects, and that blocking tau aggregation reverses these defects. Dianaminophenothiazines were identified as a class of pharmaceutically viable compounds inhibiting selectively tau-tau aggregation in vitro without affecting the normal tau-tubulin interaction. This was established using a tau-tau inhibition assay that resembles the autocalytic propagation of tau capture and aggregation arising within affected neurons in AD. To determine whether these compounds might be effective clinically in preventing tau aggregation, we sought to use these transgenic mice models to test the physiological effects of such drugs. Our findings show that the substances, MTC (rember™) and the follow-up molecule LMTM™, a compound with a more enhanced efficacy profile in our in vivo experiments, both act as Tau-Aggregation-Inhibitors (TAIs) in the transgenic mice models. Both, tau pathology and behavioural phenotypes are reversed by these inhibitors at concentrations consistent with in vitro activity.

CONCLUSIONS
These data provide prominent support to the rationale for treating AD using TAI therapy, and after MTC had successfully passed a phase 2 clinical trial, LMTM™ now has entered into phase 3 clinical trials to treat behavioural variant Frontotemporal Dementia (bvFTD) and AD, two neurodegenerative diseases where a prominent tau pathology is observed.
REFERENCES


Figure 1. Fluorescent immunostaining of a wild type spinal chord section (left) vs. immunostaining of an L66 spinal chord section (right). Note the speckled pattern of Tau signals in some of the neuronal cell bodies, indicated by arrows.
Biomarker discoveries using mass spectrometry-based metabolomics platform
Masahiro Sugimoto

Abstract—To understand the complex molecular interactions inter and intra cells, various omics technologies have been developed for simultaneous profiling of multiple molecules. Genomics, transcriptomics, and proteomics are already well established techniques for analyzing genome, ribonucleic acids (RNA), and proteins. Meanwhile, metabolomics is relatively recently emerged omics that can identify and quantify low weight organic molecules (no strict definition but usually the molecules < 1500 Da) named metabolites. Currently, mass spectrometry (MS) is the dominantly used as a profiling instrument with separation systems prior to MS. We have developed analytical techniques using capillary electrophoresis-time-of-flight-MS (CE-TOF-MS) for profiling charged metabolites, which suits understanding of primary metabolism as well as biomarker discoveries. Here, we introduce the recent clinical applications using this technique, and the bioinformatics topics, such as data-processing of raw data and interpretation using multi-omics datasets including both metabolomics and transcriptomics data.

INTRODUCTION

Advent of high-throughput molecular profiling techniques, omics analysis, enabled us the simultaneous quantifying hundreds molecules in living systems, which provide us holistic views that help understanding of cellular functions. DNA sequencing (genomics or genome), gene expression profiling (transcriptomics or transcriptome), protein interactions (proteomics or proteome), and metabolic pathways (metabolomics or metabolome) are major omics research fields. We have developed measurement protocols of metabolomics, and conducted a variety of application studies, such as novel biomarkers for probing cancer and liver diseases. We also have developed data processing tools, databases, and pathway visualization tools in metabolomics.

What is metabolomics?

In a biological system, active response to environmental and/or intrinsic changes involves numerous molecular interactions. The central dogma of molecular biology stipulates that information is transferred from DNA to RNA and subsequently to proteins. Metabolites are regulated by proteins, the end product of the central dogma. Thus, among the omics fields, metabolomics resides at the end of the downstream of the central dogma. Therefore, the information available using metabolomics are considered to be closely relates to the phenotypes of living systems. Therefore, we have focused on this molecular set which might contain informative properties reflecting the current status of a living system.

Analytical techniques of metabolomics

Unlike those of other omics technologies, the analytical technologies of metabolomics have not yet matured. Because of their wide variety of molecular chemical properties, metabolites cannot be measured by a single method. Thus, in a strict sense, no methodology exists for conducting metabolomic analysis i.e. the simultaneous and quantification of small molecules comprehensively. Based on the target and purpose, we have to select or combine the suitable methods.

Currently, nuclear magnetic resonance (NMR) and mass spectrometry (MS) are the most commonly used in metabolomics field. NMR provides definitive advantage that enables non-destructive sample treatment, which facilitate metabolic screening of any samples [1]. However, even in the newest one, NMR shows only low sensitivity and there this method can detect only a few set of metabolites. Therefore, this method does not suit the novel biomarkers while pattern recognition techniques are usually used for discriminating the type of given samples.

By contrast, MS is more sensitive and can observe a greater variety of metabolites [2]. MS is usually preceded by a separation system such as gas chromatography (GC), liquid chromatography (LC), and capillary electrophoresis (CE). GC-MS is a common and well established method for profiling volatile compounds [3], while LC-MS has become more popular for profiling diverse neutral or low charged molecules, such as lipids and saccharides [4]. However, depending on the target analytes, MS requires a suitable LC column and optimal sample treatment and measurement conditions, i.e. we have to optimize many options. CE-MS, an emerging technology among these hyphenated-MS methods, simultaneously identify and quantify charged metabolites using only two modes; one for cationic (positively charged) metabolites, the other for anionic (negatively charged) metabolites [5].

How are measurement techniques selected?

GC-MS is ideal for analyzing foodstuff and beverage flavors because perfume and aromatic molecules have volatile properties. The metabolites of secondary metabolic pathways are also frequently profiled by GC-MS, but the derivations required for non-volatile metabolites limits the throughput of these methods. LC-MS can profile a wide range of metabolites and is commonly used for lipid and sugar profiling. Representative examples are immunological reactions such as macrophage activation [6], and the compositional analysis of foodstuffs and beverage [7, 8]. CE-MS is suitable for profiling charged metabolites in primary pathways, such as glycolysis, the tricarboxylic acid (TCA) cycle, the pentose phosphate pathway (PPP), amino acid synthesis and degradation, and nucleotide pathways. In fact, 80% of primary pathway metabolites are charged and can thus be simultaneously quantified by CE-MS [5]. Recently, metabolomic anomalies in energy metabolism have been reported in the
principle pathways of cancer cells; therefore, CE-MS is a valuable tool for cancer-related research [9, 10] (Fig. 1). It is also suitable for detecting compounds in foodstuffs, such as amino acids and organic acids [8, 11].

**BIological applications**

**Warburg effect in tumor cells**

Tumor cells undergo aggressive proliferation by unexpected signal transductions derived from accumulated mutation of oncogenes. Therefore, tumor cells consume much higher quantities of the energy source ATP than do normal cells. Tumor cells are expected to be chronically hypoxic because of insufficient oxygen supply in their microenvironment. Therefore, tumor cells activate the glycolysis pathway to generate ATP, despite the low efficiency of glycolysis compared to oxidative phosphorylation. However, Warburg [14] found that, in the tumor cells, glycolysis is preferentially activated over oxidative phosphorylation in the presence of sufficient oxygen.

One possible reason for this unusual behavior is that glycolysis yields not only ATP but also NADPH. The latter is used for generating palmitate, one of the fatty acids constituting the cellular membrane, which is required for cell proliferation. Thus, glycolysis activation induces concurrent activation of the PPP cycle to fulfill these requirements. However, these findings have been observed in cancer cell lines and have yet to be confirmed in situ.

We conducted metabolomic analysis of tumor tissues in patients with colon and gastric cancer [10]. To prevent individual variation, healthy tissues were obtained close to the tumor tissue. The observed profile showed characteristic signs of glycolysis activation; namely, the decrease of glucose and accumulation of lactate. In addition, the concentration of hypoxanthine was raised in tumor tissue, indicating that carbon resources were supplemented by autophagy rather than by glucose. Interestingly, most of the amino acid end products of many metabolic pathways significantly differ between tumor and matched control samples, although the energy change within both tissue types is unexpectedly similar. The mechanism behind this phenomenon has yet to be elucidated.

**Metabolite biomarker discovery**

Drug-induced liver diseases constitute a critical medical problem. Acute hepatitis can be induced by unexpected sensitivity to acetaminophen (APAP), a common ingredient in cold medicine. Soga et al [15] compared the metabolic profiles of livers in control and APAP-overdosed mice, and found that reduced glutathione (GSH), a detoxification metabolite, was depleted in the latter group, accompanied by increased oxidative stress. Therefore, APAP stress seems to activate GSH synthesis from cysteine, with concurrent synthesis of ophthalmic acid by the same enzyme. Because these metabolites are secreted into the blood, the presence of ophthalmic acid in blood indicates oxidative stress in the liver. Human livers secrete γ-glutamyl dipeptides as well as ophthalmic acid. Collectively, these metabolites can successfully discriminate various liver diseases, such as drug-induced injury, hepatitis C, and hepatocellular carcinoma [12]. Such mechanism-based biomarker discovery is more reliable than data mining of the molecular profiles extracted from biofluid; however, the difference between model species, e.g. mouse and human, is frequently problematic and its resolution requires systematic analytical algorithms.
In collaboration with Dr. D.T. Wong at UCLA University, we discovered biomarkers of oral cancer in saliva [9]. In comparison of saliva from the patients with oral cancer, breast cancer, pancreatic cancer, and periodontal diseases, as well as healthy controls, cancer-specific aberrance of 54 metabolites were observed. However, the single metabolites cannot identify a disease from the other groups clearly, we have developed classification model using multiple metabolites. Although the multiple logistic regression model incorporating several metabolites clearly discriminated the cancer patients from healthy controls, the salivary metabolites were also sensitive to other environmental factors [13]. Thus, robust prediction methodologies should be developed for small sample sizes and noisy data.

**CONCLUSIONS**

Bioinformatics plays a key role in metabolomics. To fill the gap between analyst and biologist, novel algorithms for data processing, data interpretation tools, and database integration should be developed. The difference between microarray data and metabolomic profiles is a primary consideration in this development.

**ACKNOWLEDGMENTS**

The author thanks Drs. Soga and Tomita for fruitful discussions. The author also thanks the funding agency for its financial support, JSPS KAKENHI Grant Number 2471025 and research funds from Yamagata Prefectural Government and Tsuruoka City.

**REFERENCES**

Modelling of Long Bone Microstructure and Ultrasonic Assessment of Healing Mechanisms

Vassiliki T. Potsika, Vasilios C. Protopappas, Ioannis F. Spiridon, Demosthenes Polyzos and Dimitrios I. Fotiadis*

Abstract—Bone is a heterogeneous and composite medium with a complex, hierarchical structure. Several research groups have used ultrasonic methods in order to investigate experimentally and computationally the structure of bone at different hierarchical scales. Recently, realistic computational models of long bones have been developed using powerful computational tools which make efficient use of memory and computation time. In addition, the availability of high-resolution two-dimensional and three-dimensional imaging data of bone’s microstructure has significantly contributed to the development of computational models which incorporate its complex and porous nature. Thus, nowadays the investigation of bone status in simulated cases of pathologies or traumas such as osteoporosis and fracture healing using computational techniques can provide insight into complicated wave propagation phenomena that cannot be observed using experimental setup. This study presents a comprehensive state of the art as well as the current research progress in the scientific domain of computational modeling of ultrasound propagation in pathologic bones so as to highlight the significant monitoring and diagnostic role of ultrasound.

INTRODUCTION

Osteoporosis is responsible for approximately 2 million fractures that occur annually, including hip, vertebral, wrist and other fractures [1]. Fractures caused by the weakness of the skeleton due to osteoporosis may lead to diminished quality of life, disability and even death. In addition, the financial costs for the monitoring and treatment of osteoporosis and fracture healing are tremendous. Specifically, by 2025 the annual costs from osteoporosis are expected to reach approximately $25.3 billion [1].

The main cause of osteoporosis is hormonal deficiency, and thus the frequent disease is post menopausal osteoporosis [2]. At least 40% of post menopausal women over the age of 50 and 15–30% of men will sustain one or more fragility fractures [2]. The bone resorption starts from the trabecularisation of the inner cortical layer and provokes the thinning of the cortex. In clinical practice, the assessment of osteoporosis has been mostly based on bone mineral density measurements by using Dual energy X-ray Absorptiometry (DEXA) scan, which, however, is an inconvenient and expensive method [3].

Fracture healing is one of the most remarkable repair processes since it does not result in a scar, but in the actual reconstitution of the injured tissues in the original material properties and structural integrity. It involves sequential stages of callus formation and consolidation and is normally completed within some months. However, in about 5-10% of the fractures complications such as delayed unions and non-unions are encountered [4]. As a series of cellular and molecular interactions are evolving, it is not to be expected that the mechanisms controlling such a complex process can be easily elucidated and monitored. The assessment of fracture healing is conducted by serial clinical and radiographic examinations, which strongly depend on the orthopedic surgeon’s experience and judgment. Thus, the development of more objective and quantitative means to diagnose and monitor bone pathologies is extremely important.

To this end, several research groups worldwide have investigated the effectiveness of quantitative ultrasound (QUS) to evaluate bone status by using experimental and computational techniques [1-26]. Due to the relative low cost and portability, QUS is considered as a promising non-invasive and non-radiating diagnostic and monitoring tool which could potentially contribute to the prediction of osteoporotic fractures [3]. More recently, the exploitation of numerical simulations as a complementary tool to experimental techniques has extended our knowledge on the underlying wave propagation mechanisms since the sensitivity of ultrasonic features to material and structural changes in bone can be independently investigated. Moreover, in computational studies the role of the different types of ultrasonic waves can be studied more easily and complications such as experimental artefacts can be avoided [3]. The development of imaging techniques such as micro-computed tomography (μ-CT) and scanning acoustic microscopy (SAM) has enhanced the development of realistic computational models which incorporate bone microstructure based on high-resolution two-dimensional (2D) and three-dimensional (3D) imaging data. This study aims to review the milestones in the field of computational modelling of ultrasound wave propagation in osteoporotic and healing long bones and presents the current status of knowledge.

BONE STRUCTURE

The mechanical rigidity of bone depends on its composite, porous and hierarchical structure as well as on the material properties which vary at different scales of the macrostructure, microstructure and nanostructure level. Thus, the first step for the development of realistic computational models of bone requires the investigation of the mechanical properties of its component phases and the structural interaction of the different levels of hierarchical organization.

* D.I. Fotiadis is with the Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, Greece, and with the Foundation for Research and Technology-Hellas (FORTH) (email: fotiadis@cs.uoi.gr).
V.T. Potsika, V.C. Protopappas and I.F. Spiridon are with the Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, Greece, and with the Foundation for Research and Technology-Hellas (FORTH) (email: vprotopappas@gmail.com, spiridon@cc.uoi.gr).
D.K. Polyzos is with the Department of Mechanical Engineering and Aeronautics, University of Patras, Greece (e-mail: polyzos@mech.upatras.gr).
Concerning the macrostructural level, the main components of bone are the periosteum, osseous tissue, marrow, blood vessels and nerves. The osseous tissue is consisted of two different tissue types, namely the cortical and the trabecular bone, which differ on their degree of porosity and density. Specifically, the trabecular bone with porosities ranging from 50−95% is found in the inner layer of bone, surrounded by the cortical bone [5]. From a microstructure point of view, trabecular bone is composed of randomly distributed cylindrical structures filled with marrow, called trabeculae, with a thickness of about 100 μm. Cortical bone covers the external bone layer and its porosity ranges from 5−10% [2, 3]. The thickness of cortical bone ranges from few millimetres to even centimetres in the mid shaft [5]. In the microstructure scale, it is composed of cylindrical structures known as the Haversian canals which are arranged in parallel to the long axis of the bone and their diameter is approximately 10-500 μm [2, 3]. Additionally, the Volkmann canals are arranged transverse to the long axis of the bone which connect the Haversian canals.

The main components of bone in the nanostructure level are collagen and hydroxypatite which consist of collagen molecules organized in fibrils [2, 3]. The compression strength and stiffness depend on the properties of the hydroxypatite crystals, while the collagen fibers, proteoglycans and osteocalcin determine the tension properties.

Therefore, it is evident that the development of realistic computational models of bone depends on the incorporation of the microstructural properties and geometrical features. The proper exploitation and integration of data such as μ-CT and SAM images have opened new horizons for the computational investigation of bone status and health at different hierarchical levels.

MODELING AND ASSESSMENT OF HEALTHY AND PATHOLOGIC LONG BONES

Computational Modeling of Intact and Osteoporotic Bones

The first research groups that investigated the diagnostic role of ultrasound in osteoporotic bones developed simple 2D computational models of cortical bone [6-16]. The main quantitative indicators for the assessment of material and structural changes were calculations of the first arriving signal (FAS) velocity and attenuation. In addition, the propagation of guided waves has been studied as an advanced means that can provide insight to ultrasound propagation phenomena occurring at inner cortical bone layers. Recent studies have investigated the influence of cortical porosity and pores’ size on the propagation of ultrasound [13-15] using SAM images. Also, 3D computational models have been presented in order to apply more realistic conditions [16].

Our research group has studied analytically and computationally [11, 12] the propagation of guided and Rayleigh waves in 2D bone models with microstructural effects using the Mindlin Form II theory of gradient elasticity. In [11], intrinsic parameters were incorporated which correlate the bone’s macrostructure with microstructure. It was found that the microstructure of bone has a significant influence on the propagation of guided waves and should not be neglected. Additionally, in [12] both the simple and the general Mindlin Form II of gradient elastic theory were used to investigate the effect of bone’s microstructure on the propagation of Rayleigh waves. It was shown that Rayleigh waves are dispersive only when the microstructural effects are represented by different shear stiffness and inertia internal length scale parameters [12].

Our current study [14, 15] investigates the influence of cortical porosity on the propagation of FAS and guided waves. SAM images have been used in order to incorporate realistic conditions of porosity distribution. Initially, in [14] simple 2D numerical models were presented in which cortical bone was modeled as a 2D plate with different porosities. Then, in [15] realistic 2D computational models of osteoporotic and healthy bones were developed based on SAM images corresponding to a well-characterized set of human tibia SAM images presenting: a) a highly porous bone with large cortical thickness, b) an osteoporotic bone with low cortical thickness, c) an osteoporotic bone with large cortical thickness, d) a healthy bone with low porosity and large cortical thickness. The preliminary results have shown higher FAS velocities for healthy bone and lower values for osteoporotic bones. Moreover, for the excitation frequency of 1 MHz the analysis of the propagation of guided waves indicated that the A1 mode is detected in the lower frequency range for osteoporotic bones, while the same mode was observed for higher frequencies for healthy bone [15]. However, further research is needed to understand deeply the influence of cortical porosity on the propagation of guided waves.

Computational Modeling of Fracture Healing

Several non-invasive quantitative techniques have been reported in the literature for the monitoring of fracture healing such as bone densitometry, vibrational analysis, acoustic emission, and the attachment of strain gauges to external fixation devices. However, these methods depend on extrinsic bone properties and most of them require a clinical setting together with the intervention of a specialist to configure the experimental set-up [4]. To this end, our research group [17-21,26] has studied experimentally and computationally the monitoring role of QUS in fracture healing. In this section we present the computational models which have been developed to study wave propagation as well as the most significant findings in the field.

In our first computational study [17], we presented 2D isotropic models of cortical bone and callus. The material properties of callus were varying so as to simulate the healing process in seven stages. Also, the periosteal and endosteal formation were incorporated in the callus geometry. It was shown that the FAS velocity decreases
during the first and the second healing stages and gradually increases at later healing stages. However, the FAS velocity could not give insight to material and structural changes occurring within the endosteal callus tissue at different healing stages. To this end, analysis of the propagation of guided waves was also performed. Signal analysis was carried out in the time-frequency (t, f) domain based on the reassigned smoothed pseudo Wigner-Ville (RSPWV) energy distribution. Then, the Lamb wave theory was introduced to derive the group velocity dispersion curves, which were superimposed in the (t, f) diagrams in order to investigate the evolution of the dominant modes. It was found that changes in the callus during fracture healing had a significant influence on the propagation of guided waves.

In [18] we studied the influence of the soft tissues surrounding cortical bone on FAS velocity and guided waves. Three different cases of fluid loading boundary conditions were considered. Callus was modeled as a non-homogeneous 2D medium and fracture healing was simulated as a three-stage process. It was found that the different fluid loading boundary conditions had a minor effect on FAS velocity, while their influence is significant on guided waves.

Our study was later extended to investigate the propagation of ultrasound when the transducers are placed on the pins of an existing external fixation device [19]. Callus was modeled as a 2D non-homogeneous material consisting of six ossification regions. Two stainless steel pins were implanted in cortical bone models considering different inclination angles. Higher velocity measurements were found due to the incorporation of the pins as the metal material of the pins has a higher bulk velocity than that of cortical bone. In addition, we observed that the FAS velocity increases at later healing stages and this behavior is not influenced by the different inclination angles.

Recently, realistic computational models of healing bones have been developed that incorporate callus porosity based on SAM images [20]. In Fig. 1 the SAM images are shown which represent longitudinal sections of 3-mm osteotomies in the tibia of female Merino sheep after 3, 6 and 9 weeks of consolidation [21]. In [20] an iterative effective medium approximation [22] was applied to evaluate theoretically and numerically wave dispersion and attenuation. Callus was modeled as a composite medium consisting of a matrix with spherical inclusions of blood. It was shown that the phase velocity decreases with increasing frequency, while the attenuation coefficient increases exponentially. Also, the scattering, material dispersion and absorption phenomena were found to be more significant during the early healing stages due to the callus porosity.

Other research groups have presented 2D computational studies using different material and geometrical features to model the callus [23, 24, 25]. The main estimated parameters were the FAS velocity and attenuation [23, 24]. In [23], the FAS propagation time was found to decrease during healing, while the callus composition could not well explain the changes in energy attenuation. In [24], the results clearly indicated that a large net loss in the signal amplitude was produced for all the examined geometries, while the arrival time and the signal amplitude followed a different behavior depending on the receiver position and the fracture geometry. More recently, in [25], the mode amplitude variations of the two fundamental guided waves were examined as functions of the gap-breakage width and depth. It was shown that reflection energy and transmission coefficients of the S0 and A0 modes can be used to evaluate long bone fracture status.

Although 2D studies have contributed to the interpretation of real bone measurements, the development of 3D computational models of healing long bones is necessary. To this end, in [26] a 3D computational model of healing long bones was developed to account for the irregular geometry of cortical bone (Fig. 2). Different types of material symmetry were considered and callus was modeled according to [18]. It was observed that guided modes are sensitive to material and structural changes in contrast to the FAS velocity [26].

**DISCUSSION**

Ultrasonic evaluation of bone is a complicated procedure due to its porous nature and hierarchical structure. Assessment techniques based on QUS have attracted the interest of many research groups as they are promising non-invasive and non-ionizing diagnostic and monitoring means.

Computational studies on ultrasonic wave propagation in pathologic bones have contributed to the interpretation of experimental findings. The main estimated parameters are the FAS velocity and attenuation as well as the dispersion of guided waves. Concerning intact and osteoporotic bones, it has been found that: a) bone’s microstructure and anisotropy affects the propagation of guided modes, b) cortical porosity influences significantly the FAS propagation [6-16]. In the context of fracture healing, a
decrease of the FAS velocity has been observed during the first healing stages followed by a constant increase as healing progresses [17-26]. However, the FAS velocity was not sensitive to different types of material symmetry and the presence of the soft tissues. Thus, the study of guided waves is considered crucial.

In conclusion, computational findings have contributed significantly to the interpretation and validation of experimental observations. Nevertheless, several issues need to be further addressed such as multi-scale modeling, the ultrasound configuration, the simulation algorithm and the estimated parameters in order to incorporate more realistic conditions and represent the clinical problem accurately.

ACKNOWLEDGEMENT

This work is part funded by the State Scholarships Foundation (IKY) and the German Academic Exchange Service (DAAD).

REFERENCES

Employing proteomic analysis to study female and male hearts in mice: insights into sex differences in cardiovascular diseases

Franz Theuring, Karima Schwab

Abstract - Proteomic techniques based on high resolution two-dimensional electrophoresis and mass spectrometry provide a very effective approach to explore differentially expressed proteins in various physiological and pathophysiological processes in a non-biased way. With the introduction of transgenic technologies for altering the genetic make up of mice very precisely and efficiently, and the adaptation and miniaturization of cardiovascular techniques to the mouse format, we are now able to address fundamental questions related to cardiovascular disease appropriately not only to the rat, but also to this animal model. The combination of these state-of-the-art technologies will contribute to a better understanding of the mechanisms of cardiovascular diseases and their perturbations and thereby will help in reducing human morbidity and mortality.

INTRODUCTION

The two-dimensional electrophoresis (2-DE) can be traced back some 50 years when Smithies and Poulik developed this technology in 1956 by combining filter paper electrophoresis and starch gel electrophoresis for the first and second dimension, respectively (1). Human serum proteins could be separated into 20 spots by this method. In the following years, a number of other 2-DE technologies were developed, which finally led to the two independent publications in 1975 by Klose (2) and O’Farrell (3) describing for the first time the attempt to resolve highly complex protein mixtures from total cell extracts derived from mouse tissues and E.coli, respectively. Since then the high resolution 2-DE technique has been tremendously modified and improved and even more importantly, micromethods for the molecular analysis of proteins from 2-DE gels became available for partial sequencing, amino acid composition determinations, and for the mass spectrometry determination of peptide compositions (for a review see 4). Technological development has been a major driver of opportunities in proteomics, as it has been in genomics. By combining these approaches with the introduction of the world wide web (www) and the establishment of appropriate data bases there is now a powerful technology available for the endeavour first to understand and then to treat human diseases.

Proteomics-based approaches, which examine the expressed proteins of a tissue or cell type, complement the genome initiatives and are increasingly being used to address biomedical questions. High resolution 2-DE represents a unique non-biased method for large scale protein characterization and discovery. By comparing 2-DE protein patterns from different disease states, different tissues, cell fractions, and developmental stages, proteins can be characterized and cataloged according to different biological parameters. Proteomics, which examines the unique protein repertoire of a sample, including posttranslational modifications and subcellular localization is capable of identifying events at the protein level that change during disease progression (for review see 5, 6). Such disease-associated changes may identify new diagnostic markers, disease fingerprints and new drug targets for therapeutic intervention.

Cardiovascular disease remains a prominent feature of basic science and clinical investigation throughout the developed world. Understanding the disease processes requires clarification of the underlying mechanistic changes in organ phenotype over time, the influence of genetic variations, and the effects of pharmacologic, surgical, and interventional treatment. It is clear, however, that knowledge of the DNA sequence, although essential, is not sufficient. A more meaningful understanding of gene expression can be achieved through characterization of the products of that expression, i.e. the biological active proteins that are the essential biological determinants of disease phenotype.

Gender, age and phytoestrogen related comparative proteomics analysis of mouse heart

Cardiovascular diseases are the number one cause of morbidity and mortality, both for men and women (7). Premenopausal women are at a lower risk for cardiovascular disease as compared with age-matched men, but this risk increases dramatically after menopause (8), indicating that sexual hormones, such as androgens and estrogens, may contribute to the pathogenesis of these disorders. This notion is supported by findings from animal models, in which removal of ovarian hormones results in diastolic dysfunction with interstitial cardiac fibrosis (9), a decrease in cardiac contractility and the development of ventricular dilatation (10). Estrogen replacement on the other hand exerts vasoprotective effects by multiple mechanisms, e.g. decreased ventricular remodeling, a decrease in low-density lipoprotein levels, an increase in high-density lipoprotein levels, modulation of vascular tone and inhibition of vascular growth (11-13). In humans, studies with hormone replacement therapy (HRT) yielded conflicting results. Observational studies found a reduction in mortality and the incidence of cardiovascular diseases in postmenopausal women receiving HRT as compared with women who did not. However, randomized controlled trials failed to confirm these findings (14). Thus alternatives for traditional HRT are intensely discussed with phytoestrogens representing a class of potential candidate compounds. Their application as dietary supplements has a long history of use in traditional medicine, particularly in Asia (15). Genistein, one of the best studied phytoestrogens, has agonistic activities for both estrogen receptors (16) and was shown...
to exert beneficial effects on serum cholesterol levels, artery occlusion, endothelial function and ventricular hypertrophy (17-19).

An often ignored pitfall in cardiovascular studies carried out in rodents is the fact that the routinely used standard rodent chow exhibits high and variable levels of phytoestrogens, which were shown to influence the cardiovascular system (20-22). Thus, benefits of estrogens and/or phytoestrogens may interfere (23) and studies employing the administration of defined amounts of phytoestrogens to animals both in the absence and presence of endogenous sex hormones are needed in order to better understand the impact of these compounds on cardiovascular physiology. Moreover, global studies employing the normal, non-failing mouse myocardium are rare. Identifying and characterizing protein patterns involved in the maintenance of normal heart physiology will help understanding disease conditions. In this context, a proteomic approach, i.e. protein separation by 2-DE- and MS-based identification, is the first choice to realize a hypothesis-free approach. Moreover, enormous advancement in this field makes combined proteomics techniques indispensable to relatively quantify proteins at their protein species level. The protein species is defined chemically (24) including all post-translational modifications and genetic variations.

The aim of the present study is to characterize an animal model for alternative hormone replacement with genistein as a natural estrogenic compound. We performed a 2-DE/ESI-LC-MS approach in order to identify protein species varying with genistein receipt and sex in their relative abundance in the healthy murine heart (http://www.mpiib-berlin.mpg.de/2D-PAGE). Oral genistein treatment revealed a substantial effect on the relative abundance of both estrogen receptors. Several enzymes of the fatty acid metabolism and their transcriptional regulators varied differentially in male and in female animals, at the transcript and/or the protein species level. Increased levels of enzyme species involved in the oxidative phosphorylation and generation of ROS were accompanied by decreased amounts of antioxidants in male mice receiving genistein compared with control males, which have been previously associated with various pathological conditions. Exposure of female animals to genistein provoked an increased abundance of two species of LIM domain-binding protein and one species of desmin. These proteins have been associated with cardiac hypertrophy and our data warrant caution for the use of them as molecular markers, since the animals did not exhibit any histological signs of cardiac hypertrophy (25, 26).

There is evidence that isoflavones, such as genistein, can directly or indirectly improve lipid profile and lower blood pressure and hence exert cardiovascular protection. It is further believed, that genistein attenuates vascular contraction and thus vascular tone and blood pressure through altering the phosphorylation of the regulatory myosin light chain (MLC) probably via the myosin light chain kinase (MLCK) or the RhoA pathway. However, the direct role of genistein in the myocardium is poorly reviewed. Therefore, we investigated the effect of dietary genistein intake on the myocardium in an ovariecotomized mouse model, performing 2DE and ion-trap Fourier transform ion cyclotron resonance MS/MS (LTQ-FT-ICR-MS/MS) proteomics (27).

REFERENCES


Whole Genome Sequencing of Turkish Genomes Reveals Functional Private Alleles and Impact of Genetic Interactions with Europe, Asia and Africa

Can Alkan1,2, Pinar Kavak3,4, Mehmet Somel5, Omer Gökçumen6, Serkan Uğurlu7, Elif Dal2, Kuyaş Buğra-Bilge7, Tunga Güngör5, S. Cenk Sahinalp6, Nesrin Özören7*, Cemalettin Bekpen7*

1Department of Genome Sciences, University of Washington, Seattle, WA 98195, United States
2Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey
3Department of Computer Engineering, Boğaziçi University, İstanbul 34342, Turkey
4TÜBİTAK - BILGEM - UEKAE (The Scientific and Technological Research Council of Turkey, Informatics and Information Security Research Center, National Research Institute of Electronics and Cryptology), Gebze, Kocaeli 41470, Turkey
5Department of Integrative Biology, University of California, Berkeley, CA 94720, United States
6Brigham and Women’s Hospital, Harvard Medical School, Boston, MA 02115, United States
7Department of Molecular Biology and Genetics, Boğaziçi University, İstanbul 34342, Turkey
8School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Abstract—Turkey is a crossroads of major population movements throughout history and has been a hotspot of cultural interactions. Several studies have investigated the complex population history of Turkey through a limited set of genetic markers. However, to date, there have been no studies to assess the genetic variation at the whole genome level and without ascertainment bias. Here, we present whole genome sequences of 16 Turkish individuals resequenced at high coverage (32x-48x). Our study provides the first map of common genetic variation in western Asia. We show that the genetic variation of the contemporary Turkish population clusters with South European populations, as expected, but also shows signatures of relatively recent contribution from ancestral East Asian populations. In addition, we document a significant enrichment of non-synonymous private alleles, consistent with recent observations in European populations. A number of variants associated with skin color and total cholesterol levels show frequency differentiation between the Turkish populations and European populations. Furthermore, we have analyzed the the 17q21.31 inversion polymorphism region (MAPT locus) and found increased allele frequency of 31.25% for H1/H2 inversion polymorphism when compared to European populations which shows about 25% of allele frequency. Our data will help develop population-specific experimental designs for studies investigating disease associations and demographic history in Turkey.

INTRODUCTION

High throughput sequencing technologies have prompted sequencing of human genomes at the population level. These projects provide us with immense amounts of information regarding human genomic variation and the functional properties of such variation. Still, the coverage of world-wide variation remains limited, and virtually no whole genome resequencing data is available involving populations in western Asia, a region encompassing the eastern Mediterranean basin and the Middle East [1].

FURTHER INFORMATION

Western Asia has been the main corridor through which initial out-of-Africa migrations have populated Eurasia. Sequencing studies not only provide insight into the demographic history of the population, but also helping to gain prominence for identifying variants that are associated with complex, sometimes population specific diseases. Such assessments are not possible using array technologies, as array probes are designed based on common variation and hence suffer from ascertainment bias [2]. However, rare or private variants have stirred recent attention and may explain heritable diseases and local adaptation. Such variants may explain the genetic components of diseases that occur at unusual frequencies in Turkey, including Behçet’s [3], familial Mediterranean fever [4], and beta thalassemia [5].

CONCLUSIONS

This study provides the first map of common genetic variation from 16 individuals in western Asia, thus it helps fill an important gap in analyzing natural human variation and human migration. Our data will help develop population-specific experimental designs for studies.

ACKNOWLEDGMENT

The authors thank Boğaziçi University Research Funds (BRF/BAP-Infrastructure), the FP7-REGPOT project MBG-BRIDGE and two EMBO Installation Grants (EMBO ALTF 1475–2010).

REFERENCES


*Nesrin Özören (e-mail: nesrin.ozoren@boun.edu.tr).
*Cemalettin Bekpen (e-mail: cemalettin.bekpen@boun.edu.tr).
Three dimensional structures in signaling pathways towards revealing mechanisms in diseases
Özlem Keskin Özkaya, Atila Gursoy, Ece Acuner-Ozbabacan, Billur Engin

Abstract—3D Structural information of proteins in signaling pathways are important because they provide insight into signaling mechanisms; help understand the mechanism of disease-related mutations; and help in drug discovery. While extremely useful, common pathway diagrams lacking structural data are unable to provide mechanistic insight to explain oncogenic mutations or SNPs. Here we focus on the construction of structural pathways and map oncogenic mutations and SNPs to complexes in this pathway. Our results indicate that computational modeling of protein-protein interactions on a large scale can provide accurate, structural atom-level detail of signaling pathways in the human cell and help delineate the mechanism through which a mutation leads to disease. We show that the mutations either thwart the interactions, activating the proteins even in their absence or stabilize them, leading to the same uncontrolled outcome. Computational mapping of mutations on the interface of the predicted complexes may constitute an effective strategy to explain the mechanisms of mutations—constitutive activation or deactivation.

INTRODUCTION
Cancers are mostly due to somatic mutations and environmental factors, and chronic inflammation is implicated by most of these risk factors. Chronic inflammation, due to autoimmune diseases or infections, causes tumor development via several mechanisms, including oncogenic mutation induction. Oncogenic mutations and single nucleotide polymorphisms (SNPs) are key players in inflammation-related cancers and it is crucial to map the mutations/SNPs on the corresponding 3D structures of the proteins to gain insight into how they affect protein function. SNPs that cause diseases, if not in the core of the protein, are frequently located in protein-protein interface regions rather than elsewhere on the surface. Structural knowledge can clarify the conformational and functional impact of the mutation/SNP on the protein. The effect of a functional mutation can be expressed by a change in the specificity of the interactions between a mutated protein and its partners. Quantitatively, a mutation changes the binding free energies of the mutant’s interactions with its partners with respect to the free energies of its interactions in the native form. The functional impact of the mutation on the specificity differs. The mutation can destabilize the protein and/or its interaction, leading to ‘loss-of-function’; or can lead to a change in the specificity of protein-partner interactions, resulting in a ‘gain-of-function’, or can gain new binding partners and hence a new biological function, i.e. result in a ‘switch-of-function’.

Here, we construct the IL-1 signaling pathway by combining the related pathways and information from the literature (Figure 1). We observe that there are approximately 100 interactions between proteins that have experimentally identified 3D structures. However, only 15 of the interactions have structures of protein-protein complexes in the Protein Data Bank (PDB, http://www.rcsb.org/pdb/). The structural coverage of the pathway is under 20%. Our major aim is to expand the structural apoptosis pathway [1], and map oncogenic mutations and SNPs to reveal their mechanism.

IL-1 signaling pathway reconstructed by combining related pathways and information from the literature. This detailed map of IL-1 signaling presents the protein-protein interactions and the resulting cellular events. The colored nodes represent proteins having experimentally identified 3D structures and the white nodes are the proteins without 3D structures. The edges represent protein-protein interactions (straight/dashed arrows relate to available/unavailable 3D structures of proteins) or associations leading to cellular events such as cell cycle or gene expression (dashed arrows beginning with circular heads).

CONCLUSIONS
By predicting protein-protein complexes throughout the pathway using the PRISM algorithm, the structural coverage of the pathway increased up to 71%. The distributions of oncogenic mutations and SNPs in the predicted structures of protein-protein complexes indicated that they significantly correspond to interface and adjoining residues, and more importantly, in some cases to

REFERENCES
1. This work is from Acuner-Ozbabacan et al, Plos Comp Biol 2013.
**Computational drug repositioning from herpes to cancer**

Michael Schroeder
Biotec, TU Dresden, Germany

Abstract—BVDU is a herpes drug, which is on the market since the 80s. Recently, there is evidence that it may help to treat cancer. I will shed light on the mode of action of BVDU in herpes and in cancer and show how computational drug repositioning with structural data can systematically uncover novel drug-target relationships.

BVDU is a thymidine analog, which was developed in the 80s and early 90s as herpes treatment. After being introduced on the market as herpes drug, it was tested on cancer patients suffering from herpes. Subsequently, it emerged that it may be beneficial in cancer. Over the last two decades, BVDU was further developed culminating in a phase II clinical trial on BVDU as pancreas cancer treatment.

How can a drug work against herpes and cancer? A side effect in one indication may be a desired effect in another. Drugs are promiscuous and often bind multiple targets, which is one reason for side effects, but also an opportunity for repositioning. Thus, the question arises what the targets of BVDU are. In the herpes indication, BVDU acts by disturbing the replication of the virus. Being a thymidine analog, BVDU is phosphorylated and integrated into the viral DNA. The viral thymidine kinase (TK) is therefore a target of BVDU. But nonetheless, the interaction of BVDU and the TK is vital, since the mode of binding is understood. There is a protein structure showing exactly how two phenylalanine rings hold BVDU in the TK’s binding pocket. Thymidine kinases also exist in human, so is a human TK the main target that helps in cancer treatment? No, it turned out to be a human heat shock protein, Hsp27, which is a key target in cancer. It is surprising since the viral TK and the human Hsp27 do not share any evolutionary relationships and there is not the slightest sequence similarity. However, it may be of course, that there is structure similarity despite missing sequence similarity. Therefore, we modeled Hsp27. Part of the protein, the so called alpha crystalin domain are solved and could therefore be used for modeling. However, a large flexible and unstructured part was not available and had to be modeled without template. The model of Hsp27 does not share structural similarity with the viral TK, so BVDU is a drug binding two targets without sequence- and without structural similarity. However, something is similar. We identified pockets on the surface of the Hsp27 model to find potential binding sites. The second deepest pocket in Hsp27 turns out to bear resemblance to the binding site in TK. In Hsp27, there are the two phenylalanine rings and three more residues in exactly the same position as in the TK. So, TK and Hsp27 share a binding site for BVDU and a handful of residues appear to be the key for binding. This hypothesis was confirmed in mutation experiments, where the two phenylalanines were mutated and indeed the drug does not bind any longer.

In a nutshell, binding site similarity is key. This raises many questions. Do these remote similarities evolve by chance? Or is there a functional requirement directing their independent evolution? Is binding site similarity generally key for drug promiscuity? Many authors have studied properties of drugs such as hydrophobicity or molecular weight for the correlation to drug promiscuity.

We investigated this question systematically. We selected all drugs in the protein data bank PDB that bind more than three distinct targets. We tested whether their degree of promiscuity correlates with weight, hydrophobicity, drug flexibility and binding site similarity. The result was clear. Only binding site similarity correlates. This finding has some important implications: Algorithms that can identify drug promiscuity by pinpointing remotely similar binding sites can be used to screen for targets and therefore pave the way to understand and quantify potential side and desired effects of a drug. Furthermore, the knowledge about the binding site may help to optimise drugs and identify stronger and better binders than the original drug considered.

ACKNOWLEDGMENT

I’d like to thank the bioinformatics group, who developed these wonderful findings and EU, BMBF, BMWi, DFG who support our research.

REFERENCES


*Corresponding Author (ms@biotec.tu-dresden.de).
How to Assess Quality of Electronic Health Record Systems

Alfred Winter*

Abstract—Electronic Health Record Systems (EHRS) are crucial for patient centered care processes in modern health care. The quality of EHRS cannot be assessed by simply assessing a software product for managing the content of the respective health record. Moreover outcome, process and structural quality have to be taken into account. This results in various quality criteria concerning the three layers of Health Information Systems, the three scopes of information management and the resources for information management. Last but not least it turns out that education in Medical Informatics is a vital resource and thus quality criterion for EHRS as well.

INTRODUCTION

Patient related data is stored in a respective record. If the record contains data representing information in computer processable form it is called an electronic record. An electronic record containing a patient’s medical data from one health care institution like a hospital, clinic or practice is called an Electronic Patient Record (EPR). If the record includes a patient's entire health history, then it is called an Electronic Health Record (EHR). Thus an EHR does not contain only medical data from one institution, but health related documents and data from all institutions which provided care to the patient and even wellness data [1, 2]. The system "for recording, retrieving and manipulating information in electronic health records" is called the Electronic Health Record System (EHRS) [3].

EHRS and therefore EHRS are essential for modern health care which is moving from isolated procedures in a single healthcare institution (e.g. a hospital or a general practice) to patient-oriented care processes spreading over institutional boundaries.

Due to this importance, high quality and innovative EHRS are needed.

But what does high quality mean with respect to EHRS? Is it sufficient to buy and install a particular high quality software product?

This paper’s first objective is to show, that an EHRS is not simply one software product but is more or less identical to a complex transinstitutional Health Information System (tHIS) [1, 4]. Thus quality of EHRS depends strongly on the quality of its corresponding tHIS.

The second objective is to give a broad range of examples for different quality aspects related to tHIS and thus to show that assessing the quality of tHIS and EHRS will not result in a scalar grade but in a multidimensional assessment.

EPRS, EHRS, HIS, and tHIS

According to [4] an information system is that socio-technical subsystem of an institution that processes and stores data, information, and knowledge. In a health care institution the respective information system is called the Health Information System (HIS); in case of a network of health care institutions the respective information system is called the transinstitutional Health Information System (tHIS).

Hence an institutional HIS is the system for recording, retrieving and manipulating information of the institution’s EPR, i.e. the institution’s Electronic Patient Record System (EPRS). The transinstitutional HIS (tHIS) however is the Electronic Health Record System (EHRS) of the respective health care network. This paper will often refer to HIS as a more general term including both, institutional as well as transinstitutional HIS.

QUALITY OF HIS, tHIS AND EHRS

A HIS can be considered as the product, i.e. the outcome of information management. Information management in an institution is planning, monitoring and directing the institution’s information system [5]. In many institutions a CIO is responsible for this task.

Taking Donabedian’s notion of “quality of medical care” [6, 7] into account we can differentiate three dimensions of quality, which are as well applicable to HIS in general and tHIS in particular. The dimensions are outcome quality of information management, i.e. the quality of the product HIS itself, process quality of information management, i.e. the quality of HIS production processes and finally structural quality, i.e. the quality of resources available for production.

Outcome quality of information management: Quality of the product HIS

HIS can be described by concepts on three layers [8]: The domain layer describes a HIS by the enterprise functions or business and health care processes it supports and the data which are processed. The logical tool layer concentrates on application components supporting the processes and storing the data. Computer based application components are installed software. The physical tool layer consists of physical data processing systems (like personal computers, servers, switches, routers, etc.) on which the application components are installed.

Each of these layers has to be subject of quality assessment:

Domain layer

Assessing quality of an EHRS at the domain layer means assessing the quality of both processes and data.

Quality of data, i.e. the content of an EHRS, has many aspects. Data has to be complete with regard to different types of documents (e.g. relevant reports, findings and images from all examinations) and to different sources, locations and institutions (e.g. documents from in- an outpatient encounters). Data should be structured - otherwise it could not be used e.g. for the improvement of medication safety [9]. Data should be annotated by semantic information to ensure correct interpretation and to avoid ambiguity. Besides health care, data of an EHRS...
should be usable for medical research as well. Ensuring data privacy is a major issue especially in the medical domain. It requires a lot of measures to be checked at the logical and physical tool layer of the HIS. Classical methods for access control turn out to be insufficient. For high quality EHRS data usage control is needed, “that extends authority not only to who may access which data, but also to how the data may or may not be used or distributed afterward” [10].

The EHRS provides data for enabling processes. Quality of health care processes as e.g. medical documentation can be examined from a medical professional’s perspective. Criteria like timeliness of certain clinical documents, their availability, time needed, user satisfaction, completeness/correctness of finished clinical documents, compliance with legal regulations, usability etc. can be applied and measured. [11]

But processes in health care look different if viewed from a patient’s perspective. Consumers in general and patients in particular want to be self-determined and therefore may want to select and compile even complex health services self dependently. High quality HIS have to support – at least in near future - such kind of customer enabled orchestration of complex services. This will have widespread consequences since many needs and areas of life are affected. [12]

Logical tool layer
At this layer HIS have to have an appropriate, integrated architecture providing all applications needed for meeting the aforementioned quality requirements at the domain layer.

For being able to provide complete data from all relevant sources and locations a HIS has to encompass the institutional HISs of the region with their local application systems ranging from patient management systems and enterprise resource management systems to PACS and CPOE. [4]

So called independent health record banks (IHRB) are intended to provide patient/individual-centric, longitudinal, cross-institutional and interoperable health records. Its key principles are to have independent/trusted organizations that curate and manage the records on behalf of patients and make it available to all authorized parties, to aggregate the records of each patient logically in one place and to give patients greater control over their own records with no need for globally unique patient’s id. [13] Hence IHRBs are excellent means for integrating medical data in an tHIS in order to present a high quality EHR.

Since data shall be usable for clinical research as well, a comprehensive clinical research database with semantically annotated data should be integrated too. Besides a conventional relational database medical data warehouse software like i2b2 is useful to serve as an easy to use frontend for medical doctors. This way patient recruitment for clinical studies, cohort identification and quality assurance in daily routine can be supported. [14]

As has been seen before, the demand for supporting patients’ self-determination affects many areas of life. Especially for elderly people their houses and apartments have to be considered. Communication architectures are necessary based on common definitions of communication scenarios, application components and communication standards. Respective tHIS architectures at their logical tool layer can be designed using reference models. [15]

Usage of standards is a key criterion for high quality logical tool layers of tHIS. The Integrating the Healthcare Enterprise initiative (IHE) [16] provides profiles describing usage of standards for many scenarios. An enterprise architecture modeling tool like 3LGOS [17] can provide a respective reference model supporting the design of standards based data integration in complex tHIS. [18]

Physical tool layer
At the physical tool layer the availability of high performance computers and communication networks is among the quality criteria. The level of achievable performance depends on the respective environment and quality criteria have to be adjusted respectively. For example in rural areas of developing countries usage of SMS in cell phone networks may be considered as a very efficient and thus performant solution for data exchange [19].

But besides performance effective security measures guaranteeing privacy are crucial. EHR demand for nationwide secure infrastructures; but unfortunately they are at best in quite early stages [20, 21].

Process quality of a Health Information System: quality of information management

“If health information systems are not systematically managed and operated, they tend to develop chaotically. This, in turn, leads to negative consequences, such as low data quality, resulting low quality of patient care, and increasing costs. Systematic information management can help to prevent such HIS failures and contribute to a high-quality and efficient patient care.” [4] Strategic, tactical and operational information management can be distinguished.

In total these different scopes of information management describe how an HIS or a tHIS is planned, directed and monitored, i.e. how they are produced and their quality is managed.

Strategic information management is the basis of information management and has to translate the business strategy of a health care institution or a health care network into an appropriate information strategy. A respective strategy plan should contain a project portfolio as a step-by-step path from the current HIS state to the planned state. It should assign resources and priorities to the projects [5]. The existence of such a strategic IT plan is an important quality criterion for benchmarking HIS [22]. Further criteria for good strategic information management are also provided e.g. by COBIT [23].

To execute the projects of a strategic IT plan is task of tactical information management. Again systematic planning is required. High process quality requires that projects for updating a HIS follow strict plans as e.g. described in [24].

ITIL provides a comprehensive set of good practices for high quality operational information management [25]. If operational services for EHRS are based on that, they can
be compared with those of other institutions and potentials for improving them can be detected more easily.

**Structural quality of a Health Information System:**

Quality of resources for information management

Good information management needs appropriate resources like money and experienced and skilled staff for producing good EHRs. This does not only hold for the installation of new components of the respective HIS. Since quality of EHRs is also a matter of sustainability, those resources are needed for its operational management during the entire life cycle.

Skilled personnel is among the most important resources. Unfortunately those persons don’t fall out of the blue sky. Hence especially on a national scale the availability of Medical Informatics education e.g. according to the IMIA guidelines [26] is an important criterion for structural quality of EHRs.

**CONCLUSIONS**

Assessing quality of EHRs means to assess the quality of the entire underlying HIS or even tHIS which is needed to provide an Electronic Health Record. This assessment is multidimensional and comprises aspects dealing with the content of the EHR, with the architecture of the HIS and tHIS and its level of integration, with systematic strategic, tactical and operational information management and with sufficient resources especially for a sustainable operation.

In summary it means that for achieving high quality Electronic Health Records selecting and installing excellent software is not enough. Moreover continuous and systematic information management, a careful view on a huge and heterogeneous information system and skilled personnel are needed.

**REFERENCES**

Identifying in proteins functionally and evolutionary important residues by means of computational biology

Jan-Oliver Janda¹, Markus Busch², Patrick Löffler¹, Maximilian Plach³, and Rainer Merkl¹*  

Abstract— The reliable identification of functionally and structurally important residues of a protein by means of computational methods is still an open problem and most state-of-the-art classifiers predict only one or at most two different categories. Here we present the concept of the classifier CLIPS-4D, which predicts for each residue-position of a protein in a mutually exclusively manner a role in catalysis, ligand-binding, or protein stability. We further demonstrate how these predictions can be supplemented by means of H2rs, which identifies evolutionary and functionally important residues based on an orthogonal approach.

INTRODUCTION

An important goal of molecular biochemistry is the detailed analysis of protein characteristics like functionality, stability or dynamics. However, this is a laborious and time consuming task due to the many aspects of protein function and the enormous spectrum of experimental methods to be considered for their determination. In the optimal case, one would characterize experimentally the contribution of each individual amino acid residue, which is however not feasible for larger proteins. This is why the biochemical assessment of proteins has to concentrate on a relatively small number of residues. In enzymes, these are the residues directly involved in catalysis (catalytic sites) and substrate binding (ligand binding sites). If the structure of the protein is known, resulting annotations can be found in dedicated databases like PDBsum. However, there exist no equivalent databases one can consult to identify residues which are important for stability or other characteristics.

Due to the enormous success of genome sequencing projects, the sequences of more than 17 600 protein families (InterPro Version 48) are known at date and thus, methods of computational biology are of utmost importance to support their characterization. A large number of in silico approaches are at hand to identify important residues, see [1] and references therein. Often, a family-specific multiple sequence alignment (MSA) is the main data source to elucidate the role of the residues; for a review see ref. [2]. Most effective is the assessment of residue conservation deduced from the corresponding MSA columns. The success of these analyses is due to the biochemical properties of the residues: For example, in most cases only one residue-type fulfills all critical requirements at catalytic sites, which prohibits a mutation. Accordingly, strict residue conservation is a strong indicator signaling functionally important residues. In contrast, a prevalent but not exclusively found amino acid is often important for protein stability, which similarly holds true for ligand-binding sites.

In the following, we distinguish four types of residue-positions, i.e. CLASSes. These are sites directly involved in catalysis (CAT_sites), ligand-binding sites (LIG_sites), sites important for protein structure (STRUC_sites) and all other residue-positions (NOANN_sites). To exploit the above mentioned patterns of residue conservation for the assignment of these roles, we have implemented CLIPS-4D [1].

Interestingly, less conserved residue-positions may bear a pattern indicative of dependencies in the occupancy of two or more positions. The importance of these correlation signals and their consequences have long been realized and quite different approaches have been introduced to identify correlated residue pairs. To exploit these signals, we recently designed an improved method named H2rs [3], which utilizes the von Neumann entropy for the analysis of correlated mutations.

COMMON PRINCIPLES FOR THE CHARACTERIZATION OF RESIDUE-POSITIONS.

As explained above, a characteristic property (feature) of functionally important sites is the conservation of a residue. It follows, that one needs a larger set of sequences which represent homologs having the same function. Databases like InterPro offer for each protein family such a set, which is already arranged in an optimal way named multiple sequence alignment (MSA). Each column of the MSA corresponds to a site in the protein structure (left side, backbone presentation). Here, the mapping of two columns onto two residue-positions k and l is shown. Generally, one representative protein structure is chosen as template.

A more detailed analysis of enzymes and related MSAs makes clear that conservation is not limited to functional sites. In many cases, the local sequence neighborhood is conserved as well. Moreover, at catalytic sites the distribution of residues is skewed: Due to their biochemical properties some residues like histidine are significantly over- and some other residues like proline are significantly underrepresented at CAT_sites.
These observations propose to characterize each residue position by computing several parameters. In the context of machine learning, these properties are named features. If probability distributions are needed to compute the numerical value of a feature, these have to be deduced beforehand by analyzing cases which are known to possess the property under study. In the case of functionally important sites, we used the annotation of the catalytic site atlas of the EBI to determine the CLASS-specific probabilities.

**SEQUENCE-BASED FEATURES RELEVANT FOR CLASSIFICATION**

During our testing of several features to be integrated in CLIPS-4D, the following sequence-based features turned out to improve classification quality.

**Conservation of a residue-position** We chose JSD(k) as it performs better than other conservation measures.

\[
JSD(k) = \frac{H(f_{\text{obs}}^k) - H(f_{\text{back}}^k)}{2} = \frac{1}{2} H(f_{\text{obs}}^k) - \frac{1}{2} H(f_{\text{back}}^k) \tag{1}
\]

\( f_{\text{obs}}^k \) is the probability mass function for residue position \( k \) approximated as \( f_{\text{obs}}^k(aa_l) = f_l(aa_l) \) by the amino acid frequencies observed in the respective column \( k \) of the MSA; the mean amino acid frequencies deduced from the SwissProt database served as background frequencies \( f_{\text{back}}^k \). \( H(\cdot) \) is Shannon’s entropy, see below. For classification, we used a z-score; see [1].

**Conservation of a sequence neighborhood** For functional sites, sequence neighborhood is conserved as well. Therefore we determined \( \text{cons}_{\text{loc}}(k) \):

\[
\text{cons}_{\text{loc}}(k) = \frac{1}{|\text{Neib}|} \sum_{l \in \text{Neib}} w_l \text{cons}_{\text{loc}}(k+l) \tag{2}
\]

\( \text{Neib} = \{-3,-2,-1,1,2,3\} \) determined the set of neighboring positions. The weights were: \( w_{-3} = w_{+3} = 3, w_{-2} = w_{+2} = 2, w_{-1} = w_{+1} = 1 \). Note that conservation of position \( k \) did not contribute to \( \text{cons}_{\text{loc}}(k) \).

**Propensities of catalytic sites, ligand-binding sites, and positions important for structure** The distribution of amino acid residues at functionally important sites is skewed. Thus, three scores named \( \text{abund}_{\text{loc}}(k,\text{CLASS}) \) were computed:

\[
\text{abund}_{\text{loc}}(k,\text{CLASS}) = \sum_{l=1}^{20} f_l(aa_l) \log \frac{f_{\text{CLASS}}^{\text{loc}}(aa_l)}{f_{\text{back}}^{\text{loc}}(aa_l)} \tag{3}
\]

\( f_{\text{back}}^{\text{loc}}(aa_l) \) were the above background frequencies and \( f_{\text{CLASS}}^{\text{loc}}(aa_l) \) were the frequencies of residues from one set \( \text{CLASS} \in \{\text{CAT sites}, \text{LIG sites}, \text{STRUC sites}\} \).

**Scoring propensities of a neighborhood** The orchestration of a class-specific neighborhood of a site \( k \) can be assessed by

\[
\text{abund}_{\text{neib}}(aa_i^k,\text{CLASS}) = \frac{1}{|\text{Neib}|} \sum_{l \in \text{Neib}} \sum_{i=1}^{20} f_{l,i}(aa_i) \log \frac{f_{\text{CLASS}}^{\text{neib}}(aa_i)}{f_{\text{back}}^{\text{neib}}(aa_i)} \tag{4}
\]

Here, \( aa_i^k \) is the amino acid \( aa_i \) occurring at site \( k \) under consideration, \( f_{l,i}(aa_i) \) is the frequency of \( aa_i \) at position \( l \) relative to \( k \) and \( f_{\text{CLASS}}^{\text{neib}}(aa_i) \) is the conditional frequency of \( aa_i \) at the same positional offset deduced from the neighborhood of all residues \( aa_i \) of a set \( \text{CLASS} \). \( \text{Neib} \) is the \( \pm 3 \) neighborhood.

**STRUCTURE-BASED FEATURES RELEVANT FOR CLASSIFICATION**

What distinguishes the localization of functionally important residues and those of the enzyme’s core? Functionally important residues are in contact with the substrate to be processed, whereas the most often hydrophobic residues, which are important for structure, are buried in the protein’s core. Thus, it is useful to determine the solvent accessibility of the residues as a further feature. Moreover, it is known that functionally important residues are located in pockets, which proposes a second 3D feature.

**Computing the relative solvent-accessible surface area** The relative SASA (\( r\text{SASA} \)) was deduced from the protein 3D structure for each residue \( aa_i \):

\[
r\text{SASA}(aa_i) = \frac{\text{SASA}(aa_i)}{\text{SASA}_{\text{max}}(aa_i)} \tag{5}
\]

Here, \( \text{SASA}_{\text{max}}(aa_i) \) is the maximally possible SASA of the amino acid.

**Assessing pockets** pocket is one of the best methods for the identification of pockets in proteins. It scores cavities of the protein surface based on a Voronoi tessellation and alpha spheres. We determined a normalized score to compensate for the protein-specific number of pockets:

\[
n\text{Pocket}(aa_i) = \frac{\max(\text{PocketScore})}{\text{PocketScore}(aa_i)} \tag{6}
\]

\( \max(\text{PocketScore}) \) is the largest score deduced for any pocket of the considered protein and \( \text{PocketScore}(aa_i) \) is the score of the pocket in which \( aa_i \) is allocated. We assigned a score of -1 to all residues which did not belong to pockets or whose \( r\text{SASA} \) value was less than 4%.

In our hands, these seven sequence- and two structure-based features gave an optimal combination for the classification of residue-positions into four \( \text{CLASS}\)es.

**CLIPS-4D: CLASSIFYING BY MEANS OF A MC-SVM**

For a classification, these features have to be combined to come to a decision. Machine learning, which is a specific branch of computer science, offer several algorithms that can be trained in an initial phase and used afterwards to classify hitherto unseen cases. For example, the weights of a neural network can be fixed by means of the back propagation algorithm. However, neural networks suffer from the problem of being trapped in local minima, which
can degrade classification performance. A very robust algorithm to be used for a binary classification is a support vector machine (SVM). Again, positive (P) and negative (N) cases have to be presented during a training phase. Then, the SVM “learns” the characteristic combinations of features that distinguish P and N cases. After training, the SVM has found a hyperplane that optimally separates the two classes with respect to criteria from learning theory. Usually, each classification is further characterized by means of a p-value.

When designing CLIPS-4D, it was our aim to distinguish four ClassEs of residues with respect to their function. Thus, we implemented and trained a multiclass support vector machine (MC-SVM) that classifies residue-positions based on the above introduced seven sequence-based and the two structure-based features.

How is the performance of this program? For a dataset containing 840 CAT_sites, 4466 LIG_sites, 3703 STRUC_sites and 19 223 NOANN_sites, we determined the MCC-value, which is considered a fair measure of classification performance. This correlation coefficient takes values in the range of [+1, -1] and is zero, if the classification is completely random. Table 1 lists the results for CLIPS-4D and of ConSurf, which is an alternative classifier.

**TABLE 1 MCC-VALUES OF CLASSIFIERS FOR CRUCIAL RESIDUE-POSITIONS**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>CAT_sites</th>
<th>LIG_sites</th>
<th>STRUC_sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. 3D-Features only</td>
<td>II. 0.31</td>
<td>III. 0.22</td>
<td>IV. 0.43</td>
</tr>
<tr>
<td>V. 1D-Features only</td>
<td>VI. 0.34</td>
<td>VII. 0.12</td>
<td>VIII. 0.67</td>
</tr>
<tr>
<td>IX. CLIPS-4D</td>
<td>X. 0.43</td>
<td>XI. 0.27</td>
<td>XII. 0.68</td>
</tr>
<tr>
<td>X. ConSurf</td>
<td>XIV. 0.30</td>
<td>XV. 0.46</td>
<td></td>
</tr>
</tbody>
</table>

In all cases, MCC-values for the classification of CAT_sites, LIG_sites, and STRUC_sites are listed. The lines 3D-Features only and 1D-Features only list the performance, if classification is exclusively based on these features. The alternative classifier ConSurf does not distinguish catalytic and ligand-binding sites. Therefore, we merged the sets CAT_sites and LIG_sites prior to classification. For details see [1].

What can the user expect for an individual case? We analyzed the aminodeoxychorismate synthase (PabB) which belongs to a heterodimeric complex. For the structure of PabB (E. coli, 453 residues) and a respective MSA, functional residues were predicted and compared to the PDBsum entry. 20 residues are involved in catalysis or ligand binding. 11 were predicted as functionally important, but also 38 false positives (FP). CLIPS-4D predicted 12 CAT_sites and 37 LIG_sites. The CAT_sites contained the 2 known catalytic residues. Among the 10 FP predictions were 3 ligand binding sites and 2 residues belonging to the protein-protein interface. The 37 predicted LIG sites contained 6 of the known ligand-binding sites, 7 were not detected. The role of the 68 STRUC_sites is unknown, 2 are binding the ligand.

**A FURTHER SIGNAL ELUCIDATING IMPORTANT RESIDUES: CORRELATION PATTERNS**

The features exploited by CLIPS-4D assess the conservation of residues or the abundance of pairs but do not quantify signals caused by dependencies in the occupancy of two or more positions. Again, these patterns are due to the physical/chemical properties of the residues, see Figure 2.

For instance, the occurrence of a specific amino acid at a given position of a protein may crucially depend on the local environment, which can impose restrictions with respect to the size or chemical properties of the neighboring residues. Thus, an amino acid replacement at one position is tolerated only together with a complementary residue substitution at a correlated site. As a consequence, the frequencies of particular residues at adjacent positions in the structure of a protein can be interrelated. This coupling can be detected by a column-wise correlation analysis of the MSA based on information theory; see [3] and references therein.

Frequently used concepts for the assessment of these signals are based on Shannon’s information theory; e. g., one can compute for each pair of residue-positions k, l the term \( U(k, l) \) according to

\[
U(k, l) = 2 \frac{H(k) + H(l) - H(k, l)}{H(k) + H(l)} \quad (7)
\]

Here, \( H(k) \) is the entropy of an individual column k

\[
H(k) = -\sum_{i=1}^{20} p(a_i^k) \ln p(a_i^k) \quad (8)
\]

and \( p(a_i^k) \) is the probability of amino acid \( a_i \) at position \( k \). The entropy \( H(k, l) \) of two residue-positions \( k \) and \( l \) is

\[
H(k, l) = -\sum_{i,j} p(a_i^k, a_j^l) \ln p(a_i^k, a_j^l) \quad (9)
\]

and \( p(a_i^k, a_j^l) \) is the probability of the amino acid pair \( (a_i, a_j) \) at positions \( k \) and \( l \). Previously, we have shown that \( U(k, l) \) values allow the identification of further, functionally important residue-positions [4].

![Figure 2. Patterns to be assessed by means of a correlation analysis. Size and shades of the circles represent different properties of the residues. For example, the replacement of a large residue with a small one may cause a cavity that has to be filled by a subsequent mutation in the direct 3D neighborhood. This situation is illustrated for pos \( k \) and \( l \) belonging to the same protein A. Such compensatorial effects may also occur in proteins complexes, as illustrated for pos \( r \) and pos \( s \) that belong to different proteins.](image-url)
AN IMPROVED METHOD FOR CORRELATION ANALYSIS: H2rs

Although successful in many cases, the concept of Shannon’s entropy has one severe limitation. It does not consider the similarity of the objects to be compared. Therefore, we opted for the von Neumann entropy (vNE) which is a generalization of the classical Shannon entropy and introduced a novel $U_{vNE}(k, l)$ term to replace $U(k, l)$.

The core concept of the vNE is the utilization of a so-called density matrix $\rho_{ij}$ that is a positive definite matrix whose trace (the sum of the diagonal elements) equals to 1. $\rho_{ij}$ can be computed for each pair $k, l$ according to:

$$\rho_{ij} = P_{ij} A P_{ij}$$

(10)

Here, $P_{ij} = \text{diag}(\sqrt{p_1}, \ldots, \sqrt{p_{400}})$ and $p_1 \ldots p_{400}$ are the pairwise amino acid probabilities $p(a_i^k, a_i^l)$ specified in Formula (9). These probabilities satisfy the normalization condition $\sum_{i=1}^{400} p_i = 1$. $A$ is a $400 \times 400$ matrix that assesses the similarity of residue pairs and it is this matrix that allows us to model substitutions more precisely. If $A$ is equal to the identity matrix, then the vNE is equal to the Shannon entropy, that is, $\text{vNE}(k, l) = H(k, l)$. Based on $\rho_{ij}$, the von Neumann entropy $\text{vNE}(k, l)$ can be calculated as

$$\text{vNE}(k, l) = \text{vNE}(\rho_{ij}) = -\sum_{i=1}^{400} \lambda_i \log \lambda_i$$

(11)

by means of the eigenvalues $\lambda_i$ of $\rho_{ij}$. H2rs used an implementation of this method to characterize individual residue-positions.

The analysis of a large benchmark dataset confirmed that specificity and precision were higher for H2rs than for the predecessor program and two other methods of correlation analysis; for details see [3].

REFERENCES


See also the references cited in [1] to [4].
**Abstract**—Breast cancer is known to be a heterogeneous group of a number of disease subtypes, which show distinct variances in biological and clinical behavior and in response to systemic therapies. These subtypes, which are distinguished according to their molecular differentiation, are named: luminal A, luminal B, normal breast-like, basal-like and HER2. In this study, systems biology techniques seek to add a new level of understanding of one of these groups—the basal like breast cancer group—by viewing it at a systems level, through analysis of gene expression micro-array data using Artificial Neural Network (ANN) based data mining and network inference techniques. This contributes towards a greater understanding of the biology of these subtype systems, and provides information that will lead to a better understanding of this disease system.

**INTRODUCTION**

Basal-like breast cancer (BLBC) has attracted much attention in recent years ([1], [2]). This may be because of its clinically aggressive nature and its association with poor prognosis and outcomes ([1], [3]). BLBC is so-called because its cancerous cells express genes usually present in normal basal/myoepithelial cells of the breast; such as cytokeratins (CK 5/6 and 17), and the epidermal growth factor receptor (EGFR) ([4], [5]). Importantly, it is usually characterized by negative ER, PR and HER2 based on immunohistochemistry analysis; therefore, it is sometimes clinically defined as a triple-negative phenotype and would not be expected to respond to antiestrogen hormonal therapies or to trastuzumab ([1], [4]).

A number of challenges exist in the molecular definition of BLBC. Previous classifications such as that of Perou [5] have been largely based on ordination techniques such as hierarchical clustering or linear parametric approaches such as Classification And Regression Trees (CART). Such approaches are not well suited to the classification of biological systems because of their inherent linear nature and do not perform well at classifying biological groups because of the inability to cope with the high dimensionality of the data sets being analyzed. Furthermore these approaches do not explore the inherent system being modeled but merely define a list of classifier markers.

**Rationale behind approach**

The approach adopted here uses a constrained architecture early stopping on test data and a sigmoidal transfer function. This is coupled with a Monte Carlo Cross Validation (MCCV) used in the ranking of the classification ability of the probe-set and for data mining out key classifiers. Some of the key issues around data mining gene expression data are the risk of false discovery, lack of generalization of solutions and selection of markers with poor biological representation due to the use of linear functions. The approach adopted here seeks to overcome these issues; constrained architecture and early stopping will prevent over fitting. MCCV will produce generalized solutions and sigmoidal transfer functions will facilitate good biological representation.

**Aims**

The classification of BLBC at the molecular level was one of the major steps forward in breast cancer studies, and that was based on a gene expression profiling technique combined with hierarchical cluster analysis as one of the systems biology techniques. Because the latter technique is considered to be a linear approach, herein, the study aims to apply ANNs as a non-linear systems biology approach in BLBC.

**METHODOLOGICAL APPROACH**

**Data source**

Data from a previous microarray experiment has been downloaded from the ArrayExpress website (http://www.ebi.ac.uk/arrayexpress/), which is a respiratory tool for gene expression arrays; related to the European Bioinformatics Institute. The ArrayExpress ID for the dataset is E-GEOD-21653. This dataset was originally generated in a study published by [6] using whole-genome DNA microarrays (HG-U133 plus 2.0, Affymetrix) (n=2045). These data contained a definition of the molecular subtype based on the Perou Classification scheme [5]. From these classifications the data were divided into basal and non-basal classes.

**Algorithm and architecture**

The ANN modeling undertaken used a supervised learning approach applied to a three-layer multi-layer perceptron architecture [7]. The initial weights matrix was randomized with a standard deviation of 0.1 to reduce the risk of over fitting the data. The ANN architecture was initially constrained to two hidden nodes in the hidden layer also for this reason. Hidden nodes and the output node incorporated a sigmoidal transfer function. During training weights were updated by a feed forward back propagation algorithm. Learning rate and momentum were set at 0.1 and 0.5 respectively. The output node was coded as 0 if the sample was non-basal, and 1 if it was basal.

**Monte Carlo Cross validation**

This was achieved using a stochastic data selection approach incorporating Monte Carlo cross validation. Prior to ANN training, the data was randomly divided into three subsets; 60% for training, 20% for testing (to assess model performance during the training process) and 20% for

---

* Corresponding Author (e-mail: graham.balls@ntu.ac.uk).

Authors One, Two and Three are all from John van Geest Cancer Research Centre, Nottingham Trent University, Clifton Lane, Nottingham. NG11 8NS. United Kingdom.
validation (to independently test the model on data completely blind to the model). This Monte Carlo cross validation also contributed to the reduction of over-fitting to the data.

Parameter selection for Network inference

The normalized intensity of each gene was used as an individual input in the ANN model, creating 54,164 individual models, which was the number of genes in the data source. The 2045 cases in the data were then split into three subsets (described above) and trained. This random resampling and training process was repeated 50 times to generate predictions and associated error values for each sample with respect to the validation (blind) data. Inputs were ranked in ascending order based on predictive error and the gene that performed with the lowest error was selected for further training. This process was repeated 10 times and the results aggregated.

Network inference

The network inference algorithm is an ANN based approach, which was initially developed and presented by Lemetre et al [8]. The purpose of this method was to identify the nature of interaction between a subset of genes associated with a particular subclass of disease, signifying the potential biological interaction network for this subclass. The approach identifies the key drivers of a given system by determining the magnitude of a probes influence on other probes in the defined set.

The main concept of this approach is based on the iterative calculation of the influence that a number of genes (i.e. markers) may have on a single gene. In other words, the algorithm investigates how the level of expression of every single gene, as the output for the ANN modeling, can be inferred using all the other genes in the original dataset as the inputs, and repeating the process for all of them [8].

The Interaction algorithm utilizes the ANN algorithm described above in terms of the ANN architecture and its parameter settings. However in this instance the algorithm Output is the expression level of another gene within a set defined by parameterization. Thus genes are used to predict genes.

At each step of the model’s development, the generated weights of the trained ANN, with the lowest predictive error that is associated with the gene in the input vector and the output gene, is used in a scoring algorithm to derive a score indicating the intensity of the interaction between these two genes. Therefore, the final output from the interaction algorithm is a Microsoft Excel spreadsheet containing the interaction score between each pair of genes. That is, between each gene as target and the other genes in the dataset as sources and vice versa, signifying an interaction map of this gene’s dataset. These scores could be negative or positive, suggesting inhibition or stimulation respectively. The analysis could be set to be repeated independently over ten times from which the average interaction scores can be calculated [8]. This analysis created a matrix of potential gene to gene interactions comprising 9900 ((100x100)-100) interactions. This matrix was sorted and the top 100 strongest interactions were selected for visualization in Cytoscape (www.cytoscape.org).

RESULTS

Parameter selection for Network inference

Initial data mining to determine the transcriptomic expression probes identified a subset of probes that were able singularly to classify BLBC. The optimal subset of BLBC predictive markers was identified from a rank order of single probes based on the MSE of the classifier for test data (20% n cases, Figure 1) and the top 10 based on Mean Squared Error (MSE) are listed in Table 1. The distribution of probe performances indicated that there were a small subset of probes relative to the number in the total probe set, that were associated with BLBC. Key genes found are evaluated in the discussion. A number of repeated probes from the same genes were found in the top 100 reinforcing the importance of those genes

Network Inference

Within the network inference study, one major hub (ERBB4) and two minor hubs were identified (TTC6 and Hs.658745) (Figure 2). Hubs are defined as probes having strong connectivity with a large number of other probes. Such probes are likely to be highly influential in a given system given their strong connectivity.

FIGURE 1 DISTRIBUTION OF GENE PROBE PERFORMANCES AND SELECTION BASIS FOR TOP 100 GENE PROBES.

DISCUSSION

The aims of this study were, through the use of ANN based data mining and network inference, to model molecular elements of the BLBC system. This both identified key molecular classifiers of the system and identified the most influential molecular drivers.

The results reveal that the first best transcript that can distinguish BLBC from other types with a median accuracy of almost 93 % and an error of 0.037 has not been annotated yet, as shown in Table II.
The present study has investigated the systems of breast cancer molecular subtypes in order to identify the most influential genes by modeling the gene interaction network for these systems, and identifying the most predictive genes for each disease using Stepwise and Interaction analyses as non-linear systems biology approaches to ANNs.

On the other hand, the second transcript, which is derived from the MLPH gene, has a basal-like prediction accuracy of almost 94% with a slightly higher error (0.038). This gene encodes a member of the exophilin subfamily of Rab effector proteins [9]. In fact, there is no direct association reported between this gene and breast cancer in the literature. However, it has been associated with some other types of cancer such as lung cancer and prostate cancer ([10], [9]).

In addition to the MLPH, the results have shown that the transcript of the FOXA1 gene has a median accuracy of nearly 93% when discriminating between BLBC and other breast cancer subtypes. FOXA1 is a member the fox family of transcription factors, and they are expressed in the breast, liver, pancreas, bladder, prostate, colon and lung where they can bind to the promoters of more than a hundred genes, and serve as regulators of cell signaling and the cell cycle ([11],[12]). In accordance with the current study’s findings, it has been revealed in a recent study that FOXA1 works as a repressor to the molecular phenotype of basal breast cancer cells [13]. In addition, it has been reported that FOXA1 is inversely correlated with basal-like phenotypes, which includes the positive expression of CK5/6 and CK14 [14]. Furthermore, FOXA1 was found in association with the expression of the AR gene in triple negative tumors, signifying a good prognostic effect in these tumors (Rakha, et al. 2007). This is in agreement with the current study’s findings since the BLBC is considered to be one of these cancer types, as mentioned previously. In addition, the results show that the AR gene is one of the BLBC biomarkers, as demonstrated in Table 1. In general, the expression of FOXA1 has been reported in many cancer types in addition to breast cancer, such as lung and prostate cancer [12].

### TABLE 1: THE TOP 100 GENE PROBES IDENTIFIED BY THE ALGORITHM FED INTO NETWORK INFERENCE (REPEAT PROBES OMITTED)

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene symbol</th>
<th>Gene description</th>
<th>Median % accuracy (Test Data)</th>
<th>MSE (Test Data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>229150_at</td>
<td>Hs.659093</td>
<td>Not identified</td>
<td>92.6%</td>
<td>0.037</td>
</tr>
<tr>
<td>218211_s_at</td>
<td>MLPH</td>
<td>Melanophilin</td>
<td>93.5%</td>
<td>0.038</td>
</tr>
<tr>
<td>237086_at</td>
<td>FOXA1</td>
<td>Forkhead box A1</td>
<td>92.8%</td>
<td>0.039</td>
</tr>
<tr>
<td>220192_s_at</td>
<td>SPDEF</td>
<td>SAM pointed domain containing ETS transcription factor</td>
<td>92.2%</td>
<td>0.042</td>
</tr>
<tr>
<td>204667_at</td>
<td>FOXA1</td>
<td>Forkhead box A1</td>
<td>91.3%</td>
<td>0.044</td>
</tr>
<tr>
<td>228241_at</td>
<td>AGR3</td>
<td>Anterior gradient 3 homolog (Xenopus laevis)</td>
<td>89.8%</td>
<td>0.044</td>
</tr>
<tr>
<td>211621_at</td>
<td>AR</td>
<td>Androgen receptor</td>
<td>90.0%</td>
<td>0.045</td>
</tr>
<tr>
<td>214404_s_at</td>
<td>SPDEF</td>
<td>SAM pointed domain containing ETS transcription factor</td>
<td>90.9%</td>
<td>0.046</td>
</tr>
<tr>
<td>228989_at</td>
<td>AGR2</td>
<td>Anterior gradient 2 homolog (Xenopus laevis)</td>
<td>89.7%</td>
<td>0.047</td>
</tr>
<tr>
<td>233888_at</td>
<td>AK022350</td>
<td>Not identified</td>
<td>85.5%</td>
<td>0.048</td>
</tr>
</tbody>
</table>

### TABLE 2: THE TOP TEN BASAL-LIKE PREDICTIVE GENES WITH DESCRIPTION

<table>
<thead>
<tr>
<th>ADAMTS15</th>
<th>DACH1</th>
<th>GATA3</th>
<th>MSX2</th>
<th>TFF3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AGR2</th>
<th>DEGS2</th>
<th>GFRAl</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AGR3</td>
<td>DNAU1</td>
<td></td>
<td>Hs.161354</td>
<td>PGR</td>
</tr>
<tr>
<td>ANKRD30</td>
<td>A</td>
<td></td>
<td>Hs.596643</td>
<td>PRR15</td>
</tr>
<tr>
<td>ANX9</td>
<td>ENSG00000235687</td>
<td>Scube2</td>
<td>WFS1</td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>ENSG00000260932</td>
<td>SID1</td>
<td>XP1</td>
<td></td>
</tr>
<tr>
<td>CA12</td>
<td>ENSG00000261179</td>
<td>INP4B</td>
<td>SLC4A4</td>
<td>ZNF703</td>
</tr>
<tr>
<td>CAPNB</td>
<td>ERBB4</td>
<td></td>
<td>KIAA1324</td>
<td>SLC7A8</td>
</tr>
<tr>
<td>CLSTN2</td>
<td>ESR1</td>
<td></td>
<td>KRT16</td>
<td>SPODEF</td>
</tr>
<tr>
<td>CMBL</td>
<td>FAM1748</td>
<td></td>
<td>LOC10017093</td>
<td></td>
</tr>
<tr>
<td>CXXC5</td>
<td>FOXA1</td>
<td></td>
<td>MAGED2</td>
<td>TBC1D9</td>
</tr>
<tr>
<td>CYP2B6</td>
<td>GALNT7</td>
<td></td>
<td>MLPH</td>
<td>TFF1</td>
</tr>
</tbody>
</table>

FIGURE 2: THE BLBC GENE INTERACTION MAP
Importantly, the study has modeled the interaction gene network of the BLBC system, as illustrated in Figure 2. The gene interaction map has revealed that the most interesting genes that drive the BLBC system are ERBB4 and TTC6, in addition to un-annotated locus, as illustrated in Figure 2 by the red colored nodes. The results indicate that the expression of the ERBB4 gene is inhibited by many genes in the BLBC system, including the identified predictor genes, FOXA1 and MLPH. The ERBB4 gene described as V-erb-a erythroblastic leukemia viral oncogene homolog 4, encodes a member of the tyrosine kinase receptor family, which is also called the epidermal growth factor receptor family; it contains four receptors: ERBB1, ERBB2, ERBB3, and ERBB4.

In agreement with the results, the ERBB4 gene has been associated with breast cancer; however, the significance of its biology in breast cancer remains poorly understood across the literature. In other words, many studies have examined the impact of ERBB4 on breast cancer, but the results have been inconsistent, signifying that ERBB4 is oncogenic in addition to the function of tumor suppressing [15]. The current results support the first findings- that is ERBB4 may be an oncogene. To be more precise, it has been revealed that a mutation within the ERBB4 kinase domain might contribute towards the development of breast cancer [16]. In addition, a study by Kim et al [17] indicates that a SNP (rs13393577) at chromosome 2q34, located in the ERBB4 gene is associated with breast cancer risk. Many studies on breast cancer have revealed that the expression of ERBB4 is generally low in ER-negative breast cancer, and indicated that the down-regulation of ERBB4 expression is one of the BLBC’s characteristics [19].

Furthermore, the Tetratricopeptide repeat domain 6 gene (TTC6) has been shown by the current study to be one of the most influential genes in the BLBC. Similar to the case with the ERBB4 gene, the expression of the TTC6 gene is inhibited by a group of genes according to the results (Figure 2). However, there is no direct association reported between this gene and cancer in general, or breast cancer in particular, in the literature. In addition, very little information has been discovered in the literature on the nature of this gene.

**Study limitations**

As is the case with any study or analysis, this study has some limitations. For example, specifying the influential genes in a system for any of the breast cancer subtypes has been based on individual observations of the gene interaction map. On the other hand, if a computational approach is used to figure out the number of edges (i.e. interaction) linked to each gene, the genes with more intensity in interaction will be specified more accurately.

Despite this, the approach seems adequate to give an overview of the biology of each system, although studying a higher number of these genes would certainly provide more insight. In addition, for validating the results of gene interactions, the study has relied on thorough searching of the published literature using ordinary databases; however, mapping the results to one of the curated databases such as STRING or Ingenuity Pathway Analysis, would provide a more precise and powerful understanding of the gene of interest and its reported association with cancer if present.

Another limitation is associated with the origin of the dataset that the study was based on. Although the microarray dataset was chosen carefully, it was still difficult to ensure the accuracy and quality of the gene expression measurements, which no doubt has affected the study’s results. Therefore, it is suggested that separate microarray experiments should be conducted to generate the dataset needed for further study, if possible.

**ACKNOWLEDGMENTS**

The authors would like to thank Sabatier, et al for providing the data sets used in this study. They would also like to thank the John and Lucille van Geest foundation for its financial support.

**REFERENCES**

14q13 is amplified and overexpressed in esophageal and lung adenocarcinomas. Cancer Res 62:5273-5279


Nanotechnology Applications in Genetics

Mehmet Tevfik DORAK*

Abstract - The developments in nanotechnology has revolutionized genetics research. Biochips, integrated microfluidic devices, nanochannels and nanoparticles have had impacts on how PCR and sequencing are performed and provided speedy, low cost, highly sensitive and specific methods for genotyping, gene expression analysis, sequencing and epigenetics analysis.

INTRODUCTION

Nanotechnology uses devices and materials in nanoscales (1 nm = 10⁻⁹ m). Despite being a very young discipline, it has had a major impact on many areas including chemistry, biomedical engineering, medicine and biology.

Nanotechnology has also revolutionized the way genetic analyses are performed. These range from simple genotyping to epigenomic analyses. In this paper, the focus is already commercialized microfluidic devices used for genotyping and nanochannels/nanopores used for third generation sequencing. DNA nanotechnology that exploits the self-assembly ability of DNA molecules for material use of DNA, synthetic biology and extreme genetic engineering are not covered in this review.

DEFINITIONS

Biochips consist of immobilized biomolecules spatially addressed on planar surfaces, fixed in micro- or nano-channels or wells. Thanks to nanotechnology, a single chip may possess more than a million features with supersensitive detection abilities compared with conventional methods. While microarrays or microchips have been in use for some time, nanochips are gradually replacing them as predicted [1,2]. Microchips made genome-wide association studies (GWAS) possible, and nanochips are now in use for third-generation genome sequencing projects.

Microfluidics emerged in the beginning of the 1980s and aims to design mechanical flow control devices like pumps and valves. Microfluidics deals with the behavior, precise control and manipulation of fluids that are geometrically constrained to a sub-millimeter scale. Microfluidic devices are a result of advanced fabrication technologies that created the micro-/nano-features to reduce the volume of working solution of biological experiments to nanoliters (10⁻⁹ L) or femtoliters (10⁻¹⁵ L). Various temperature cycling systems, thermal cycling times, flow-rates, and cross-sectional areas have been integrated into microfluidic chips to reduce the amplification time. Isothermal amplification techniques such as loop-mediated isothermal amplification (LAMP) have also emerged. Isothermal amplification techniques have several advantages over conventional PCR such as (i) moderate incubation temperature, (ii) higher amount of amplification products, (iii) direct genetic amplification due to the superior tolerance to substances that typically inhibit PCR, (iv) higher specificity and sensitivity, and (v) rapid amplification (usually 10 to 20 min) [3]. LAMP is isothermal, fast, specific, sensitive and robust so that it does not require the DNA template to be purified or denaturized.

Microdevices that integrate multiple laboratory functions on a single chip of no more than a few square centimeters in size are called "lab-on-a-chip" (LOC). LOCs handle small fluid volumes down to less than pico liters. LOCs are a subset of micro-electro-mechanical systems (MEMS) devices also called "micro total analysis systems" (µTAS) [2]. A highly developed LOC may include functional components for solid-phase DNA extraction, PCR, capillary electrophoresis, analysis and readout which allow a complete sample-in-answer-out functionality [1,4]. There are already commercialized examples of LOCs in use (Table 1).

In a miniaturized device consisting of a large number of micro- or nano-wells, cross-contamination becomes a real threat. Microfabrication techniques and surface chemistry (surfactants) have been combined to prevent contamination within the microdevice. These microfluidic chip-based nanolabs use a single template molecule in a 10 nL reaction volume and produce the same concentration of amplicons as 1000 copies of starting template achieve in a 10 μL volume. Integrated microfluidic systems or LOCs have several advantages over conventional methods including high speed, high throughput, high sensitivity, low reagent consumption, reduction of instrument size and contamination probabilities. In a LOC, the analytical process is integrated in a single miniaturized device with efficient connections among functional components so that the loss and dilution of samples is minimized [5].

APPLICATIONS

The advent of LOCs has revolutionized genetics research. The main contributions have been to the areas of DNA sequencing, gene expression analysis, pathogen detection, and forensic analysis (short tandem repeat (STR) typing). More recently epigenetic analysis has also become possible on integrated microfluidic devices.

DNA Sequencing

A microfabricated platform integrates sample preparation with electrophoresis to achieve rapid and low-cost DNA
sequencing from even a single copy of DNA template. The first such device that integrated thermal cycling, sample purification and capillary electrophoresis was produced in 2006. Using improved Sanger chemistry, the first DNA sequencing LOC was able to do DNA sequencing from only 1 femtomole of DNA template in less than 30 min with an average 556-base read-length and 99% accuracy. Since then, Sanger sequencing has been achieved using 100 attomol of a 624-bp product in a microfluidic device.

Another nanoprocedure for DNA sequencing is based on DNA stretching in nanochannels, which was first developed as a method called “optical mapping” by which the stretched and stained single DNA molecules are photographed using a fluorescence microscope. In nanopore sequencing, DNA is detected as it passes through a nanometer-sized pore which mimics the functions of natural ion channels. Each nucleotide passing through the nanopore may also be detected by distinctive changes in ionic current. Nanopore sequencing offers sequence lengths of up to tens of kilobases, minimal sample preparation and high sequencing pace at low cost [6]. One of the biologic nanopores is the engineered Mycobacterium smegmatis porin A, MspA, which can distinguish DNA nucleotides and resolve single-nucleotides in single-stranded DNA [7]. DNA sequencing by the use of nanopores has been commercialized (Table 1).

Genotyping
Genotyping requires both PCR and capillary electrophoresis in a microdevice, which have now been perfected. Besides integrated microfluidic devices used in genotyping, including pathogen detection and STR typing as discussed below, special nanoparticles are also used to enhance the efficiency and specificity of PCR. The mechanism of the improvements in nanomaterial-assisted PCR (nanoPCR) is not yet known [8], but both gold nanoparticles (AuNPs) and carbon nanopowders (CNPs) exert drastic effects on PCR performance. It is likely that these nanoparticles show their effects by adsorbing polymerase and modulating the amount of active polymerase in PCR, adsorbing primers and increasing the melting temperature difference between matched and mismatched primers during the annealing step, and adsorbing amplicons to facilitate their dissociation in the denaturation step. It appears that it is the surface interaction of nanoparticles with PCR ingredients that make the difference. So far, nanoparticles have been used like PCR enhancers for improved efficiency (like DMSO, betaine, glycerol, formamide, and single-stranded DNA-binding proteins in conventional PCR) or PCR additives for specificity improvement (like betaine, formamide, single-stranded DNA-binding proteins or tetramethylammonium chloride). They also modify the performance of PCR as happens in hot-start or touch-down PCR. The overall efficiency of nanoPCR results in not only more efficient and specific genotyping, but also allows long PCR and haplotyping [9].

Solid phase amplification is another approach in which two PCR primers attached to the glass surface as used in microarrays. The template DNA hybridizes to the surface-bound primers and the primers are elongated with the DNA polymerase as in solution-based PCR. This design provides a major advantage of avoiding interference by primer dimers as they do not come into physical contact with each other. The design also allows multiplexing.

Pathogen detection
Microchip technology has made point-of care analysis for pathogen detection with results available within one hour and a sensitivity level of a single copy of the pathogen. In 2006, a ‘sample in-answer-out’ type integrated microfluidic genetic analysis system was developed capable of carrying out DNA extraction, PCR amplification and electrophoretic separation [4]. The successful analyses of Bacillus anthracis (anthrax) in less than one microl of whole blood and of Bordetella pertussis in 1 mL of nasal aspirate was possible in less than 24 minutes.

Forensic analysis
Forensic tests on samples from crime scenes are particularly challenging due to compromised samples, including degraded DNA, low copy number DNA and mixtures. Since STR typing is preferred over SNP typing in forensic analysis, and STR typing is a more demanding process than SNP typing, this adds a further level of complexity. In 2007, the first microdevice for STR typing integrating PCR and capillary electrophoresis along with a portable analysis instrument containing all the electronics and optics for chip operation and four-color fluorescence detection were developed [10]. The same group improved the performance of the original device developing a portable integrated genetic analyzer for on-site forensic STR typing the next year. The improved device was able to 9-plex autosomal STR typing. This device had a 160 nL PCR chamber with a microfabricated PCR heater and a temperature sensor for on-chip PCR, a co-injector for the injection of a sizing standard, and a 7 cm long separation channel for electrophoresis. The analysis was complete in 2.5 hours and can be done at the crime scene. The latest fully integrated microfluidic chip for human identification by STR analysis includes a unique enzymatic liquid preparation of the DNA, microliter non-contact PCR and a polymer for high-resolution separation [11]. This sample-in-answer-out microchip completes the analysis in two hours.

Gene expression analysis
Quantitative analysis of gene expression on the single cell level is crucial to understand the true mechanisms behind cellular processes. Reverse transcription PCR (RT-PCR) is a valuable method suitable for single cell gene expression analysis and it has been converted to microfluidic formats. Carrying out single cell gene expression analysis on an fully integrated microsystem is now possible incorporating single cell capture, RT-PCR, post-PCR product capture,
inline injection and capillary electrophoresis separation with a sensitivity of 11 mRNA molecules.

Other applications
Nanotechnologic advances are also revolutionizing epigenetic analysis, microRNA detection and DNA damage detection mainly by using nanopore-based methods.

CONCLUSIONS
Nanotechnology is a discipline which brings together engineers, chemists, physicists, biologists and medical professionals. Only started to develop in the 1980s, nanotechnology has already made an amazing impact and has not yet reached its peak. The microarray era in genetics was amazing enough, and we are now in the breathtaking lab-on-a-chip era. The near future is full exciting new developments and will surely bring more awesome methods and products.

EXAMPLES OF COMMERCIALLY AVAILABLE NANOTECHNOLOGY PRODUCTS FOR APPLICATIONS IN GENETICS

<table>
<thead>
<tr>
<th>Product</th>
<th>Application</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold nanoparticles</td>
<td>Detection of pathogens; genotyping</td>
<td>Nanosphere</td>
</tr>
<tr>
<td>Nanopores</td>
<td>DNA sequencing</td>
<td>BioNano Genomics; Oxford Nanopore Technologies</td>
</tr>
<tr>
<td>Microfluidics</td>
<td>Genotyping</td>
<td>Fluidigm</td>
</tr>
<tr>
<td>LabChip</td>
<td>Sizing, quantification and quality control of DNA, RNA, proteins and cells (automated electrophoresis system)</td>
<td>Biorad (Experion)</td>
</tr>
<tr>
<td>LabChip</td>
<td>Sizing, quantification and quality control of DNA, RNA, proteins and cells (automated electrophoresis system)</td>
<td>Agilent Technologies (2100 Bioanalyzer)</td>
</tr>
</tbody>
</table>

REFERENCES
Bioinformatics for processing, interpreting, analyzing and validating metabolomics data

Masahiro Sugimoto

Abstract—To understand the complex molecular interactions inter and intra cells, various omics technologies have been developed for profiling of hundreds of molecules simultaneously. Genomics, transcriptomics, and proteomics are already well established techniques for analyzing genome, ribonucleic acids (RNA), and proteins. In addition to the measurement instruments, bioinformatics tools to analyzing these data has been intensively developed. Metabolomics is relatively recently emerged omics that can identify and quantify low weight organic molecules named metabolites. Currently, mass spectrometry (MS) is the dominantly used as a profiling instrument with separation systems prior to MS. Because the development of measurement instrument and protocols are still active, bioinformatics tools are not matured. This tutorial provide the overview of bioinformatics topics in metabolomics field, including data processing, interpretation, pathway analysis, and statistical test. Database and data standardization to share and reproduce the profiled data is also introduced.

INTRODUCTION
Metabolomics aims to conduct the simultaneous determination and quantitative analysis of intracellular metabolites. Since metabolomics is concerned with small molecules that are the substrates and products, of cellular activity, it allows to explore in a direct and immediate way the biological systems [1-4]. Therefore, metabolomics is playing an increasingly important role in systems biology, and widely used in many applications including microbiology, diagnostic biomarker discovery, food and beverage analysis, and drug discoveries [5-7].

Metabolomics deals with ‘big data’ like other omics. Therefore sophisticated bioinformatics technologies are essential for efficient and high-throughput analysis by eliminating unexpected bias and noise which leads successful exploring biologically significant findings. In this tutorial, bioinformatics topics in the metabolomics field, especially the issues of mass spectrometry (MS) data will be introduced. We also recommend the previous excellent review papers in this field [8-27].

DATA PROCESSING
Overview of data processing
Typical data processing procedures for MS data was well documented in the ref. [9], and a variety of software packages [28-33] are also available.

*Corresponding Author (e-mail: msugi@sfc.keio.ac.jp). Masahiro Sugimoto is with Advanced Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan, (e-mail: msugi@sfc.keio.ac.jp).
on operation systems and these applications. Therefore, we developed an original binary file format that facilitates rapid access to chromatographs or electropherograms and mass spectra, which works independent on vendor-provided systems [40]. However, our data format was optimized only to CE or LC with QMS or TOFMS. The development of more versatile format that allows rapid access and realizing compact file size are required.

Peak picking

Peak picking or feature detection is the core of whole processing of MS data. Typically, extracted chromatography or electropherograms (EIC or EIE) is firstly generated, i.e. 3-dimensional data (retention or migration times, m/z and intensity) are converted to piles of two-dimensional chromatography or electropherograms, by integrating data points within a specific range along the m/z axis (ion extraction or data binning).

Second, background (or baseline) will be eliminated to reduce false positive detection. Third, local maxima are found as peak top candidates, or a mathematical model, such as Gaussian curve, is fitted to a part of chromatographs or electropherograms. Peaks whose peak area or signal-to-noise (S/N) ratio is more than user-specified threshold are remained and used for the subsequent analyses [28-30, 40, 41]. As a mathematical model to describe peak shape, wavelet transformation and Gauss-curve (named matched filter) is a commonly used [29, 41]. However, actual peak shape frequently becomes more skewed and complex with unexpected noise and the current methods does not work well in many cases. As alternative ways, interactive optimizing parameters of data processing have been implemented [28, 40] or parameter-independent robust methods is also developed [42]. The development of robust and versatile algorithms are necessary.

Alignment to generate data matrix

High throughput hardware facilitate us a large-scale comparison. Nowadays, thousands of datasets were measured in a single study and these datasets are simultaneously compared. In comparison to GC/MS, alignment of LC-MS and especially CE-MS is quite difficult because of its non-linear retention time or migration time drift among multiple samples. Therefore, currently many alignment techniques are available [43]. Here we review the representative three methods.

Time correlation optimized warping

Time correlation optimized warping (COW) was the most conventional methods among variety of algorithms. Chromatograms are divided into small segments and shifts individual segments to maximize the correlation coefficient between a reference and test chromatograph. A larger number of segments leads to greater accuracy, but raises the risk of dividing the targeted metabolite peaks. The use of genetic algorithms to optimize the degree of segmentation has been proposed [44].

Parametric time warping

This method aligns a given chromatogram with a reference chromatogram using second degree polynomial functions, called warping functions [45]. Coefficients in warping functions are tuned to minimize the time difference between selected matched peaks in reference and aligned chromatograms. Therefore, the method relies on the presence of a number of known matched peaks among the samples to be aligned. The addition of internal standards (IS) is the simplest way. However, addition of ISs decrease the quantification quality of the obtained data.

Dynamic time warping

This methods consists of two steps. Firstly, it finds the matched peaks among multiple datasets automatically to produce warping functions. Dynamic programming (DP) has historically been used in homology searching of genes or genomes, and has been used for matching peaks [46]. The parameters that characterize DP results, such as gap penalty, make this method parametric. Thus, empirical reiterative multi-step optimization of these parameters has been used in CE-MS data processing software [30] and interactive graphical user interfacing [40]. In contrast, recent modifications to DTW using multiple chromatograms with different m/z, instead of one-dimensional information available from total ion chromatography, reduced the impact of the parametric problems embedded in the original DTW algorithm.

Calibration of mass values

Time-of-flight (TOF) type MS produce exact masses (Mass-to-charge ratio (m/z) values). Several factors influencing m/z shifts can change even during the course of a single run [47]. The abundance of ions, e.g. peak intensities itself also affect m/z value [48], using the location of background noise observed throughout the measurement [49], and using statistical approaches with multiple datasets [42]. These methods should be integrated to normalize m/z values.

Normalization of quantified values

The unexpected systematic bias should be eliminated for subsequent analyses to identify significant metabolites. For example creatinine is used to normalize overall urine metabolite concentrations [50]. However, this method is not always sufficient to eliminate systematic bias, and a recent mouse metabolomic study revealed a correlation between overall urinary metabolites and several physical parameters. The use of multiple internal standards to normalize closely eluting peaks with similar m/z values has also been reported [51]. Otherwise, normalization methods based on several statistical models (unit norm [52] median [53] and quantile [35]), scaling methods (auto scaling, range scaling, Pareto scaling, vast scaling and level scaling) [37]. These methods are, however, inferior to the internal standard-based methods [35].
Identification of metabolites

However, metabolite identification is the most important step in data processing. Several tools that estimate compound composition using isotope distribution or fragmentation patterns in the mass spectrum have been developed [54-57]. A theoretical study estimated that the mass spectral information available from mass spectrometers with accuracy approaching 1 ppm is not sufficient to identify the metabolites from the large public databases [58], e.g. The Human Metabolome Database (HMDB) has already identified more than 4,000 putative endogenous metabolites [59]. Instead, many studies use tandem MS (MS/MS), which generates more informative spectra including many fragment peaks, for compound identification [60, 61]. Retention time information is also used to reduce the number of possible candidates [62-65] or theoretical simulation, which predict the retention/migration times from the metabolite structure. A computational prediction techniques have been developed [66]. More sophisticated tools are required for better identification.

Quality control

Various parameters of peak detection and alignment are to characterize the quality of data processing [34]. The selection of the best algorithm, and the best parameters, to analyze the datasets obtained is interactively conducted. QC evaluation based on various benchmark tests is performed to understand the features of each algorithm and their parameters [67]. A comparison of peak detection algorithms of LC-MS data using centWave [41], matched filter implemented in XCMS [29] and MZmine [32] showed that there was only a partial overlap in the results obtained with these methods, and a number of peaks were not overlapped [41]. Even with the same algorithm, the use of different parameters strongly affected peak detection performance [35]. Evaluation of the alignment of LC-MS data using six freely available software packages, including XCMS [29], MZmine [32], msInspect [68] and OpenMS [31], concluded that no single software perfectly aligned the datasets [69]. The annotation of metabolite identities using fixed confidence thresholds has been recommended for data reporting, as has quantitative assessment of the annotation quality using the false discovery rate (FDR) [70]. Graphical user interface (GUI) enables specific steps of data processing to be rerun using different parameters [28]. Another possible means to improve performance entails the development of an iterative analytical framework with machine learning methods that allow the program to be trained to tune parameters using the difference between automated and manual data processing [34]. The integrated software including, such QC process and GUI, are quite important.

DATA ANALYSIS

Visualization

Visualization to facilitate the interpretation of large metabolomic datasets is mandatory step. Several examples are shown in Fig 2. A heatmap and pathway form facilitates comprehension of the systematic metabolomic change. MetaboAnalyst visualizes experimental metabolomic data using heatmap visualization and offers common statistical analyses, such as principal component...
analysis (PCA) [71, 72]. Pathway Project [73] visualizes data in the form of several graph types, such as bar graphs [74]. BioCyc [75] are also available [76]. Both tools take advantage of Google Map API zoom and search functions, which can be helpful for exploring the interesting details in large metabolomic datasets. The editable pathway tool is also useful when new molecular interactions is visualized [77].

**Standardization**

Standardization of data file is important for share the profiled data. This would facilitate experimental replication, interrogation and comparison over multiple investigators and studies. The Metabolomics Society has established guidelines for reporting standards [78]. The Chemical Analysis Working Group, a member of the Metabolomics Standards Initiative, proposed a set of minimum information that should be provided when reporting chemical analyses [79]. Attempts to define standards for data reporting have been made but not widely used now [10, 78, 79]. A Metabolomics Standard Initiative was recently initiated by the Metabolomics Society, and aims to develop standards for data exchange, ontology and guidelines for data reporting to solve some of the current issues. Not only standardization but also data processing tools should be developed.

**CONCLUSIONS**

Computational data processing, data analysis, and software, as well as database and standardization using bioinformatics techniques plays a key role in metabolomics. However, the current technologies are not matured and more sophisticated one should be developed.

**ACKNOWLEDGMENTS**

The author thanks Drs. Soga and Tomita for fruitful discussions. The author also thanks the research funds from Yamagata Prefectural Government and Tsuruoka City.

**REFERENCES**

Recent Developments in State of the Art Analysis of Metabolomic Data
Margarethe Zahn, Jochen Hochrein, Helena U. Zacharias, Claudia Samol, Rainer Spang, Peter J. Oefner and Wolfram Gronwald*

Abstract—Extracting relevant biomedical information from large NMR and mass spectrometry based metabolomic datasets is of considerable complexity. Therefore, the application of optimal analysis techniques is of prime importance. Here we will summarize recent developments of our group in this regard. One important task in non-targeted data analysis is clustering of the data. There are many different algorithms both hierarchical and non-hierarchical in combination with different distance measures commonly used in metabolomic research. However, it is currently unclear which ones are ideally suited for such data. In a systematic investigation, we evaluated the performance of well-established cluster methods such as Ward and PAM as well as relatively recent methods such as Affinity Propagation on several data sets derived by means of NMR-spectroscopy and mass spectrometry. We could show that correct detection of underlying biological groups critically depends on the selection of a suitable algorithm.

INTRODUCTION
The aim of metabolomics is generally the comprehensive analysis of the flow of small organic molecules such as amino acids, sugars, organic acids, bases, lipids, vitamins, and various conjugates of absorbed substances of exogenous origin through bioenergetic and biosynthetic pathways. For this they are qualitatively and quantitatively analyzed in cells, tissues, organs, organisms and body fluids. Metabolomics finds widespread application including such diverse topics as the screening of milk of dairy cows [1] or monitoring of clinical organ transplantations [2]. Investigations are mainly conducted by employing hyphenated mass spectrometry or nuclear magnetic resonance (NMR) spectroscopy.

Here, we will focus on the application of solution NMR spectroscopy to biological fluids and tissue extracts. NMR is a powerful tool for metabolite identification and quantification, as it allows the simultaneous detection of all proton-containing metabolites present at sufficient concentrations in a given biological specimen. As NMR signal volumes scale linearly with concentration, quantitative metabolic analyses are readily performed. NMR requires very little sample pretreatment and, typically, no prior chemical derivatization of compounds. On the downside of NMR spectroscopy is its comparatively poor sensitivity in the lower micromolar range [3]. However, novel techniques such as dynamic nuclear polarization allow a substantial increase in sensitivity of NMR measurements of selected compounds. Coverage of the metabolome may be increased by combining NMR with liquid-chromatography or gas-chromatography coupled mass-spectrometry (LC-MS or GC-MS), which yields lower limits of quantification in the low nanomolar to picomolar range depending on the type of mass analyzer used.

To obtain relevant biomedical information from metabolomic data, proper data analysis is of prime importance. We will highlight some recent developments of our group in this regard. Especially, we will focus on the evaluation of different data clustering methods.

MATERIALS AND METHODS
Following NMR data acquisition the next step includes spectral processing including Fourier transform and baseline correction. Prior to subsequent statistical data analysis NMR spectra have to be corrected for variation in signal position due to differences in pH, salt concentration and/or temperature. A frequently used robust method is spectral binning or bucketing, whereby a spectrum is split into a number of segments called bins, buckets, or features followed by integration of data points inside every bucket. The whole spectrum is then represented as a vector of feature integrals.

The goal of data normalization includes the reduction of technical biases and measurement noise as well as natural, non-induced biological variability, summarized as unwanted sample-to-sample variations, while maintaining biological variations of interest. Additionally, often the variance of the measured variables correlates with the corresponding mean concentration levels leading to considerable heteroscedasticity of the data. Normalization methods removing unwanted sample-to-sample variations include Quantile Normalization, Probabilistic Quotient Normalization, Variance Stabilization Normalization (VSN) etc. Normalization methods adjusting the variance of the data include Auto Scaling, Pareto Scaling, VSN etc. Note that VSN includes tools for both removing unwanted sample-to-sample variations and for adjusting the variance of the data. In a systematic investigation on numerous data normalization methods, we found that Quantile Normalization and VSN generally work quite well for NMR derived metabolomic data [4].

However, the above holds only true under the general assumption that only a relatively small proportion of the metabolites is regulated in approximately equal shares up and down, thereby, keeping the average total spectral area across groups comparable.

Typically, the number of features extracted from a single NMR experiment considerably exceeds the number of available samples, rendering proper statistical data analysis and visualization difficult. Methods for the clustering and classification of data can generally be divided into unsupervised and supervised methods [5, 6].

*Corresponding Author (e-mail: wolfram.gronwald@ukr.de)
In unsupervised approaches, no information about the underlying groups is used. Therefore, observed group separations are purely data-driven. This renders these approaches, in contrast to supervised classification algorithms, insensitive to overfitting in case of small sample numbers. These algorithms are often used for an initial analysis of potential group separations or in cases where too few samples are available for classification with rigid cross-validation. In the following, we will discuss some of the most commonly used approaches with special emphasize on data clustering.

One can basically divide the different methods available for data clustering into hierarchical-agglomerative and partitioning methods. Generally, data-clustering consists of three basic steps:

- Determine distances between individual objects.
- Application of a suitable fusion algorithm.
- Determination of a suitable number of clusters in case of hierarchical clustering. For partitioning methods the number of clusters has to be known beforehand.

Algorithm and Datasets

We evaluated three different cluster algorithms, namely Ward-clustering [7], partitioning around medoids [8] and affinity propagation [9]. Ward clustering is a hierarchical-agglomerative approach, where first each object defines its own cluster, next the two clusters that are most similar to each other will be joined. As similarity measure a variance criterion will be used. As result a hierarchical structure is obtained that may be represented as a dendrogram (Figure 1). The next investigated method is partitioning around medoids (PAM). Here, the optimal medoid of a cluster is the data point within a cluster showing the smallest distance to all other data points in the cluster. The algorithm starts with a preselected number of medoids randomly distributed within the data. The algorithm then swaps medoids until an optimal solution is reached. PAM is similar to k-means but more robust. The third investigated method is affinity propagation (AP). In this approach clusters and their respective members are computed by passing real-valued messages between the different samples of the dataset. The messages describe the affinity that one data point has for selecting another as its cluster center. In contrast to other clustering algorithms such as PAM or k-means, AP is not based on a random initialization of cluster centers at the start of the algorithm. Therefore, reliable results can be obtained in only one run of the clustering procedure. In addition AP determines to some degree number of clusters within its algorithm. The user can determine whether a larger or smaller number of clusters shall be used. We tested two different settings (AP.a and AP.b) leading to larger and smaller cluster numbers, respectively. As distance measures we employed the Euclidean norm, Manhattan norm, Canberra norm, and a Random Forests based distance measure.

Each cluster algorithm was assessed using two 1H NMR metabolite fingerprinting datasets of different origin. The first one was obtained in relation to an investigation on acute kidney injury (AKI) of patients who underwent cardiac surgery with cardiopulmonary bypass use [10]. It consists of binned 1D 1H NMR data of 85 human plasma specimens. Of all 85 patients 52 patients developed no AKI (group 1), 25 patients developed slight signs of AKI (group 2), whereas 3 (group 3) and 5 patients (group 4) developed medium and severe cases of AKI, respectively.

In the next data set the influence of prolonged road transport on animal (sheep) metabolism was investigated. It consists of binned 1D 1H NMR data of 80 sheep serum specimens. The dataset consists of 4 groups of 20 specimens each. Specimens of groups 1 and 2 were collected before transport with 2 different trucks, while specimens of groups 3 and 4 were collected after 48 h of transport [11].

Evaluation Strategy.

Based on different indices like Silhouette score $S(x_i)$

$$S(x_i) = \frac{b_i - a_i}{\max\{b_i, a_i\}}$$

with $a_i$ denoting the average distance between data point $x_i$ and all other data points within the same cluster and $b_i$ indicating the average distance between data point $x_i$ and all other data points within the next cluster. For validation of a cluster solution the average silhouette score over all data points is calculated. In addition, agreement with biological groups is assessed.

RESULTS AND DISCUSSION

Figure 2 shows a PCA analysis of the AKI data set where the 4 different biological groups are indicated in black, red, blue and green, respectively. Data show a slight separation between non-AKI patients shown in black (group 1) and all other patients.
For this data set the best cluster results according to the applied validation criteria were obtained by Ward clustering and Manhattan distances. Here, the best solution was obtained by cutting the tree at 3 different clusters. For visualization of results, members of the 3 different clusters are color coded in the PCA shown in Figure 3. The clusters shown in black and red correspond mainly to non-AKI patients, while the blue cluster corresponds to the majority to AKI patients. Therefore, cluster analysis yielded in this case a partial separation in biological groups. The results of all investigated cluster algorithms obtained for the AKI dataset are shown in table 1.

Table 1. Ranking of results of AKI data set according to validation criteria.

<table>
<thead>
<tr>
<th>Method</th>
<th>Euclid</th>
<th>Manhattan</th>
<th>Canberra</th>
<th>Random Forest</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ward</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>PAM</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>AP.a</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>AP.b</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

The sheep dataset consisted of 4 different biological groups, where groups 1 and 2 (before road transport) and groups 3 and 4 (after road transport) were highly similar to each other. Figure 4 shows a PCA of this dataset with color coding according to biological group membership.

The PCA shows a relatively clear separation between the specimens before (groups 1 and 2) and after (groups 3 and 4) road transport in PC2, while a separation between the 2 trucks is evident in PC1. Also for this dataset Ward clustering in combination with Manhattan distances yielded the best results according to the applied validation criteria (Figure 5). In this case the best solution was obtained by cutting the trees at 5 clusters.

Results showed that cluster analysis separated the data before and after transport quite well, while a partial separation between the 2 trucks could be obtained. Table 2 shows for the sheep data set a summary of the different investigated cluster approaches.
Table 2. Ranking of results of sheep data set according to validation criteria.

<table>
<thead>
<tr>
<th>Method</th>
<th>Euclid</th>
<th>Manhattan</th>
<th>Canberra</th>
<th>Random Forest</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ward</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>PAM</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>AP.a</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>AP.b</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

CONCLUSIONS

Overall the best results were obtained employing Ward clustering with Manhattan distances leading to well defined clusters that agree in most cases well with corresponding biological groups. Disadvantage of Ward clustering is that the user has to decide where to cut the resulting dendrogram. In contrast AP, the second best method, determines the optimal number of clusters by itself. Results also show that the optimal distance measure depends on the data set in question, none of the measures performed best in all instances. However, on average best results were obtained using Manhattan distances.

ACKNOWLEDGMENT

The authors thank the German Federal Ministry of Education and Research, German Research Foundation, BayGene and ReForM (Medical Faculty at the University of Regensburg) for their financial support.

REFERENCES

"Identification of marker pathways from analysis of integrative 'omics' data"

Osman Uğur Sezerman

Abstract— Unlike Mendelian diseases, complex diseases, like Cancer and Neurological diseases, cannot be caused by a single determinant. Several factors should go wrong for the development of a complex disease. In order to understand the disease etiology one should identify which pathways are affected in most of the diseased individuals rather than searching for a common SNP or a common gene. We developed a method called PANOGA that analyzes the 'omics' data in a pathway related context. Even though there were very few common targeted genes in diseased individuals majority of the top 10 affected pathways were common proving the marker pathway concept for complex diseases.

INTRODUCTION

The identification of the variants that explain familial risk of a specific disease is important since it would facilitate the development of genetic risk prediction tests, diagnostic tools and personalized therapeutical applications. Identification of alterations in the activity of biological pathways commonly found in diseased individuals may enable understanding of disease etiology. While each of these alterations changes slightly the likelihood of having the disease, they work together to give birth to the disturbances in normal biological processes. In this paper we employed a method (termed PANOGA, Pathway and Network Oriented GWAS Analysis) that identifies functionally important pathways through the identification of genes within these pathways, where these genes may be targeted by SNPs obtained from the GWAS data analysis or found to be significantly differentially expressed in transcriptomics studies or also found significantly differentially present in proteomics studies[1-4].

The methodology presented in this proposal is tested on several complex diseases including rheumatoid arthritis, intracranial aneurysm, epilepsy, and Behçet's disease; and proved to be useful. In these tests, PANOGA found several disease related pathways, which were also identified in previous studies, and additional pathways, where the relations of the genes in these pathways to the trait were confirmed.

In this study we applied PANOGA to several cancer, neurological and autoimmune data sets obtained from GWAS database.

We identified commonly targeted pathways in all cancer and neurological data sets.

RESULTS

In all cancer cases breast cancer, prostate cancer, pancreatic, liver cancer, blood cancer, esophageal cancer, colorectal cancer, bladder cancer commonly altered pathways are TGF-Beta signaling pathway, Wnt Signaling pathway, regulation of actin cytoskeleton pathway, base excision repair, cell cycle, calcium signaling pathway, focal adhesion, adherens Junction, ECM-Receptor Interaction and complement and coagulation cascades pathways.

In neurological diseases we analyzed Alzheimer, Parkinsons and Schizophrenia GWAS data. Commonly targeted pathways are neurodegenerative disease pathway and Alzheimer pathway.

In autoimmune diseases we analyzed Multiple Sclerosis, Arthritis and Asthma GWAS data. Commonly targeted pathways in all diseases are allograft rejection, autoimmune disorders and proteasome pathways.

CONCLUSIONS

Analysis of 'omics' data in a pathway related context enabled us to identify functionally altered pathways in association to several types complex diseases. All the disease specific pathways are in line with the pathways reported in the literature and shed light into the understanding of the disease development mechanism. Analysis of common altered pathways in cancer revealed common malfunctionalities occurring in the cell and different pathways revealed cancer specific pathways. There are many more common pathways in cancer than in neurological diseases and autoimmune diseases. Many cellular functionalities should be affected in cancer while very few common disruptions should occur for neurological diseases and autoimmune diseases.

REFERENCES


* (e-mail: ugur@sabanciuniv.edu).
Bioinformatics Resources for Genetic Epidemiologists

Mehmet Tevfik DORAK*

Abstract - Genetic epidemiologists use bioinformatics resources during the design of their studies and after the results are obtained to assess functionality of genetic variants. Before the study, functional SNPs are favored for inclusion and after the study, functional variants are attributed causal roles. The internet is full of easily accessible resources. In this review, some guidance will be provided for the best resources for different purposes.

INTRODUCTION

Bioinformatic analysis is used before and after a genetic association study. The approach is the same in both phases. The aim is either to select most functional genetic variants, most commonly single nucleotide variant polymorphisms or SNPs, to include in a study, or assess the causality of found associations after a study by examining functionality of SNPs modifying disease risk. In this review, both these approaches using accessible resources are discussed from an applied bioinformatics point of view. In the context of genetic epidemiology, bioinformatics is the "research, development or application of computational tools and approaches for expanding the use of biological data." Attention is also drawn to existing empirical data not to exclusively rely on computational predictions.

Genetic variation is widespread in the human genome. It is estimated that there are more than 40 million SNPs. Some of these SNPs are correlated with one another and genotyping all of them will not provide any additional information over the one that is able to represent the set it is correlated with. To avoid wasting resources, it is important to avoid this redundancy. In an association study, it is crucial to locate the causal SNP rather than a proxy for it so that this information may be used to design preventive interventions or to develop new drugs. Thus, to avoid redundancy and to increase the likelihood of finding the causal SNP, a SNP selection process should be applied before the study.

An association study usually finds multiple SNPs in a genomic region associated with a disease. Common practice is to use the P values to rank them, but this practice is not always helpful due to the P values being correlated with statistical power and power being variable depending on allele frequencies. A more objective and robust method is to assess the functionality of SNPs to support a causal role for them. While functional replication may be used to assess functionality of SNPs, this approach uses a lot of resources and the cell type used in these in vitro assays is crucial for correct interpretation of results. Bioinformatic prediction of functionality of genetic variation has reached a point that each and every nucleotide in the human genome now has been assessed for its function and the potential effect of any variation at each position has been computed. These predictions have also been supplemented by enormous amount of experimental data to help with functional annotations of genetic variants.

FUNCTIONAL SNPs

Gene expression is the most common intermediate phenotype that mediates SNP associations with disease. The initial results of the Encyclopedia of DNA Elements (ENCODE) project increased our knowledge on gene expression regulation and more broadly, genome biology. It has become clear that even SNPs in intergenic regions may have important roles in gene expression regulation. The most common mechanism for the involvement of a SNP in gene expression regulation is the alteration of a transcription factor or a microRNA binding site (usually in the promoter or 3' untranslated region or UTR, respectively). Most SNPs have been assessed for these effects and there are also empirical data from different cell types which can be used for assessment of functionality. SNPs that show correlations with mRNA levels are called expression quantitative trait loci (eQTL) and there is a lot of interest in identifying more of those to help with functional assessments.

Protein structure alteration is another mechanism by which a SNP may contribute to disease development. In earlier genetic association studies, it was thought that this was the predominant mechanism presumably because of the experience from Mendelian disorders where the causal mutation is almost always a missense type that changes the amino acid sequence. However, missense, nonsense or frameshift mutations have not been found to be common contributors to multifactorial disease risk in the first wave of GWAS perhaps because of their rarity. More recent studies focusing on rare variants are reporting more associations with missense variants. Synonymous or silent exonic SNPs are not always non-functional. They may still alter the function of an exonic transcriptional enhancer.

Splicing is an important mechanism in genome biology that contributes to the extreme diversity of proteins and is the main reason to have more than 1 million proteins from around 20 thousand genes in humans. Splicing variation is natural and changes the composition of proteins without amino acid substitutions. However, SNPs may introduce unexpected splicing variation or result in truncation of a polypeptide by altering splice sites just inside introns (Figure 1).

Besides these three main mechanisms by which SNPs may alter susceptibility to diseases, other mechanisms like polyadenylation or mRNA stability changes, tendency to induce epigenetic changes and involvement in post-translational modifications are also possible.
APPLICATIONS

First point to consider in the assessment of functionality of a SNP is its location (Figure 1). Polymorphisms altering the function of any of the crucial gene structures are likely to be functional variants and may show disease associations. Because they are the ones associated with disease susceptibility, functional polymorphisms are also called pathogenic variants.

Location of a SNP may provide hints for its functionality (Figure 2). Most eQTLs are located in promoters or intergenic regions; intronic SNPs are most pathogenic when located in splice sites; and exonic SNPs are deleterious when they change amino acid sequence with implications on protein structure (like abolishing a disulfide bond due the substitution of cytosine). The location of a SNP is included in the algorithms used in overall assessment of functionality.

Since variants modify disease susceptibility most commonly via their effects on gene expression, their eQTL status is an important factor in variant selection or causality assessment. Currently, the most promising resources for empirical information on functional SNPs are RegulomeDB [1] and GTEx (Genotype-Tissue Expression). RegulomeDB uses data from published reports, the ENCODE project, and the NCBI Gene Expression Omnibus (GEO) to annotate SNPs with known and predicted regulatory elements in the genome. Known and predicted regulatory DNA elements included are regions of DNase hypersensitivity, transcription factor binding sites, and promoter regions especially in intergenic regions. The scores range from 1 to 7, 1 indicating the highest functionality for gene expression regulation and 7 the least evidence. RegulomeDB is searchable for SNPs, genes or genomic regions and provides a ranking of variants for their involvement in gene expression regulation.

The GTEx project database is a resource to study the relationship between genetic variation and gene expression in multiple human tissues [2]. The samples used are freshly obtained autopsy samples, and genotyping and gene expression studies have been performed at the high resolution using high-throughput methods. This project provides resources to correlate genetic variation with gene expression not in a single, but in multiple cell or tissue types. No other project has analyzed genetic variation and expression in as many tissues in such a large population as has GTEx. The GTEx online tool is searchable by SNP ID or chromosome region. It also incorporates GWAS results for each phenotype examined.

Another way to examine eQTL status of a SNP or especially SNPs in a genomic region is to visualize the region on a genome browser. The Pritchard laboratory has designed one such browser where SNPs in a region are marked if they are eQTLs [3].

There are far too many individual algorithms for the assessment of functionality and their details are beyond the scope of this review. However, searchable user-friendly online tools can be used for this purpose by researchers without much training in bioinformatics. Most commonly used online tools are bioinformatics suites that generate or compile all of this information in a single search. Most commonly used ones include NCBI PheGenI [4], SNPnexus [5] and GWASdb [6]. NCBI PheGenI results incorporate the assessment of coding region SNPs by PolyPhen-2 (Polymorphism Phenotyping v2) [7]. Besides these, suites like F-SNP generate scores for functionality. The GenePi Toolbox website provides a good background to the scoring algorithms and various tools to assess SNPs for different functionalities [8].

The latest addition to the compilations of bioinformatic data on genetic variants is the Combined Annotation-Dependent Depletion (CADD) approach [9]. The CADD algorithm objectively integrates many diverse annotations into a single measure (C score) for 8.6 billion possible sequence variant in the human genome. The C scores for each actual or potential SNP is normalized for comparison and these values are provided along with raw C scores. This is not a searchable database, but a request has to be submitted using the chromosome coordinates of the variant and the result is returned as a text file. It is also possible to download the complete results for all 8.6 billion nucleotide substitutions as a very large (218GB archived file) text file.

Most GWAS data come from common variants included in GWAS chips that are presumed to be in linkage disequilibrium with unknown causal variants. ENCODE Consortium studies suggested that for most GWAS top hits, the functional SNP most strongly supported by experimental evidence is a SNP in LD with the associated SNP rather than the associated SNP itself [1]. To assess whether a SNP is in LD with a functional SNP, HaploReg can be used [10]. Like RegulomeDB, its extension HaploReg also focuses on non-coding region SNPs. Besides HaploReg, GWAS Integrator generates lists of SNPs in LD with the SNP of interest [11].

CONCLUSIONS

A genetic association study generates a statistical correlation that is assessed by a $P$ value. This result is no more than a correlation, and correlation does not mean causation. The ultimate aim of genetic epidemiology is to identify causative risk factors and this can be achieved by use of bioinformatics approaches some of which are outlined in this review. With each and every nucleotide in the human genome has now been assessed for functionality and the computational predictions can also be supplemented by already existing empirical data, there is no excuse just to report $P$ values as the output of genetic association studies. Each result should be assessed for causality and should be used to locate the causative marker if the original SNP that has shown the association is not a functional one by bioinformatic analysis.
URL List
- Combined Annotation Dependent Depletion (CADD):
  http://cadd.gs.washington.edu
- eQTL.UChicago.Edu (Pritchard Laboratory):
  http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl
- F-SNP: http://compbio.cs.queensu.ca/F-SNP
- GenEpi Toolbox: http://genepltoolbox.i-med.ac.at/?page_id=329
- GTEx (Genotype-Tissue Expression): http://www.gtexportal.org/home
- GWASdb: http://jjwanglab.org:8080/gwasdb
- GWAS Integrator:
  http://hugenavigator.net/HuGENavigator/gWAHitStartPage.do
- Haploreg:
  http://www.broadinstitute.org/mammals/haploreg/haploreg.php
- Regulome DB: http://regulomedb.org
- SNPNexus: http://www.snp-nexus.org

References
Figure 1: Steps in gene expression from transcription to post-translational modifications. Abbreviations: UTR: untranslated region; TSS: transcription start site; TIS: translation initiation site (the first amino acid of the peptide is coded here, which is usually methionine); TES: translation end site (codes for the last amino acid of the peptide); SS1 and SS2: splice site 1 and 2 corresponding to acceptor and donor sites.

Figure 2. A functional variant influences gene activity via different mechanisms depending on its location.
Reconstructing the sequences of ancient proteins by means of computational biology

Kristina Heyn and Rainer Merkl*

Abstract—The reconstruction of proteins from extinct species is a technique that was successfully applied to elucidate evolutionary problems, which are otherwise intractable. Moreover this method has great potential in enzyme design as a tool to generate initial templates allowing for the alteration of substrate specificity. We present an application and describe in silico methods, which turned out to be reliable and were used to reconstruct ancient protein sequences.

INTRODUCTION

Paleobiology is the specific branch of biology, which is aimed at characterizing ancestral proteins or genes. In this context, an ancestral protein is from an extinct organism. Unfortunately, fossil DNA is rare for very ancient times and therefore fossils are no means to reconstruct genes or proteins from very early phases of evolution. However, due to Darwin’s theory, which postulates i) the existence of a common ancestor of all species that exist or have existed on Earth, and ii) the speciation solely by differentiation, one can - at least in principle - reconstruct protein and gene sequences with the help of methods developed for phylogenetic analyses; for an overview see [1]. This approach is quite successful: One of the earliest reconstruction experiments was done in the lab of Steven Benner in 1990. In this case, the protein sequence of a ribonuclease from an extinct ruminant was determined, which lived approximately 5 - 7 million years (Mya) ago. The protein was synthesized and characterized biochemically. It turned out that the kinetic behaviour is identical to modern RNAses [2]. In 2002, a visual pigment of an archosauroid, which lived 240 Mya, was reconstructed and surveyed [3]. In 2012 a Precambrian thioredoxin (existence approx. 4 Gyr ago) was analyzed [4] and recently an enzyme complex from the Paleoarchean era was reconstructed and biochemically characterized by us [5].

During the course of evolution, mutations alter gene sequences steadily. From a long term experiment, it has been deduced that e.g. in the genome of Escherichia coli, the mutation rate is approximately $9 \times 10^{-11}$ per base-pair per generation [6]. Thus, amino acid sequences from less related extant species differ to a great extent, when compared pairwise. Consequently, the sequence of a common ancestor cannot be deduced by a simple comparison of modern sequences and a parsimonious approach, if one is interested in the characterization of proteins that existed millions or billions of years ago. This is why more sophisticated methods have to be utilized. In the following, we will explain the basic principles of sequence reconstruction and present a successful application.

COMPUTING THE LIKELIHOOD OF A PHYLLOGENETIC TREE

A prerequisite for the reconstruction of ancient sequences is a mathematical model of evolution that explains the mutations observed in a given set of protein or gene sequences. Usually, these sequences have been chosen beforehand. Then, a phylogenetic tree has to be computed, based on the chosen method and parameters and by using an algorithm able to determine the topology of the tree.

![Figure 1](image-url)  
**Figure 1** Principle of computing the likelihood of inner nodes according to Felsenstein. Indices n1 and n2 indicate corresponding terms related to two different subtrees.

Due to their reliability, maximum likelihood approaches are chosen nowadays in most cases. According to general principles introduced by J. Felsenstein [7] as a recursive problem, the computation of a phylogenetic tree is an optimization problem. Thus, it is the aim to maximize the likelihood $L(T)$ of a tree $T$, given a set of sequences and model parameters. The likelihood $L_i$ for a site $i$, which is occupied at a given state $T_n$ by amino acid $aa_i$, is:

$$L_i(T_n, aa_i) = \begin{cases} 1 & \text{if } T_n \text{ is a leaf with } aa_i \text{ at site } i \\ 0 & \text{if } T_n \text{ is a leaf with } aa_i \neq aa_i \text{ at site } i \end{cases} \quad (1)$$

In all other cases (see Figure 1), the likelihood can be computed according to

$$L_i(T_n, aa_i) = \sum_{a_{n1}} p_{a_{n1}, a_{n2}}(t_{n1}) L_i(T_{n1}, aa_{n1}) \cdot \sum_{a_{n2}} p_{a_{n2}, a_{n3}}(t_{n2}) L_i(T_{n2}, aa_{n2}) \quad (2)$$

Here, $p_{a_{n1}, a_{n2}}(t_{n1})$ is the probability that residue $aa_{n1}$ mutates to $aa_{n2}$ within the time period $t_{n1}$, which is given by

$$p_{a_{n1}, a_{n2}}(t_{n1}) = e^{-\frac{t_{n1}}{\theta_{a_{n1}}}} \quad (3)$$

$Q$ is a matrix of mutation rates, which is part of the chosen model. As made plausible by Figure 1, this computation relies

1) on the time reversibility of evolutionary processes and

---

*Corresponding Author: Rainer Merkl, (e-mail: rainer.merkl@ur.de).
Both authors are from the group “Computational Protein Design and Evolution” at the University of Regensburg, 93040 Regensburg, Germany.
2) the assumption that mutations occur independently from each other.

Assumption 1 is required to allow for a construction of the tree starting from the leaves, which are occupied by recent, i.e. known sequences and residues. Thus, these states are well defined. Assumption 2 allows us to determine the likelihood \( L_i(.) \) of inner nodes by multiplying respective terms related to subtrees; see Formula (2) and Figure 1.

Note that only for the extant proteins, the occupancy of sites (i.e. positions) with residues is known. For all internal nodes, the residues are unclear, and therefore the sum over all possible transition probabilities \( p(aa, aa) \) has to be considered. For binary trees (Figure 1), two terms have to be multiplied for each node, which explains Formula (2).

Extending this concept to all sites 1 … \( m \) of a protein sequence of length \( m \), the likelihood of the full tree can be computed by combining all terms.

**FINDING A PLAUSIBLE TOPOLOGY**

The calculation introduced so far, requires to choose a certain topology of the tree \( T \). However, the most plausible one is unknown beforehand and cannot be deduced in a straightforward manner. It is known that the number of unrooted, binary trees is \((2n-5)! / [(n-3)!2^{n-3}]\) for \( n \) sequences. Thus, already for 10 sequences, more than \( 2 \times 10^5 \) different topologies exist. It follows that heuristic methods are necessary to find appropriate tree topologies.

J. Felsenstein proposed to construct a tree in an iterative manner and to start with two sequences [7]. For each sequence that has to be added to the tree subsequently, an edge has to be selected to be split for housing the additional node. When introducing sequence \( k \), \( 2k - 5 \) edges are available, and for all topologies \( T_k \), the likelihood \( L(T_k) \) has to be computed to find the optimal one.

Unfortunately, the specific order of adding sequences determines the number of topologies to be studied. Therefore, heuristic methods are needed that alter the local topology of the tree under study. These rearrangements are stopped, if the \( L(.) \)-values reach a plateau and vary only little.

The most important method for the alteration of the local topology is Nearest Neighbor Interchange (NNI); compare Figure 2. The algorithm deletes one internal edge (here the horizontal line) and the corresponding edges in its direct neighborhood. The resulting subtrees \( A, B, C \) and \( D \) can now be rearranged in three different ways 1 - 3. For these, their likelihood value allows to decide on the optimal topology.

**CONSIDERING DIFFERENT MUTATION RATES BY MEANS OF MCMC METHODS**

The algorithm described so far, allows us to determine for a set of sequences and a chosen set of parameters (\( Q \)) the most likely tree \( T \). If mutation frequencies were constant and the same for all periods of evolution, this method would be adequate. However, it is known that these frequencies vary for different phases of evolution. Moreover, we cannot expect that mutation rates are identical for all sites of a protein. Therefore, state-of-the-art methods of phylogeny sample from a distribution of substitution rates which adds a further dimension of complexity. It is necessary to scan the space of all tree topologies by altering topology, length of edges and other parameters.

For that purpose, a Markov Chain Monte Carlo (MCMC) method is used. This algorithm accepts all alterations that improve the likelihood value and to a certain extent other combinations as well. Each of these accepted tree topologies is written to a file, which finally contains a large number (20 000 or more) of tested trees.

If set up in an appropriate manner, this algorithm will sample from all possible combinations of parameters. In the beginning (during the burn-in period) the likelihood values will vary more drastically. After \( k \) steps, the chain should reach a stationary phase. Samples chosen from this period allows us to deduce the most likely topology. Moreover, a comparison of tree topologies chosen from these samples enables the computation of a posterior probability (p-value) for each individual edge. These numbers indicate the reliability of each individual edge and can be interpreted in analogy to bootstrap values.

During the execution of an MCMC run, which can take several weeks, one can monitor the convergence of the computation. If the energy landscape of a given optimization problem contains several local minimal, it might be that the algorithm is trapped in one of them. In this case, the chain will sample only a small set of parameter combinations. Such a chain is called poorly-mixing and can be identified by tracing certain parameters; see below. In contrast, a well-mixing chain samples large areas of the parameter space, which can be deduced from monitoring as well.

---

Figure 2 Illustration of a NNI-step. 1) shows the start condition. The algorithm deletes one internal edge (horizontal line) and the four edges which are direct neighbors. The resulting subtrees \( A, B, C, \) and \( D \) can be rearranged in three different ways 1 – 3. For these, their likelihood value allows to decide on the optimal topology.
Finally, a certain number of trees sampled after the burn-in phase is the basis for the computation of the final result by means of a consensus approach, or an other method. The monitoring parameters and the p-values are good indicators for the quality of this tree.

The method explained so far, is a maximum likelihood approach for the computation of a phylogenetic tree by means of a state-of-the-art approach, which proved reliable in many cases. Generally, it can be used to clarify the phylogenetic relationship of the proteins under study. If the tree is of high quality, it can also serve as a guide tree for the reconstruction of ancient sequences.

DEDUCING ANCESTRAL SEQUENCES

The input of methods for the reconstruction of ancestral sequences is a highly reliable tree (see above) and the set of underlying protein sequences. Again, a maximum likelihood method similar to the one introduced above, can be used to deduce ancestral sequences for all internal nodes of a given phylogenetic tree. State-of-the-art methods compute for each site a probability distribution for the occurrence of all 20 residues. Frequently, the most likely residue is then chosen for each site. A severe - and only partially solved - problem is the proper positioning of indels / gaps in ancestral sequences.

**Figure 3** Software protocol for ancestral sequence reconstruction. For detail see text.

A PROTOCOL FOR ANCESTRAL SEQUENCE RECONSTRUCTION

So far, we have explained the general principles of sequence reconstruction. In the following, we describe a protocol we have successfully used to reconstruct ancestral sequences. The starting point is a protein sequence from an extant species, which we name prot_seq. Figure 3 gives an overview.

1) We use BLAST [8] and the *nr* database of the NCBI to search for the sequences of homologs of *prot_seq* in completely sequenced genomes. If *prot_seq* is an element of a larger protein complex, we concatenate the sequences of *prot_seq* and interaction partners, if our knowledge makes plausible the existence of the complex for the full time interval under study. Thus, we can increase the phylogenetic signal which helps to improve the reliability of the reconstruction. Key2Ann [9] and one2many (unpublished) are our in-house tools for the rapid compilation of data sets and for generating meaningful names for the sequences.

2) We chose approximately 100 sequences to construct a multiple sequence alignment (MSA) by means of MAFFT [10]. Afterwards, columns containing more than 50% gaps are removed by using GBLOCKS [11]. The resulting MSA is the starting point for the computation of a phylogenetic tree. A general purpose editor we use routinely for the processing of MSAs is Jalview [12].

3) Commonly, we utilize the program pb (version 3.3 of PhyloBayes, [13]) with options `-cat` `-gtr` and compute in eight independent Monte Carlo Markov Chains (MCMC) 50 000 samples each. The options `-cat` `-gtr` induce an infinite mixture model, whose components differ by their equilibrium frequencies.

4) During execution of the chains, we monitor the quality of mixing by computing the discrepancy index (maxdiff) by means of bpcomp and the minimum effective size with tracecomp. Finally, we determine a consensus tree by means of readpb. The burnin is in all cases at least 5000. These tools are elements of the software suite PhyloBayes.

5) If the tree surpasses the following minimal quality criteria, a sequence reconstruction is initiated. We only consider nodes, whose posterior probabilities (p-values) exceed 0.75. Additionally, mutation frequencies of corresponding subtrees must be < 1.0 mutation / site. If these criteria are not met, we start over from step 2) be choosing a modified set of sequences, see Figure 3. To visualize trees, we use njplot [14] and Dendroscope [15].

6) If we managed to compute a highly reliable tree, the rooted tree and the corresponding MSA are used as input for FASTML [16]. We chose the JTT substitution model and the maximum likelihood method for indel reconstruction. As representative predecessors, we select the most likely sequences related to the respective nodes of the phylogenetic tree. If the sequences contain a large number of indels, we alternatively consider PRANK [17]. This program computes an MSA based on a sophisticated probabilistic model that treats insertions correctly and tries to avoid the over-estimation of deletion events.

AN APPLICATION OF SEQUENCE RECONSTRUCTION: CHARACTERIZING LUCA–HisF

Based on the above protocol we have reconstructed and biochemically characterized ancient enzymes from the histidine and tryptophan biosynthesis pathway; manuscripts are in preparation. In addition, we reconstructed the cyclase subunit (HisF) of the imidazole glycerol phosphate synthase (ImGP-S) from the last universal common ancestor of cellular organisms (LUCA).
The crystal structure demonstrates that the resurrected LUCA-HisF enzyme comprises all specific elements of the HisF architecture. LUCA-HisF adopts one of the most ancient folds - the (βα)₈-barrel; the 3D data set is available with PDB ID 4EVZ.

The outcome of thermal and chemical denaturation experiments characterize LUCA-HisF as a protein with elevated thermostolerance; the folding mechanism is similar to that of extant (βα)₈-barrel enzymes. Biochemical tests make clear that LUCA-HisF is a highly efficient and specific enzyme, which forms a stable and functional complex with the glutaminase subunit of an extant ImGP-S.

These results suggest that the evolution of highly specific enzymes and enzyme complexes has already been completed in the LUCA era. It follows that sophisticated catalytic concepts such as substrate tunneling and allosteric communication existed already 4 billion years ago, see [5].

CONCLUSIONS: HOW RELIABLE AND LABORIOUS IS ANCESTRAL SEQUENCE RECONSTRUCTION?

Based on simulation experiments, it was shown that maximum likelihood approaches can produce accurate sequence reconstructions, even if the phylogenetic relationship is uncertain in some cases. Moreover, state-of-the-art methods are relatively robust with respect to uncertainty introduced by a less reliable phylogenetic tree [18]: The states were identical for different tree topologies.

In our hands, approximately 75% of the reconstructed protein sequences yield soluble proteins. Reconstruction failed in cases, where the MSA contained a larger number of indels, indicating that the proper positioning of gaps is not solved in all cases.

An open problem is the selection of sequences to be integrated into the MSA. We are not aware of a protocol that allows for a computer controlled compilation of data sets. This is why we work on novel methods to solve this problem.

REFERENCES

Abstract To understand the effect of weak acid stress on *Saccharomyces cerevisiae*, we integrate the transcriptome and fluxome data. For this, we focus on: (i) common genes that are differentially expressed under all types of acid stress and (ii) transcriptionally regulated metabolic reactions, based on the hypothesis that there is a linear relationship between the gene expression and the flux of the corresponding biochemical reaction and used linear regression to estimate the parameters of this hypothesis. The transcriptome data is gathered from literature, and the fluxome data is calculated from measured biomass specific rates from the same experiment, in combination with a genome scale model for *Saccharomyces cerevisiae*, INDD750. Based on the analysis, the main findings are: there are only two differentially expressed genes statistically significant under all the weak acid stresses; the fluxome data shows an increase in ATP maintenance as a response to weak acid stress and lastly the integration of both data types corroborates that there are about hundred reactions that are transcriptionally regulated during this stress.

INTRODUCTION

Weak acids, such as lactic, acetic, benzoic or sorbic acid impose environmental stress for cells by either cytosolic acidification, increasing the turgor pressure of the cell or by triggering several pathways related to apoptosis. The stress is reported to be in medical context, whereby the pH of blood (in case hypoxia, where low oxygen levels result in accumulation of CO₂ and HCO₃⁻ in blood, causing weak acid stress on red blood cells) or intercellular fluids (in case of tumor growth, as a result of Warburg effect as the lactic acidosis) decreases. Such stress can also be caused by using weak organic acid drug treatments, through which the patients are constantly under weak acid stress. Additionally, the implications of weak acids stress are also reported for food processing (as organic acids, e.g. benzoic acid, are used as preservatives in food and beverages to control microbial growth), as pesticides in agriculture, or for their use as monomers (e.g. succinic acid) in chemical industry. Still in industrial context, weak acid stress, in particular acetic acid also inhibits alcoholic fermentation process of the yeast.

Mechanism of weak acid stress is still poorly understood, yet the widely accepted hypothesis is that weak acids are found in un-dissociated form (HA) extracellularly, where the pH is around 3.5-4.5. In this form, they are lipophilic and can diffuse into the cell passively. Once they are inside the cytosol, they become dissociated (H⁺ and A⁻) and lose their membrane permeability. As a result, the cells need to use active transport to remove the excess protons to extracellular media in order to maintain the membrane potential. Membrane located proton translocating ATPase becomes active once the pH of the cytosol decreases. The inhibition of H⁺-ATPase at even lower pH results in the cytosolic acidification, causing disruptions in DNA and RNA synthesis, cellular membrane, as well as the inhibition of enzymes for biochemical reactions in central metabolism. Beside the cytosolic acidification, disruption of membrane can also triggered by the accumulation of the anions in the cytosol which yields to a high turgor pressure.

*Saccharomyces cerevisiae* is a eukaryotic organism, used successfully not only as an industrial workhorse for fermentation, in particular for winemaking, baking, and brewing since ancient times, but also ubiquitously as a model organism for higher eukaryotes. It is one of the most intensively studied eukaryotic model organisms in molecular and cell biology. Many proteins important in human biology were first discovered by studying their homologs in yeast; these proteins include cell cycle proteins, signaling proteins, and protein-processing enzymes.

Understanding the mechanism of weak acid stress on yeast cells and how the organism adapts to these conditions is important and has implications on designing drugs against tumor growth, finding new food preservatives, relaxing the toxicity of drug treatments or pesticides, improving yeast fermentation yields. The challenge in such an endeavor is that cells typically respond to weak acid stress, by changing several processes, from transcript to proteins, as well as metabolite levels, yielding, organism wide, multi omic response. This in turn, requires integrative systemic approaches, whereby we can analyze and integrate multi-omic response of the cell.

Flux Balance Analysis (FBA) is a mathematical technique to calculate and analyze all the fluxes (fluxome) for reactions within the cell through a biochemical network, under the assumption that the entire system under consideration is at a steady state. For FBA, the availability of so-called genome-scale metabolic models is required. These models contain all metabolic reactions, as well as information on genes encoding for the each metabolically active enzyme. FBA on such networks enables prediction on how target organisms behave under certain conditions. Mathematically, these models are linear system of equations, describing the mass balance for each of the metabolites in the network as:

\[
S_{\text{m} \times \text{r}} \cdot v = \mathbf{0}_{\text{m} \times \text{r}}
\]

Where \( S \) is the stoichiometric matrix of metabolic network with \( m \) rows (one for each metabolite) and \( r \) columns (one for each reaction) and the \( v \) is the flux vector for each reaction.

Microarray technology aims to measure the expression levels of all the active genes (transcriptome), in a cell. Typically, the total set of RNA is reverse-transcribed to the corresponding cDNA and hybridized to predefined oligonucleotides on microarray chips, and the expression levels are quantified using the signal intensity emitted by the hybridized chip.
In this study, we aimed to unravel the details of the response of *S. cerevisiae* cells, by integrating transcriptomic response (measured as expression levels under weak acid conditions) and fluxomic response (calculated via FBA and measured extracellular rates) of these cells to weak acid perturbations. In doing so, we search the transcriptionally regulated metabolic reactions by hypothesizing a linear relation between flux change and the change in transcript levels for these reactions. Furthermore, we aim to find cellular processes (genes, reactions) that are common to all acid perturbations.

### METHODS, THEORETICAL ASPECTS

#### Details of dataset, experimental conditions

The data used in this paper is generated by Abbot D. et al. via a chemostat experiment, whereby *S. cerevisiae* strain CEN.PK 113-7D (MATa) cells are grown anaerobically at 30°C and the dilution rate (=growth rate) 0.1hr⁻¹, until the culture reaches steady state. Additional to the reference conditions, 4 different acids (acetic, benzoic, propionic and sorbic acids) were added to the medium to impose weak acid stress. In all conditions, the consumption and production rates of extracellular metabolites are measured as biomass specific q-Rates (reported in Table I), and the expression level of all active genes are measured via Affymetrix microarrays (resulting data downloaded from Gene Expression Omnibus with the accession number GSE5926).

**TABLE I** Measured q-Rates, taken from [3], and used as constraints for FBA. All rates are in mmol/gDW/hr.

<table>
<thead>
<tr>
<th>Condition</th>
<th>q-Glucose</th>
<th>q-CO₂</th>
<th>q-Ethanol</th>
<th>q-Glycerol</th>
<th>q-Acetate</th>
</tr>
</thead>
<tbody>
<tr>
<td>No stress (Reference)</td>
<td>-6.0 ± 0.10</td>
<td>10.4 ± 0.45</td>
<td>9.52 ± 0.16</td>
<td>0.79 ± 0.02</td>
<td>0.0 ± 0.01</td>
</tr>
<tr>
<td>Acetic Acid 10mM</td>
<td>-12.2 ± 0.20</td>
<td>23.0 ± 0.50</td>
<td>21.5 ± 0.35</td>
<td>0.5 ± 0.01</td>
<td>-0.6 ± 0.02</td>
</tr>
<tr>
<td>Benzoic Acid 2mM</td>
<td>-12.1 ± 0.58</td>
<td>22.6 ± 1.45</td>
<td>21.2 ± 1.06</td>
<td>1.0 ± 0.06</td>
<td>0.1 ± 0.01</td>
</tr>
<tr>
<td>Propionic Acid 20mM</td>
<td>-13.0 ± 0.48</td>
<td>23.7 ± 0.94</td>
<td>21.4 ± 1.32</td>
<td>1.0 ± 0.03</td>
<td>0.0 ± 0.01</td>
</tr>
<tr>
<td>Sorbic acid 1mM</td>
<td>-12.1 ± 0.20</td>
<td>21.1 ± 0.28</td>
<td>21.4 ± 0.47</td>
<td>0.8 ± 0.01</td>
<td>0.0 ± 0.01</td>
</tr>
</tbody>
</table>

#### General workflow of the approach

Our integrative approach is summarized in Figure 1. Briefly, we use the transcriptome data, and search (i) for the set of genes that are differentially expressed in all weak acid stress conditions, and (ii) for the set of reactions that are transcriptionally regulated.

**Analysis of transcriptomic data**

Statistical Z test was performed on triplicate transcriptomic data to find reliable gene expression levels for all 5 conditions (*p* < 0.05). This is followed by two-sample student t-test for the reference and each of the weak acids applied as stress to determine the set of differentially expressed genes (*p* < 0.001). For each conditions, set of differentially expressed genes are collected, and this collected set is inspected for co-expression in all conditions, by calculating the correlation coefficient, R. This information is used to construct the yeast weak acid stress response network. Genes formed nodes and for each gene couple having R>0.99 are connected with an edge.

### Analysis of fluxome data

FBA was carried out by first downloading the GSMM the yeast model IND750 from the BiGG database, constraining the exchange fluxes for the measured biomass specific q-Rates. The intracellular fluxes are calculated by maximizing the flux through ATP maintenance reaction.

**Finding transcriptionally regulated reactions**

The set of transcriptionally regulated reactions are constructed by integrative analysis of fluxome and transcriptome data. For each reaction that carry non-zero flux, its flux is plotted against the mRNA level that encode the enzyme of that reaction. We hypothesize that transcriptionally regulated reactions are the ones for which a linear model can be constructed between the transcript levels and fluxes under all conditions. The coefficients and statistics of the linear model are calculated using standard linear regression techniques.

All the statistical tests were performed using Microsoft Excel and MATLAB 2013b. The COunstrains Based Reconstruction and Analysis (COBRA) Toolbox was used for flux calculations. The constructed transcriptional network was visualized using Cytoscape.

### RESULTS

Starting from complete expression data for 9335 open reading frames (ORF) on the microarray, after filtering unreliable measurements and looking for statistically significant expression, 244 differentially expressed genes were identified. The distribution of these genes among conditions is given in Figure 2. Majority of the
differentially expressed genes (207 out of 244) are found to be expressed only in single condition. Only two genes are found to be expressed in all conditions YAL062W and YBR129C.

Figure 2: The distribution of genes that are significantly expressed among different acid stress conditions.

YAL062W encodes for NADP(+)-dependent glutamate dehydrogenase enzyme, catalyzing the reaction

$$\text{glu} + \text{NADP}^+ = \alpha\text{KG} + \text{NH}_3 + \text{NADPH}$$.

Interestingly, YBR129C is of unknown function and is reported to locate in mitochondria\(^\text{[13]}\). We see from our analysis that it is differentially expressed as a response to all weak acids, and is furthermore co-expressed with YAL062W (\(R^2 > 0.8\)). Finally, the combined and consolidated list of differentially expressed genes are inspected for co-expression analysis and transcription network is constructed (Figure 3).

Flux Balance Analysis pointed that, 307 reactions carried non-zero flux in at least one of the stress conditions. The flux towards mATP expectedly increases almost twice, as a response to weak acid stress: as the cells are taking up more glucose, albeit growing at the same rate; they generate more ATP to be spent to maintenance (Figure 4).

Lastly, all reactions reported in the model to carry flux are regressed towards the corresponding gene expression. The gene-reaction connections are used from the COBRA model, as is. The underlying hypothesis is that, fluxes for transcriptionally regulated reactions will change linearly with transcript levels. The initial 1529 reaction-gene pairs were filtered according to the coefficient of variation of the regressed line. There were, 138 gene-reaction pairs having \(R^2\) greater than 0.5. All these pairs are manually inspected for false positives, resulting from numerical issues, leaving 96 gene-reaction pairs exhibiting transcriptionally regulation. Figure 5 depicts typical results: activating (Figure 5b), inhibiting (Figure 5c) transcription factors vs. no transcriptional control (Figure 5a). Finally, among the 96 gene-reaction pairs, we found 32 unique genes.

Gene ontology (GO) term analysis was performed for these 32 genes. According to the GO term analysis, a large part of the genes have roles in small molecule metabolic process. The proteins encoded by these genes have activities as oxidoreductase agents, substrate specific transporters and substrate specific transmembrane transporters. The reactions hypothesized to be transcriptionally regulated are mainly located in the mitochondria or they constitute the transport reactions from/to mitochondria.
CONCLUSIONS

We integratively analyzed the transcriptional and fluxomic response of *Saccharomyces cerevisiae* to 4 different weak acids. Taken individually, there are very few common transcription factors, which change in all conditions, and the resulting transcriptional regulation network is sparsely connected. In terms of physiology, the flux towards the maintenance ATP reaction increases with acid stress, the cell uses more ATP to survive with the speculation that cells use ATP to pump out protons resulting from dissociation in the cytosol. Lastly, the GO term analysis findings are consistent with both the fluxome data and conjectured mechanism for weak acid stress. The oxidoreductase activities are involved in acid balance of the cell. Both substrate specific transporter and transmembrane transporter activities can be linked to the H+-ATPase activity due to the acidification of cytosol. Small molecule metabolic processes include the ATP metabolism, as pointed out earlier the ATP production is significantly increased as a result of weak acid stress.

REFERENCES

Kinetic and molecular docking experiments on the inhibition of acetylcholinesterase by irgarol

Hakan Alyuruk*, Levent Cavas

Abstract— Co-biocides, such as irgarol, have so far been used in antifouling paints to prevent biofouling. However, these biocides are reported to have acute and sublethal toxicities on non-target marine organisms. In determination of sublethal toxicity, acetylcholinesterase is one of the indicator enzymes. In the present study, the inhibition of acetylcholinesterase by irgarol was investigated with kinetic and molecular docking experiments. According to the kinetic experiments results, the acetylcholinesterase was inhibited by irgarol based on mixed type mechanism. The inhibition constants ($K_i$ and $K_d$) for irgarol were found as 13.1 and 28.8 µM, respectively. Molecular docking studies showed that irgarol was able to bind acetylcholinesterase at the entrance of active site with two hydrogen bonds. The binding affinity of irgarol was estimated as -6.90 kcal mol$^{-1}$. In conclusion, newly developed co-biocides for antifouling paints should be checked for their acetylcholinesterase inhibition potentials to estimate their impact on marine organisms.

Keywords: acetylcholinesterase, antifouling, inhibition, irgarol, molecular docking.

INTRODUCTION

Acetylcholinesterase (3.1.1.7) (hereafter AChE) is an enzyme that catalyses the hydrolysis reaction of acetylcholine to choline and acetate. This reaction is essential for neurotransmission at cholinergic synapses in animals including humans [1]. AChE inhibitors have been used as therapeutic agents in treatment of neurodegenerative diseases such as Alzheimer’s disease [2,3]. AChE is also a target enzyme for pesticides that block nervous systems of insects [4]. In addition, AChEs have been used for toxicological monitoring of neurotoxic agents accumulated in non-target organisms [5,6]. The active site of AChE from *Torpedo californica* has been characterized by Sussman et al [7]. The catalytic triad and an anionic subsite are reported to be located at the bottom of deep and narrow gorge [8]. While the anionic subsite recognizes the quaternary ammonium group of acetylcholine, Glu327, His440, and Ser200 are found at catalytic triad and responsible from the hydrolysis of acetylcholine. The regulatory site of the enzyme is found at the entrance of gorge and it is called as anionic peripheral site. The key amino acids functioning at catalytic and anionic peripheral sites are Trp84 and Trp279, respectively [8,9]. Inhibition of AChE is used as an indicator of neurotoxicity caused by environmental pollutants such as pesticides, insecticides, biocides and herbicides [10].

Inhibitory potentials of inhibitors are usually determined through kinetic experiments and if the enzyme is inhibited; $K_i$ and $IC_{50}$ values for the inhibitor are presented. Another complementary approach in estimation of inhibition potential is molecular docking of inhibitor to the active site of the target enzyme. By the help of the docking method, one can predict the inhibitory effect of a toxic compound based on experimental and theoretical results obtained with a reference inhibitor. Therefore, both kinetic experiments and molecular docking studies have great importance in toxicity studies. Marine biofouling is an important problem for shipping industry. Formation of a rough surface on the ships’ hull as a result of biofouling organisms decreases the speed, increases the fuel consumption and also increases CO$_2$ emission. Therefore, antifouling coatings have currently been using to prevent the biofouling on the surfaces of ships’ hull [11]. Different types of co-biocides have been added to antifouling paints to improve their performance and service time. Many industrial antifouling paints now contain copper(I)oxide as a main biocide and co-biocides such as irgarol (2-(tert-butylamino)-4-cyclopropylamino)-(methylthio)-1,3,5-triazine). So far, many reports have been published about the negative impacts of this well-used marine antifouling co-biodecide on the aquatic ecosystems [12-16]. Acute toxicity of irgarol was first reported by Panagoula et al [17] on a model test organism, *Artemia salina*. Since a parallel observation of reduced locomotor activities of *Artemia salina* grown in irgarol containing mediums has been observed in our laboratory experiments, the inhibition effect of irgarol on the AChE was studied by kinetic experiments on purified AChE from *Electrophorus electricus* to identify the type of the inhibition and by in silico docking experiments to prove the effect of structure-activity relationship in the present study. According to the scientific literature, this is the first investigation on the inhibition kinetics of AChE by irgarol with molecular docking calculations.

MATERIALS AND METHODS

S-Acetylthiocholine iodide (hereafter ATC) and 5,5'-Dithiobis(2-nitrobenzoic acid) (hereafter DTNB) were purchased from Alfa Aesar. AChE from *Electrophorus electricus* (Type VI-S, lyophilized powder - C3389) was obtained from Sigma. Booster co-biodecide, irgarol, was kindly provided by Moravia Marine and Industrial Coatings, Istanbul, Turkey. AChE activity assays were performed on a microplate reader (ELx800, Biotek Instruments, USA). AChE activity was determined according to the method of Ellman et al [18] with modifications for microplate reader [19,20]. Briefly, 25 µL of buffer, DMSO or irgarol in DMSO were added to 2.5 mL of 1 µg mL$^{-1}$ AChE in 0.1 M Tris buffer (pH 8.0) and equilibrated in an orbital shaking water bath (GFL 1092) at 20°C. Then, 100 µL enzyme solution, 155 µL 0.6 mM DTNB solution were added to microplate wells to start the

---

*Corresponding Author (E-mail: hakan.alayuruk@deu.edu.tr).
Hakan Alyuruk is with Dokuz Eylül University, Institute of Marine Sciences and Technology, Inciralti, 35340, Izmir, Turkey, (E-mail: hakan.alayuruk@deu.edu.tr).
Levent Cavas is with Dokuz Eylül University, Faculty of Science, Department of Chemistry, Division of Biochemistry, Tinaztepe Campus, 35160, Izmir, Turkey, (E-mail: levent.cavas@deu.edu.tr).
reaction. Substrate concentrations were studied between 0.2 – 1.2 mM range. Enzymic activity of AChE was expressed as µmol acetylthiocholine converted to thiocholine by per mg AChE per minute. The molar extinction coefficient (ε) for TNB was used as 1.36 x 10³ M cm⁻¹ [20]. The inhibition type and the inhibition constants were determined by using XL-Kinetics add-in [21] for Microsoft Excel. The experiments were run in triplicate and the data were expressed as Mean±SD.

In docking studies, structural coordinates for AChE were retrieved from Protein Data Bank [22]. Since we could not find a suitable AChE template from E. electricus, AChE (PDB ID: 1VOT) from Torpedo californica was used as a template in docking studies. When amino acids of active site are considered, according to Uniprot Database [23], Ser, Glu and His are found in both enzymes from E. electricus (UniProt ID: O42275) and T. californica (UniProt ID: P04058). The selected template includes the X-ray crystal coordinates of AChE complexed with huperzine A, a competitive inhibitor of AChE, which was removed from the template before docking calculations [24]. Docking experiments were performed with AutoDock 4.2.6 suite (The Scripps Research Institute) within AutoDock Tools 1.5.6 [25,26]. The 3D structure of irgarol was retrieved in SDF format from PUBMED Compound Database (CID 91590). In order to dock irgarol, possible binding region for irgarol was determined by finding the center coordinates of the huperzine A. Binding site comparison for huperzine A and irgarol was also performed. PyRx 0.8 was used for energy minimization calculations and SDF to PDBQT file format conversion [27]. AutoDock 4 performs calculations by using coordinates belonging to the binding area of native ligand as a template in cubic geometry. Thus, each dimension of the cubic area was set to 50 points and Cartesian coordinates (X, Y and Z) of the center of cubic area were selected for each crystal structures as follows: 3.1856, 68.0136 and 63.0859 for 1VOT, respectively. Average binding affinities of inhibitors were calculated after four runs of the software by selecting the value of best pose in each run.

RESULTS

The inhibitory effect of irgarol on AChE was studied through kinetic experiments. According to the Michaelis-Menten and Lineweaver-Burk plots, irgarol has shown reversible inhibitory effect on AChE in the range of studied concentrations (Figure 1 and 2). The Michaelis-Menten constants and maximum velocities for free enzyme and irgarol were calculated by model fitting functions in XL-Kinetics add-in for Excel. The Vmax and Km values for free enzyme were found as 931.2 µmol mg enzyme⁻¹ min⁻¹ and 0.623 mM, respectively. According to the results of model fitting, the inhibition mechanism was determined as mixed type (R²=0.97084).

The apparent Vmax and Km constants were found as 1024.8 µmol mg enzyme⁻¹ min⁻¹ and 0.750 mM, respectively. The inhibition constants, Ki and Kib, were determined as 13.1 and 28.8 µM, respectively. Michaelis-Menten (1) and Lineweaver-Burk equations (2, 3 and 4) for the mixed type inhibition are given below:

$$V_0 = \frac{V_{\text{max}} [S]}{\alpha K_m + [S]}$$ \hspace{1cm} (1),

$$\frac{1}{V_0} = \frac{\alpha K_m}{V_{\text{max}}} \frac{1}{[S]} + \frac{\alpha'}{V_{\text{max}}}$$ \hspace{1cm} (2),

$$\alpha = 1 + \frac{[I]}{K_i}$$ \hspace{1cm} (3),

$$\alpha' = 1 + \frac{[I]}{K_{ib}}$$ \hspace{1cm} (4)

where Vmax is the maximum velocity of the reaction, Km is the Michaelis-Menten constant, [I] is the inhibitor concentration, Kib is the equilibrium constant for inhibitor binding to enzyme, Kib is the equilibrium constant for inhibitor binding to enzyme-substrate (ES) complex, αKib/α’ is equal to apparent Km (Km.app) and Vmax/α’ is the apparent Vmax (Vmax.app).

In docking experiments, the most probable conformations were selected from the best overall docking calculations. Binding affinity of irgarol to AChE was estimated as -6.90 ± 0.02 kcal mol⁻¹. The inhibition constant of irgarol was estimated from docking calculations as 8.76 ± 0.30 µM. The binding site for irgarol was located at the entrance of active site. It was found that irgarol
interacts with Tyr121 and Ser122 residues via two hydrogen bonds (Figure 3).

Figure 3. Molecular model of AChE docked with irgarol. A representation of interactions between side-chains of amino acids located at the entrance of active site and irgarol. Green dotted lines indicate hydrogen bonds.

It is clear from Figure 3 that irgarol interacts with phenolic O atom of Tyr121 residue by its protonated N5 atom on triazine ring. Irgarol interacts with hydroxyl O atom of Ser122 through its N atom on C4 of triazine ring. Binding site conformation analysis results showed that irgarol and huperzine A bind to AChE (PDB ID: 1VOT) from the same sites. It was observed that both compounds were closely positioned at the entrance of active site and in contact with Trp84, Gly118, Tyr130, Phe330, and His440.

DISCUSSION AND CONCLUSION

In the present study, the inhibition of AChE by irgarol was investigated with kinetic experiments and molecular docking calculations. Since irgarol reduced locomotor activity of Artemia salina nauplii in our preliminary acute toxicity tests (data not given), the possible inhibitory effect of irgarol on AChE was studied in this paper. Irgarol has been used as photosystem II inhibitor and it has been shown that irgarol interrupts the electron transport by binding to D1 protein [28,29]. This type of herbicides has been used as co-biocides in antifouling paints for prevention of biofouling on the ships’ hull. However, toxic effects of these herbicides were reported on non-target marine organisms [12-16].

Inhibition of AChE by irgarol can be attributed to its structural similarity to known AChE inhibitors. Irgarol is a symmetric triazine that has a triangular 3D structure. In the molecular docking studies, it was found that irgarol binds to AChE from the entrance of active site. Irgarol can be protonated from its nitrogen atoms belong to triazine ring and thus, it is also able to form hydrogen bonds with functional amino acid side-chains. In the literature, the inhibitory effects of triazine derivatives on AChE have been reported with enzyme inhibition and molecular docking experiments [30]. The structure of the irgarol is also similar to known AChE inhibitors and its inhibition constants (K_i) are comparable to known AChE inhibitors such as chlordiazepoxide hydrochloride, a benzodiazepine derivative [31-33]. Also, the experimentally determined inhibition constant (found from kinetic experiments) and estimated inhibition constant (found from docking calculations) were close to each other for binding of irgarol to AChE. Binding site conformation analysis showed that irgarol was able to bind AChE from the same binding site of huperzine A. The alkaloid, (-)-huperzine A, was reported to interact with aromatic amino acids on active site of AChE through hydrophobic interactions [24]. As reported by Raves et al [24], similar results were also found in the present study for binding interactions of irgarol in the active site of AChE.

In the literature, the number of studies concerning with the toxicity of irgarol on AChE are very limited and there have been only two published papers focused on two different estuarine organisms [34,35]. In these published papers, both acute toxicity and AChE levels as a biomarker for sublethal toxicity of irgarol on a grass shrimp and an estuariine fish were studied. Key et al [34] reported no statistically significant differences for AChE levels in adult shrimp but authors noted a trend for lower AChE levels as the irgarol concentrations increased. On the contrary, they reported significantly elevated larval AChE levels at the highest exposure of 2.0 mg L^{-1} [34]. In their second paper, Key et al [35] indicated that irgarol significantly changed only lipid peroxidase levels and they also stated that AChE levels were not significantly affected by irgarol. Although Key et al [34] noted the toxic and sublethal effects of irgarol were only observed at concentrations well above its environmental levels, they also warranted further studies on AChE as a possible indicator of sublethal effects. Since many reports have been published on the negative impacts of the irgarol on the aquatic ecosystems [12-16,34-39], some restrictions might be applied by the authorities on the use of this toxic co-biocide at worldwide scale in near future. Moreover, inhibition kinetics of AChE should be examined in future studies related to marine antifouling co-biocides development. Molecular docking studies can also help to understand the possible effects of candidate toxic co-biocides on the AChE in silico. We strongly suggest that the toxic biocides used in compositions of antifouling paints should be tested for their AChE inhibition potentials before their commercial uses or other applications. Otherwise, the presence of toxic co-biocides in aquatic ecosystems or closed bays where seawater currents are not sufficient would possibly affect the marine biodiversity negatively by reducing the physical activities (e.g., swimming, feeding) that are important activities in prey-predators relationships and also reproductive activities of marine organisms.

ACKNOWLEDGMENT

Hakan Alyuruk is thankful to The Scientific and Technological Research Council of Turkey for scholarship (TÜBİTAK-BİDEB 2228).

REFERENCES

Mutagenicity Assessment of Citalopram Impurities with Structure Activity Relationship Approach

Esra Emerce*

Abstract—Impurities in drugs may occur during manufacturing and storing of active pharmaceutical ingredients and drug products. The existence of these impurities may influence the efficacy, safety and quality of pharmaceuticals even at very low quantities. For the pharmaceutical industry, controlling of impurities is a crucial subject, particularly when they pose genotoxic effects, causing serious health problems. In this aspect, for minimizing the carcinogenic potential of pharmaceuticals due to impurities, more and more significant efforts are given to understand the genotoxic potential of impurities. Tools such as structure–activity relationship (SAR) can help us reveal genotoxicity in impurities while enabling predictions without additional cost, time and effort.

This study aims mutagenicity prediction of some impurities of citalopram with SAR approach. Citalopram is a selective serotonin reuptake inhibitor (SSRI) antidepressant with a wide spread use in clinical practice. Citalopram impurities EP A, B, C, D, E, F and citalopram USP related compound F were assessed by the Toxtree® knowledge-based system. The system reported no alerts for S. typhimurium mutagenicity for any of the impurities. Results have shown all assessed citalopram impurities as non-mutagenic and to be classified as an ordinary impurity according to guidelines.

INTRODUCTION

The safety assessment of pharmaceutical agents requires a highly specialized process that covers a number of toxicological assessments such as carcinogenicity, genotoxicity, and reproductive toxicity studies. Safety of pharmaceuticals is related to the toxicological properties of the active pharmaceutical ingredients as well as the characteristics of impurities within the pharmaceuticals; hence efforts on impurities which the drugs may contain have intensified in recent years. Pharmaceutical products may have residual impurities which are formed during manufacturing and formulation or from degradation of the active pharmaceutical ingredient (API) and excipients [1]. Some impurities might be defined or considered mutagens and/or carcinogens because of their reactive nature. These impurities can have serious health effects on patients if there is a chronic exposure even at very low levels [2]. Therefore, in addition to testing the drug itself, there is also a need to characterize the genotoxic potential of metabolites, degradants, impurities, and process intermediates in order to protect clinical trial volunteers and patients from the adverse effects of impurities [3]. Although regulatory bodies focus on guidance on new drug substances and new drug products during their clinical development and subsequent applications for marketing, they also recommend relevant studies to be conducted in some circumstances on new marketing applications and post approval submissions for marketed products [4].

DNA reactive substances which have the potential to directly cause DNA damage may lead to mutations even at low levels and therefore bear the potential for cancer formation. Besides, non mutagenic genotoxicants due to their thresholded mechanisms do not carry carcinogenic risk in humans as they are at level ordinarily present as impurities. Hence, in order to limit a cancer risk associated with exposure to potentially mutagenic impurities, the bacterial mutagenicity assay is used to assess the mutagenic potential/ effect and the need for controls [1,4].

In parallel, scientific community and public interest is also increased in application of in silico technologies for the prediction of preclinical toxicological endpoints, clinical adverse effects, and metabolism of pharmaceutical substances. This is also supported by further access to these technologies by scientists as well as the recent regulations granting the use of these technologies for chemical risk assessment [5]. Over the years studies on mutagenesis and clastogenesis have provided sufficient structure–activity relationship (SAR) information which may be used to establish computational models for prediction of endpoints based on molecular structure and reactivity [6,7]. To predict the potential mutagenicity of impurities, (Q)SAR or in silico tools are commonly used and current guidelines indicate that a structural assessment is fairly enough to conclude whether an impurity is non-mutagenic or not. The two categories in which most commonly used in silico tools are divided are as follows a. empirical or rule-based expert system (i.e. DEREK, ToxTree); b. statistical or (Q)SAR-based system (i.e. MultiCASE/MC4PC, Leadscope, SciQSAR). For the prediction of impurities’ mutagenicity, it is observed that a rule-based expert system complemented by expert knowledge or a second (Q)SAR model would be appropriate [8].

In this study, the aim was mutagenicity prediction of some impurities of citalopram with SAR approach. Citalopram is a selective serotonin reuptake inhibitor (SSRI) antidepressant. Food and Drug Administration approved “citalopram hydrobromide” for marketing in 1998. Since then it has been widely used in clinical practice. Citalopram impurities for which no report of genotoxicity data is available and which are determined for investigating mutagenic effects are already indicated in United States Pharmacopeia (USP) and European Pharmacopeia.

FURTHER INFORMATION

To predict mutagenicity for some impurities of citalopram, Toxtree® (v2.6.6) knowledge-based system was used. Toxtree® (toxic hazard estimation by decision tree approach) developed by Ideacounsel Ltd. (Sofia, Bulgaria) under the terms of a contract with the European Commission Joint Research Centre. Toxtree® is an open-source application that categorizes chemicals and makes
predictions for different kinds of toxicological endpoints by applying decision tree approaches. For making a prediction mutagenicity of impurities, ToxTree® program conducts the Benigni/Bossa rulebase. With the Toxtree execution of the rulebase, processing of a query chemical produces a small number of different outputs, namely: a) no structural alerts for carcinogenicity are recognised; b) one or more structural alerts (SAs) are recognized for genotoxic or non-genotoxic carcinogenicity; c) SAs relative to aromatic amines or α-unsaturated aldehydes are recognised, and followed by the chemical going through Quantitative Structure-Activity Relationship (QSAR) analysis with which the outcome may be positive or negative.

Parent compound citalopram and its impurities are shown in Fig.1.

Citalopram hydrobromide:  
(CAS No: 59729-32-7)  
1-[3-(Dimethylamino)propyl]-1-(4-fluorophenyl)-1,3-dihydro-5-isobenzofuran-carbonitrile hydrobromide

Citalopram EP Impurity A:  
(CAS No: 64372-56-1) 1-(3-(Dimethylamino)propyl)-1-(4-fluorophenyl)-1,3-dihydro isobenzofuran-5-carboxamide

Citalopram EP Impurity B (USP related compound B):  
(CAS No: 411221-53-9) 1-[3-(Dimethylamino)propyl]-1-(4-fluorophenyl)-3-hydroxy-1,3-dihydro isobenzofuran-5-carbonitrile oxalate

Citalopram EP Impurity C (USP related compound C):  
(CAS No: 372941-54-3) 3-[3-(Dimethylamino)-1-propyl](4-fluorophenyl)-6-cyano-1(3H)-isobenzofuranone oxalate

Citalopram EP Impurity D (USP related compound D):  
(CAS No: 144025-14-9) 1-(4′-Fluorophenyl)-1-(3-(methylamino)propyl)-1,3-dihydro isobenzofuran-5-carboxamide

Citalopram EP Impurity E (USP related compound G):  
5-Chloro-1-(4-fluorophenyl)-1,3-dihydro-N,N-dimethyl-1-isobenzofuranopropanamine hydrobromide

Citalopram EP Impurity F (USP related compound H):  
(CAS No: 64169-39-7) 5-Bromo-1-(4-fluorophenyl)-1,3-dihydro-N,N-dimethyl-1-isobenzofuranopropanamine hydrobromide

Citalopram USP related compound F:  
(CAS No: 55011-89-7) Dimethyl-(1-methyl-3,3-diphenyl-allyl)-amine hydrochloride

CONCLUSIONS

In this study, when the SAR model applied to citalopram impurities EP A, B, C, D, E, F and citalopram USP related compound F, the system reported no alerts for S. typhimurium mutagenicity for any of the impurities. Results have shown all assessed citalopram impurities as non-mutagenic and to be classified as an ordinary impurity according to guidelines. Determination of the investigated impurities as ordinary impurities is a positive outcome for patients using this antidepressant. It is also important to determine toxicity via computational toxicology methods for other impurities where there is no toxicological data in the literature as in Citalopram impurities. With SAR systems, it is possible to conduct faster analysis which enables high throughput while saving time and resource, at varying accuracies, in addition to allowing reduction of animal studies. It is expected that with these advantages and improvements on the subject, the in silico approaches would find more grounds in scientific world.
REFERENCES


Abstract— Recent pandemics increased the need for efficient diagnostic tools that can provide rapid results for a number of pathogens without compromising sensitivity and specificity. To address this, a number of multiple-pathogen detection systems for simultaneous, multiple, rapid pathogen detection have been developed. Among them, systems that are RT-PCR based using TaqMan low-density array cards (TAC) are demonstrated to be an effective surveillance tool for this purpose.

However, when multiple labs that may be geographically distributed collect samples and execute TAC runs, it is essential to standardize and store TAC results and configuration parameters in a central database that can be accessed remotely. Such a design not only conveniently enables scientists to upload, share and compare 'Run Results' with other labs, but also enables epidemiologists to rapidly access and evaluate the results for in-depth investigation.

Therefore, this paper describes the building blocks and development of an integrated data entry and analysis system that; i) stores TaqMan (RT-PCR) Array Card (TAC) experimental setup and run results, ii) enables epidemiologist to retrieve and evaluate run results.

INTRODUCTION

Among the leading causes of death and disability worldwide perhaps infectious diseases still remain at the top [1]. This is because, as population increases and we live and work close to each other in urban centers, pandemics such as 2009 H1N1 Influenza A, spread quickly and become deadly for some. Despite the impact of infectious diseases on global human health, we are still at the early stages of development of efficient diagnostic tools that can provide rapid results for a number of pathogens without compromising sensitivity or specificity.

In recent years, improved nucleic acid amplification and detection technology has facilitated the identification of multiple pathogens simultaneously that had previously proved labor intensive and/or impossible to detect using traditional culture or immunofluorescent techniques [2, 3, 4].

In 2011, the CDC reported the evaluation of the Taqman array card (TAC) for 21 respiratory viruses and bacteria using singleplex quantitative PCR assays [3]. In the evaluation by Kodani et al, TAC was designed to test 8 samples per card in duplicates with one negative control sample for testing 292 clinical specimens against 21 different pathogens. The pathogen composition of each TAC may vary, and each card consists of 8 ports, each may contain 2x24 connected wells, giving 384 unique data points.

It was shown that rapid pathogen detection method based on TAC improved capacity to detect etiologic agents for acute infections during outbreaks. It was also useful for routine surveillance, while identifying co-detections of multiple pathogens and establishing rates for bacterial and virus co-infection.

However an additional component to collect and process detailed epidemiologic information about samples and patients was needed. Considering the various layers, there is a tremendous volume of data that must be processed, stored and analyzed systematically for interpretation in any such study.

Even for the small number of clinical specimens studied in the initial TAC evaluation, it was understood that approximately 40 TAC were used and over 15000 results were stored as Microsoft Excel spreadsheets [3]. Traditionally next steps to analyze data involve the development and execution of scripts to parse the results of RT-PCR assays. Few statistical methods are proposed in the literature for the data analysis. Although, using SAS or any other scripts for analyzing the data seems statistically powerful, these scripts need to be re-written or modified for any non-standard or new file formats, new instruments, new card layout, new probe, and pathogen design.

Further the difficulty of analyzing data increases exponentially with the number of sites participating in any surveillance project. We developed the TAC Data Management System (DMS) to address the need for assimilating data from several TAC related surveillance projects in a secure manner with a user friendly web-interface and controlled access. In this paper, we a) review requirements and uses cases, b) discuss considerations, limitations and design decision, c) explain the software architecture and modules.

METHODOLOGY

An important task in creating any software program is gathering of the requirements. Traditionally, this is done before the software development begins [5]. Skilled and experienced business analyst and software programmers identify the needs and develop the uses cases while recognizing the incomplete, ambiguous, or even contradictory requirements at this point. However, in this TAC DMS reference implementation project as in any bioinformatics scientific research, our requirements changed very rapidly, requiring us to follow a composite approach namely Agile methodology [5].

Current Workflow

In a typical lab, workflow of TAC data analysis may be outlined as below.

* Epidemiologist or program coordinator receives case report and samples
* TAC card is designed with pathogen and probe information
* Samples are sent to lab for analysis

---

*Corresponding Author (e-mail: mgunay@akdeniz.edu.tr).
Melih Günay, Akdeniz University, Antalya/Turkey.
Rajarajeswari Balasubramaniyan, rajikarth@gmail.com
* TNA extraction is done on samples in batch manner
* Extracted samples are analyzed using ViiA-7 Realtime PCR machine
* Run results are verified manually
* Results are saved as .txt or .xls files
* Analyst write/modify parsing scripts to parse the .txt/.xls files
* Epidemiologist merge TAC results with patient information to draw conclusions.

**Requirements**

The main purpose of TAC DMS is to automate and store much of the information outlined above. Furthermore, the following high level requirements are identified to be critical for the reference implementation of TAC DMS:

- At any time, the system must handle data from multiple surveillance projects. One surveillance program can have multiple sites/labs associated with it, and the need to capture and organize data from every site/lab is crucial for effective surveillance.
- The need to implement authentication mechanism is a vital one, restricting individual users to limit access to the data based on the project/site and their role in the project.
- The system should be accessed remotely, providing access to multiple sites/labs involved in the program.
- The system needs to provide efficient data sharing as needed between various labs and epidemiologists.
- Detailed information about the collected samples, extractions used, pathogens tested, TAC card layout and design and TAC run results need to captured and stored.
- Raw RT-PCR reaction data from each well with Ct values should be captured and stored. The system should be flexible as it allows assigning new Ct values for individual PCR reactions based on subjective evaluation.
- The system should include validation mechanism to validate the individual RT-PCR Ct value results with respect to positive and negative controls.
- Approval mechanism should be provided by which the TAC-run results are restricted until it is approved by an authoritative person.
- The system need to keep complete track of audit trials performed.
- The system needs to provide efficient search programs to help browse through the data. The data should be browsable per run, sample, reaction and pathogen, and the results should be downloadable for further analysis.

- The system needs to provide analysis tools using which basic spatial and temporal analysis with respect to pathogen and/or sample can be done.

**SOFTWARE ARCHITECTURE AND DESIGN**

Once the system requirements has been identified, the next step towards centralizing the TAC data and provide access to the users from different sites is to build the system software architecture in a client/server configuration as shown in Fig.1.

**Figure 1. Software Architecture.**

In this setup, the server is a high performance computer with 4 cores and 16 MB RAM where most of the computationally intensive executions occur. The computer which hosts Tomcat Web Server also acts as a gateway to the database server. In order to improve the performance and scalability in our system, the MySQL database server is hosted on another computer.

In our system, the front-end software technology platform uses Java [7] and Goggle Web Toolkit 2.6 (GWT). Data is exchanged between client and server using GWT RPC framework that is based on well-known Java servlet architecture. The server-side code that gets invoked from the client is often referred to as a service. The client code uses proxy class to make calls to the service. GWT handles serialization of the Java objects passing back and forth the arguments in the method calls and the return value [8] between the client and server. Fig.2. shows the architecture for the TAC DMS. For instance when a button is clicked, the click event is captured and an API call is made to the TACService Implementor (TACServiceImpl) on the server side through the proxy TACServiceAsync. The TACService implementor handles the client call in server side and as needed interacts with the Data Access Object Layer (DAO) for database inquiries. Once the data is gathered and computation is completed, either the result is serialized and returned back to the client or an exception is thrown from the server side. If the result is returned successfully, then the GUI is updated with the data. Otherwise a warning message is displayed for further user action.
Figure 2. Client Server Architecture based on GWT.

The TAC Service implementor is a singleton JAVA object and all of its methods are synchronized. In order to increase performance, both the DAO accesses and computationally intensive functions use threads. For instance as each reaction is independent from each other, the CT values are computed in parallel using one thread for each well. However, once the CT value of each well is calculated, the infection rate determination for the given set has to be computed in serial.

Figure 3 Overview of TAC data management modules.

While the MySQL database (Version 5.1) is configured and accessed through MyBatis [9] framework, the configuration of MyBatis is managed by the Spring Container [10]. The Spring container (See Fig. 2.) is also responsible from configuring the service layer and managing database transactions. For instance, saving of specimen may occur in 2 steps, i) create location, ii) save specimen, respectively. For some reason, if the saving of specimen fails, then the location that was just created needs to be erased. Spring Transaction Management module is configured to handle such exceptions.

Fig.3. show the overview of the TAC DMS. TAC-DMS may split into following 3 modules; i) Sample/Extraction Meta Data Entry ii) Upload of TAC Run Results and iii) Browse/Download Data for analysis.

Database Design

The TAC Data Management System (DMS) primarily consists of 7 major components namely Project, User, Pathogen, Sample, Extraction, Reaction and TAC Run as shown in Figure 3. Each run consist of multiple reactions. In addition, each sample may have multiple extractions. Hence the unit for ‘Sample/Extraction’ module is the ‘Extraction’ and ‘TAC Run Result’ module is the ‘Reaction’. Reaction holds reference to Extraction for both modules to be linked. According to design each sample is also associated with the ‘Project’. This allows software to categorize the data and limit access to it as each user may be allowed to work on one or more projects. Finally, each reaction is also associated with a pathogen.

Software Features

Each user is required to login the system to manage/view data. User authentication is also used to keep track of audit trials performed on data. To accomplish audit trail, each database table include 'modified_by' and 'date_modified' columns. Furthermore, we also included 'valid' column in each table to facilitate 'invalidation' of entry when needed. Detailed description of the software modules are given below.

User Authentication and Authorization

TAC DMS incorporates an authentication and authorization module allowing an administrator to authenticate users and assign roles to them.

Authentication is done based on project. Until an user is permitted to view the data, he/she wont be able browse through the data. Authorizing cards and editing data is further restricted, allowing only handful of people to perform the editing and authorization, keeping the security to a maximum level.

Access to the data, as assigned by role provides an additional layer of data integrity and security. While role based access to data may allow one user to modify the data and other to only view the results.

Meta data entry for sample, extraction and TAC run

The meta data entry module stores information about surveillance projects including sample meta data (type, location, date collected and received, sample type and conditions upon arrival), extraction instrument and settings, temperature and chemical treatments for sample extraction and buffers, positive and negative controls for the TAC run. For each of the above submodules, data entry and management forms are developed. Meta data for sample, extraction and TAC run are normalized and tightly linked to each other using foreign keys.

TAC-run reaction result upload

The TAC-run reaction result upload module allows user to upload TAC Run results. Additional information such as experimenter, instrument, card barcode, execution date is also captured during the upload process.

Regardless of the card layout each well reaction is associated with the sample extraction and pathogen. Ignoring card layout at this stage allows software to handle different plate configurations. Once reaction results are uploaded, user may view each well PCR curve, add additional information to the run and finally store the TAC run result.
Once the TAC-cards are processed by the PCR platform (ViiA 7), a CT value is assigned for each assay if the fluorescence reading increases above a certain threshold setting. Due to an intrinsic error while differentiating signal from noise, many times the PCR platform will assign a CT value erroneously. Thus a time consuming laboratorian’s verification is needed for validation and result assignment. In order to reduce the effort, we developed a second order differential equation based algorithm for CT value prediction. For every PCR assay, the automatic CT prediction algorithm predicts a CT value and assigns to the assay. Once a TAC-card is uploaded it will be given a status of “New” until an authentic person checks for the validity of the card and approves it by giving an “Approved” status. Any discrepancies identified between the assigned and predicted CT values and the laboratorian’s interpretation result are indicated by a red alert.

Until the user assigns a value for these discrepancies the status of the card cannot be approved. An audit trial is stored indicating individual who approved the card and the date it was approved.

Data Retrieval and Analysis Module

The data retrieval and analysis module allows users to conveniently search, manage, download and analyze TAC run results. Results can be browsed either by i) TAC Run or ii) Sample. Various search parameters, such as program, project, sample type, sample location, run status (new or approved), specimen id, TAC run id, created by and sample date collection range, may be grouped together for effective browsing. The TAC runs or specimens may also be sorted by pathogen, sample or CT values to assist browsing.

Browse by card:

Using the browse by card function, uploaded TAC runs for any project or program can be searched and retrieved. Many search parameters are available for effective browsing, including search by program, project, run id, run operator, run date and/or card approval status. Search results appear in the adjacent tab. Any card run can be selected and the following detailed information about the card run can be viewed; a) card setup information, b) CT values for card layout in tabular form c) CT values by well, d) pathogens detected for each sample, and e) validity of each sample with respect to its negative and positive controls.

Another powerful aspect of the database is the ability to review the individual fluorescence curves or PCR plot for each assay within the TAC run.

Browse by sample:

The browse by sample function allows for the analysis of sample results from any card and surveillance project. Following the search, user may select samples of interest for analysis. Search may be repeated until a set of samples are constructed. Once the sample set is prepared for analysis, specimen infections per well is reported as demonstrated in Table 2. Each cell of the table display the number of infections over the total number of wells tested for the corresponding specimen and pathogen. In addition, the Infection Rate which is calculated by Eq. 1. per pathogen for a select set. Software reports infection rate in a tabular format as well as histogram.

\[
\text{Infection Rate} = \frac{\# \text{ of Positive}}{\# \text{ of Total Samples}}
\] (1)

Conclusions

Detailed information about samples, processes, pathogens, infections, location, and patients play a crucial role in understanding the etiology of infections and are vital during outbreak and diagnostic situations.

TAC DMS stores and manages both laboratory and epidemiological data from various surveillance and epidemiological studies. To facilitate the automated analysis, TAC DMS also provides a unique algorithm that automates CT value prediction reducing the data analysis time significantly while improving CT consistency significantly. In addition, TAC-DMS web interface provides various data selection, analysis and retrieval of a pathogens infection statistics with respect to patients, location and date relevance.

Acknowledgment

The authors thanks Dr. Ashley Fowles for assisting requirements and workflow guidance. The opinions and assertions contained herein are those of the authors and are not to be construed as official or as reflecting the views of the Department of Health of Sciences or the U.S. Government.

References

10. Get Started with Spring.
Abstract— As the usage of the Internet becomes global, accessibility of the Web content is crucial for all people regardless of their disabilities. In order to achieve that, designers should consider the problems and difficulties that people with disabilities can face on the Web. Visually impaired users consider the Web as a primary source for information. Web developers are obligated, both ethically and legally, to address the needs of people with disabilities when putting content on the Web. In this study, several Web accessibility evaluation tools are evaluated based on their capability to find accessibility problems. Quantitative results are presented to emphasize the strengths and weaknesses of these evaluation tools.

INTRODUCTION

The widespread use of the Internet and Web technologies has given people many capabilities to perform their daily tasks using Web resources. People communicate with each other through social networking, read newspapers, pay bills, buy things, and search information through the Internet. The society heavily depends on using Web applications and technologies. However, people with disabilities have a considerable disadvantage when they access the Internet. There are millions of disabled people facing many challenges accessing Web resources. There are approximately 314 million people worldwide, who are visually impaired with 45 million being blind [1]. Visually impaired users consider the Web as a primary source for information; however most of the websites have accessibility issues for visually impaired and blind people. This paper presents the results of the comparative research done on the evaluation tools that are used to find accessibility issues on websites. The evaluation criteria consists of displaying images, headings, navigation, JavaScript elements, using good contrast, and text or language settings. In the analysis, first the total number of different errors found by the evaluation tools was measured, and then each tool’s percentage of success to find the accessibility errors was calculated. The comparison of each tool’s percentage of success is presented to reach a conclusion about the tools’ capabilities to find Web accessibility errors.

WEB ACCESSIBILITY EVALUATION TOOLS

Three accessibility evaluation tools are used in this study: a. WAVE [2], a free tool developed by a non-profit organization named WebAIM, b. AMP [3], a commercial application developed by SSB Bart group, and c. Firefox Accessibility Extension (FAE) [4], an open source browser extension. The comparison was done on a public website, which belongs to a non-profit architectural organization, Van Alen Institute [5]. The Website uses many images and JavaScript elements and proved to be a good test-bed for this study. Each evaluation tool provides automated evaluation capabilities, therefore allows the user to enter the website address to do the evaluation and generate a report of the accessibility issues. The report on the errors found on the selected Website by using the WAVE tool is illustrated in Figure 1.

Figure 1. The accessibility evaluation report generated by WAVE

The report generated by the AMP tool is shown in Figure 2.

Figure 2. The accessibility evaluation report generated by AMP

The report generated by the FAE is shown in Figure 3.

Figure 3. The accessibility evaluation report generated by FAE
The results of the evaluations are tabulated in Table I.

<table>
<thead>
<tr>
<th>Errors</th>
<th>Evaluation Tools</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WAVE</td>
<td>AMP</td>
</tr>
<tr>
<td>Images</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Headings</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Navigation</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Java Script</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Contrast</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Text</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Language</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total Errors</td>
<td>15</td>
<td>36</td>
</tr>
</tbody>
</table>

The majority of the accessibility issues are caused by images. The common cause of image errors is the lack of alternative text for the images. Visually impaired and blind people cannot access the graphical content if alternative text for the content is not provided. The alternative text enables the users to use a screen reader to read the textual description of the graphical components. The alternative text should illustrate the purpose of the graphic accurately. Using the alt attribute for the image tag in HTML can be used to provide alternative text for the graphical content. The second major cause of accessibility errors is the JavaScript errors. If a website uses JavaScript elements to display content, the functional text must be provided; therefore visually impaired users can access the content while the script is turned off.

The bar-chart representations of the issues found by the evaluation tools are illustrated in Figures 4, 5, and 6.

**RESULTS AND DISCUSSION**

A comparison of the Web accessibility evaluation tools based on their capability to find accessibility problems on a Website was done. In order to find out what type of errors each tool is capable of finding, evaluation results of these tools based on multiple elements, such as images or headings, were recorded. In the analysis of results, first the total number of different errors found by the evaluation tools was calculated, and then each tool’s percentage of success to find accessibility errors in the Website was computed. This allowed a comparison of the evaluation tools’ capability to find accessibility errors. The overall percentage of finding accessibility errors for the three tools we used are illustrated in Figure 7.

The evaluation tools’ capabilities based on finding specific kinds of errors were also compared. This comparison provided information in the form of, x evaluation tool is most capable of finding y type of errors. For example, all three tools are 100 percent capable of finding image errors caused by not providing alternative text for the images. The comparison of the percentages of the tools’ capabilities to find image errors is illustrated in Figure 8 below.
Next, the comparison of the percentages of the evaluation tools’ capabilities to find JavaScript errors is displayed in Figure 9.

![Figure 9. The capabilities of the evaluation tools to find JavaScript errors](image)

According to the results, AMP is the only tool capable of finding the JavaScript errors.

Next, the comparison of the percentages of the tools’ capabilities to find heading errors is displayed in Figure 10.

![Figure 10. The capabilities of the evaluation tools to find heading errors](image)

The result shows that neither WAVE nor AMP are capable of finding heading errors. On the other hand, FAE found all the heading errors. Next, the comparison of the percentages of the tools’ capabilities to find navigation or link errors is displayed in Figure 11.

![Figure 11. The capabilities of the evaluation tools to find navigation errors](image)

The result shows that while AMP was able to find 100 percent of the link errors, WAVE found only 10 percent of the link errors, and FAE found none of the link errors.

**CONCLUSION**

The most frequently used Web accessibility evaluation tools, WAVE, AMP, and FAE, were analyzed based on their ability to find accessibility problems. The investigation of the Web accessibility tools shows that different accessibility tools are capable of finding different accessibility problems. For instance, albeit all the three evaluation tools were capable of finding image errors on the selected Website, only the AMP tool was able to find the JavaScript errors. FAE was capable of finding neither the JavaScript errors nor the navigation errors. The WAVE also could not detect the JavaScript errors and it can only find approximately 10 percent of the navigation errors on the selected Website.

The evaluation results showed that some of the evaluation tools found some errors even though these were not actual errors based on our manual analysis. We believe that these false-positives occurred due to various implementation errors in these tools.

Consequently, none of the evaluation tools in this study can be considered perfect to find accessibility problems of a Website. We strongly recommend using all these three tools in conjunction to evaluate accessibility of a Website when possible.

**REFERENCES**

4. [http://fae20.cita.illinois.edu/](http://fae20.cita.illinois.edu/) [last access date: 11 Sep 2014].
The missing professional in the literature of the interprofessional collaborative practice: Speech and Language therapist

Ali Yildirim and Sezer Domac

Abstract—Interprofessional Collaborative Practice (IPCP) is a work practice that involves different health and/or social care professionals who share a team identity and work collaboratively in an integrated and independent manner to solve problems and deliver services. Interprofessional education (IPE) involves members (or students) of two or more professions, associated with health or social care, engaged in learning with, from, and about each other (Barr et al 2005). Speech and language therapy (SLT) is known to be new profession in the health and social care, predominately working with people with communication and swallowing difficulties/disabilities. SLT professionals work in diverse settings, alongside a range of other professions. There has however been very little research on SLT teams in community settings, as well as the wider range of professions such as allied health. In particular there has been insufficient research on SLT and of interprofessional practices to support children with communication disabilities. This paper aims to provide whom IPCP can contribute for the speech and language therapy and audiology departments. Following this, a detailed exploration of key literature, and critical reflection upon this will be presented. The report then sets out the methodology that will be adopted to fill this knowledge gap, the progress of the project so far and how the work will demonstrate an original contribution to knowledge appropriate.

INTRODUCTION

Interprofessional Collaborative Practice (IPCP) is a model which first gained academic popularity in the health and social care sector following the advice of Beveridge report in 1942. IPCP occurs “when multiple health workers from different professional backgrounds work together with patients, families, careers and communities to deliver the highest quality care” (World Health Organization, 2010:55). Shared team goals, ideas, decision-making, team identity, partnership and integration of team members, have been identified as some of the key elements of IPCP (Day, 2006, WHO, 2010). Although professionals’ ideology, values, self-interest and specialism often differ, partnership working within and between agencies has been found to be directly related to health service user satisfaction (Spencer, 2003, Heath Service Ombudsman for England, 2006) partly because professional’s enhanced relationships with each other directly influence their attitudes towards patients. Professionals’ communication skills and knowledge around health and social care were found to be enhanced as a result of increased integration with other professionals and patients. As illustrated above, IPCP has gained an important space within the health care system, largely due to its potential in delivering highest quality of medical care through collaboration.

FINDINGS AND DISCUSSION

The theory behind the IPCP model is that if different professional groups, which are deployed within a mutual fieldwork together, service delivery will enhance in many ways. Interprofessional education (IPE) prepares professionals for the ICPC in the pre-registration and it involves members (or students) of two or more professions associated with health or social care engaged in learning with, from, and about each other (Barr et al., 2005). Glen et al (2004) advised that IPE needs to be built on evidence base practice. Speech and language therapist (SLT) is relatively a new known health profession working with people with communication and swallowing difficulties/disabilities. SLT works in diverse settings with a range of other professions. They work closely with parents, carers and other professionals, such as teachers, nurses, occupational therapists, physiotherapists and doctors. The ability to work effectively with others requires many skills and coordination.

In practice, SLT services have undergone significant changes in recent years. The need for improved interprofessional teamwork, the increased use of psychosocial interventions and increased collaborations with users and carers are just some of the demands they have had to adapt to. In the literature of IPCP, particularly related to SLT is written by Boaden et al (2006) that stated Interprofessional Dysphagia Framework. This framework informs professionals about the strategies for developing the skills, knowledge and ability of speech and language therapists, nurses and other healthcare professionals/non-registered staff to contribute more effectively in the identification of people with in the management of feeding-swallowing difficulties. Another survey carried out in speech and language therapy field was by ASHA (American speech-Hearing Association) in 2013.

The survey was designed to provide information about health care based services delivery and to update information and date of SLT Health care survey. All date are collected six type of health care facilities: general
medical and long term acute care hospitals, rehabilitation, rehab hospital, paediatric hospitals, skilled nursing facilities, home health agencies and clients homes, and outpatient clinics and offices. As sample of 4000 ASHA certificate SLTs employed in health care settings was selected to participate in this study. In this survey, Interprofessional collaboration was defined in the survey as occurring “when individuals from different fields work together to provide integrated services in a health care environment.” Respondents were asked to identify how frequently they engaged in interprofessional collaboration with the other professionals. The result of this survey shows that in some cases of their work for example in documentation and assessment SLTs do not prefer to work interprofessionally. However, in clinical team meeting and patient/ family meeting they prefer to work interprofessionally.

In interprofessional field has a lot of research about the other professional views about the IPCP. Such a, nurse, doctors, GPs, occupational therapist, Physiotherapists. For example, Villeneuve and Shulha (2012) point out that the view of occupational therapists (OT) about the IPE in pre registration courses is that IPE is an important and helpful course for the personal development of OTs. In addition, they mentioned that adult learning theories might be helpful for the IPE course structures. There are some others researches, which show the other professionals views of the IPE and IPCP such as doctors, physiotherapists. However, in this literature the missing area is the view of the SLTs, knowledge and understanding about IPCP. Moreover, SLT’s work covers a range of environments, such as community health centers, hospital wards, outpatient departments mainstream and special schools, children's centers, day centers, clients' homes, courtrooms, prisons, young offenders' institutions and independently through private practice (Royal College of Speech & Language Therapists, 2013).

In the literature of the IPCP about workplace or work environment has significantly effects to the interprofessional working. Gregory et al (2012) stated that the space or the work affects the interprofessional learning in the daily practice. Billett (2014) mentioned that shape of work place and organization; sequencing of experience for interprofessional work can affect the understanding of concepts, procedures and values of IPCP. All this literature in IPCP field shows that work place has greater impact on the IPCP but there needs to be further exploration about the SLT’s work place and effects of it to IPCP. This is another identified gap of the both IPCP and SLT literature.

The theoretical background of IPCP is about the professional views, attitudes, team roles, pre-registration education, hierarchal barriers and professional identities however, in speech and language therapy field there is not enough research to shows this.

CONCLUSIONS

This study shows that in interprofessional education and interprofessional collaboration field there is not enough study shows the SLTs perspective about IPE and IPCP. Also in speech and language therapy field there is very little research shows the perspective of the SLTs about this. To sum up, IPE and IPCP is a new type of professional education system in the health and medical. With this system speech and language therapist have great opportunities for using and understanding of IPCP. Because they can work in health, education and social care system with in this entire place they connect with a lot of difference professionals. That’s why their perspective about IPCP is quite important for the literature.

The result of this study shows that the field of interprofessional education and interprofessional collaborative practice needs to some evidence base researches, which is about the speech and language therapist perspective. Another result of this literature searching research is that the literature of the speech and language therapy there is not enough mentioned about the Interprofessional education and interprofessional collaborative practice. On the whole Both literature and both area of research needs to more evidence base work and studies to shows the impact of the Interprofessional education and collaborative practice to speech and language therapy filed. The perspective and knowledge of the speech and language therapist about the interprofessional education and collaboration.

REFERENCES


Abstract—This paper focuses on the design and implementation of mobile device application for Speech Therapy on Android devices. The application is designed for the purpose of helping the therapist to improve their treatment session. Speech disorder, types of disorder and their therapy methods are also explained in this paper. Additionally, this paper outlines what kind of technology has been used until now for the speech disorders. The application is developed to provide a tool that is used for the therapy sessions. The application supports two languages, namely Turkish and English. It comes with English as a default language. The Turkish language package can be downloaded as an option. Therapists have the possibility to add their pictures and audio as well. The application also provides images and audio. The user can choose the language, letters and position of the word, which are initial, middle, and final. Then, the users can study the proper pronunciation of the word with this application. The application is evaluated by the therapist and non-therapist users to measure how the application is useful. We send the application to the users via the internet and asked them to fill in the questionnaire. Then we collect the answer and comment about the application. The result was useful to find some minor bugs and understand the user interface how is looking.

INTRODUCTION

Over the last decade, numerous mobile platforms have appeared in the market, such as Symbian, Windows Mobile, iPhone, Blackberry, Java Mobile Edition, Linux Mobile (LiMo) and more. Android is the one of the mobile platforms which have had a huge spike in market share. Android applications range from health care and entertainment apps to simple games and more. In addition, Android is one of the fastest growing trends in Information System Industry [1]. Taking into consideration the biggest market share and universal popularity of the Android OS, the Android platform is the best choice in terms of research and business. Android applications are used on the different smartphones and tablets around the globe [2]. As the improvement in every part of health care makes Speech Disorder Therapy easier and more accessible, the number of smart phone applications is becoming profitable. Correspondingly, it is now more common to see Articulation applications on the market, especially for IOS.

Nonetheless, it is difficult to find a well-designed articulation application on Android. The aim of the project is to introduce the android platform and develop an Android Speech Therapy application that enables Turkish Therapists to improve their treatment methods.

The application will be used as equipment for speech therapy methods by the therapist. This project will primarily engage in the research and development of the articulation application.

The main aspiration of the project is to design and implement an application for the Android platform, which will be used for articulation therapy in future work. In this area, there is no application in Turkish. In Turkey, Therapists are still doing therapy by using flashcards and audio devices. The goal of the project is to create an inclusive and flexible articulation program to practice pronunciation of Turkish and English phonemes of words and phrases. The application will contain hundreds of photos and pre-recorded audio stimuli. It assists speech therapists to use targeted words and to add patients’ own audio recordings. At the first step, the project will start in English or Turkish, and then it may be improved in other languages. The application will come with one language in it, and then preferably Turkish or English libraries could be downloaded.

BACKGROUND INFORMATION

This part defines briefly what is meant by speech disorder and Android. A key aspect of speech disorder is to distinguish between language disorder and speech disorder because the person who has a speech disorder can understand other people but not speak properly, failing to let out sounds. This chapter will clarify speech disorder and its levels. It will also give concise information about how our project constructs the Speech Therapy. Furthermore, Android and Android Architecture is explained in this chapter.

Speech disorder [3] is an articulation problem that is related to creating sounds while speaking. This view is supported by [4] who writes that language disorder is an illness that affects people’s talking. People who have a speech disorder, know what they want to say, but they have some difficulties with communicating and producing a voice [4]. It has been identified that the problems of children who have a speech disorder include spelling the incorrect sound, letting out sounds, adding sounds or distorting sounds. Several studies investigating problems of speech disorder have been divided between four main problems: stuttering, dylalia, dysphasia and other disorders [5]. However, [6] classification and description of language disorder may issue from the same aspects. Speech, language and literacy problems can be inherited [7]. In general, brain injury or strokes which damage some part of the brain related to language causes Aphasia.

Speech therapy can be divided into three main categories. They are Word Level, Sentence Level and Discourse Level and can be described as [8].

(i) Word Level: For this level, children are required to say or write the right word. Word level includes different levels from easy to difficult. In this case, students need to select a matching word from the list. An example of a complicated word matching is shown below.
Students are required to choose the position “Initial” and the letter and learn how to pronounce words that start with “B” then the users can decide some target words, and then use flashcards for the sound of vowels due to there not being a teaching method for vowel sounds. For the same reason our application will not include the vowel letters which are “A”, “E”, “I”, “U”, “İ”, “O”, “Ö”, “Ü”. We will cover the first five letters of the Turkish alphabet which come with the application, and upload them to the SD Card.

Our project will comprise word level treatment because it is the beginning of the treatment. Word level treatment consists of a set of images illustrating the words and these images can be used in a variety of ways in the therapy such as transforming, modifying, combining, and expanding. In the therapy, these images can be intoned, repeated, or combined as word level therapy. Besides, therapists raise awareness for demonstrating how a patient’s speech differs from the norm by using audio or video recording equipment and a mirror. Therapists decide some target words, and then use flashcards repeatedly in order to reduce failure. There are a variety of applications, tools and materials for speech therapy. All these applications address different types of disorder such as Aphasia. Some of them can be used for all levels of treatment such as Language TherAppy that is one of the iTunes applications, while others specialize in just one level like Conversation Cards, which is also an iTunes application. However, these tools support IOS systems and are quite costly. Furthermore, none of these applications or tools provide an option for Turkish language support. At this point, our application will support the Turkish language and run on the Android OS that has the highest rate of market share. The Speech Therapy does not cover the sound of vowels due to there not being a teaching method for vowel sounds. For the same reason our application will not include the vowel letters which are “Q”. This page also allows the user to add a new image and audio to the selected language package. After clicking the language button, the users can see the letter.xml page. This page allows us to choose the letter which the user wants to focus on. When the users make their decision about the position and the letter, then the users are able to see a b.xml which includes 5 or less than 5 pictures because some letters do not have many words such as Q. This page also allows the user to add a new image and audio to the selected language package. Next time, the users will see their images on this page. When the users choose one of the images from the picture.xml then the users will see that the picture on the fullimage.xml page. Fullimage.xml provides us with the proper pronunciation of the word and also give the chance to the user to record their voice and listen to their recording.

First of all, the users need to install the application on their Android device. When you touch the launcher icon, you will see a warning message. If the users choose “Yes”, it will start to copy the pictures and audios to the database. This is the first time the application has been run and it needs to install the default language data and database on the device. The application creates a folder for images, audios and user data on the SD Card. Then, it checks the SD card for any existing database. Otherwise, it creates a database and starts to upload the information of data to the database.

It takes a little bit of times to copy all the files. The application will also inform the user about the app program setup. In this step the application takes the pictures, which come with the application, and upload them to the SD Card.

This main.xml page contains language buttons which enable the users to initiate the application. After clicking the language button, the users reach the second page that includes three buttons: the initial, middle, and final. These buttons allow users to select the position of the letters whether at the start, in the middle or at the end of the word. This page is called an option.xml page in the code and it carries the intents, which have information about the user’s choices, to another page. The option.xml page brings the text from the database and writes the button texts dynamically. For example, if the language is Turkish, then these buttons will fill in the Turkish version such as Başta, Ortada, Sonda. When the user touches one of the positions, the application will start the next activity which is called letter.xml page.

When the decision is made about the position, users can see the letter.xml page. This page allows the user to choose the letter which the users want to study. This page consists of the letters so that users can make selections from among the letters to provide a word which contains the selected letter in a particular position. This page also still carries the intents to the next page because we need these intents while the querying the information from the database. The option.xml page brings the letter from the database and writes the button texts dynamically. This page checks how many letter the language has because the Turkish alphabet has 20 letters and the English alphabet has 21 letters (excluding the vowels). The application needs to create dynamic buttons.

The next page comprises of the images which are related to the letters that have been selected on the letter “B”. The letter.xml page allows us to choose the letter which the user wants to focus on. When the users make their decision about the position and the letter, then the users are able to see a b.xml which includes 5 or less than 5 pictures because some letters do not have many words such as Q. This page also allows the user to add a new image and audio to the selected language package. Next time, the users will see their images on this page. When the users choose one of the images from the picture.xml then the users will see that the picture on the fullimage.xml page. Fullimage.xml provides us with the proper pronunciation of the word and also give the chance to the user to record their voice and listen to their recording.
page b.xml, which includes the images, allows the user to add new items on this page and also chooses an image that the users want to study on it. On this page the user also can add his item which can be an image and audio. The user will see the add_library.xml page when the user touches the add button.

The second part of the application is adding a language package. This application comes with the default language being the English alphabet and users can also install the Turkish language alphabet. When the user downloads and installs the Turkish language package the launcher icon will be seen in the program list. The Turkish language package just allows the user to install the new language in the application database. The packaged application has just a simple interface. When the user starts the packaged application, the user will only see a button that allows them to install a new language into the database.

Figure 3 illustrates the main page of the application. Further details and demo will be illustrated during the presentation.

TESTING AND RESULT

To determine how the application reacts to different inputs, we tested our application on different devices, including ASUS MeMo Pad 10.1 and Samsung Tab 2 7.0, Samsung Galaxy S3 mini, and Android Virtual Devices in a variety of resolutions. The application performs as expected for the inputs.

The questionnaire was designed and administrated using the Google Drive. The Survey link was distributed via e-mail to the sample group. The questionnaire included thirteen relevant questions to examine the relationship between users and Speech Therapy application.

Device test results illustrate that the application supported minimum 5.1 inch screen. The application fulfils all expectations in the requirements part. The application can support pictures, allow the users to record their voice and allow the users to hear the proper pronunciation. The users can download additional language packages in any device. Pictures do not fit on screens under 5.1 inches and the application does not work on versions earlier than Android 2.2.

16 respondents answered the questionnaire. Half of them were non-therapists and the other half of them therapists. As described above, the questionnaire consists of four parts. The respondents who answered the first part of the questionnaire could easily download and install the application. However, participants, especially therapists, found the user interface somewhat complicated. In the second part of the questionnaire, we asked the participants
‘Does the application bring the solution of your problems that you face at your therapy?’ Non-therapist participants’ were not able to answer this question. However, therapists found the application very effective for their therapy. Therapists had used similar applications before, but our application is different from other applications. All participants preferred to install Android OS, just one person preferred the IOS to the Android OS. In the third part, it asked for the level of agreement about the application, how likely they would be to recommend it to a friend. Almost every participant gave point 5 or 4. In the last section of the questionnaire, we asked the participants to give an extra comment about the application. The respondents could not understand the user interface; they asked for extra direction for the application while they were using it, for clearer understanding. Also, they asked for some arrows for the navigation in the application. Furthermore, participants reported some bugs.

CONCLUSIONS

This part of paper concludes the whole process of the developing a Speech Therapy Application. This report contains background information about the Android OS and Speech Disorder, the progress of developing a prototype of a mobile application which can be used in the therapy. Additionally, the report evaluates the test results and questionnaire results.

This report gives an insight into developing an Android Application and specifically emphasizes how we can use the characterized Android Application in the Speech Therapy sessions. This report represents every step of development levels on how to each part of the project was conducted.

The implementation of these parts was demonstrated through the Speech Therapy Application and described in detail with the complete codes. The paper describes where problems occurred while implementing the application and which future work arises from this project. The project has comprised mobile application development on the Android Platform, essential components of Android applications and determined the overall architecture of the Android OS. The requirements of the project, risk and challenges and needs of the therapist helped formulate a suitable design and implementation phases. The result of the project has realised the aspirations that were introduced in the beginning. However, the scope was limited to, certain aspects such as time constraints. The thesis was limited with three main constraints. The first constraint was a time issue. Taking into consideration time limitations, the scope of the paper was set to demonstration a prototype Speech Therapy Application. Also, the time was limited to three months. The second constraint was the learning process of Android programming. It took time and getting used to the environment of Android SDK was an issue. The Android platform is not new, but is not very established either. Finding relevant code pieces required effort. The third constraint was understanding the speech disorder and finding the most frequently used therapy methods. As a result, we needed to work with a therapist. The application can demonstrate related word and audio to the users. The users can add their images and pre-recorded audio or record new audio alongside. The user can record their voice and also listen to this recording. However, the project could be enhanced to develop a fully functional application that contains different features optimized for better performance, such as comparison.

The project was completed as expected as a success for the Android operating system. Additionally, in future work, customers could create a new library themselves. Actually, we wanted to add this feature to the project, but there was not enough time to finish this part. Instead of this, the user can add images and audio to the existing language package. As further implementation, comparison features could be added after recording the audio. The application would make a comparison between newly recorded audio and pre-recorded audio. Then, the application would demonstrate the statistics of the score. Also, some exercises could be added to help people with speech disorders to improve.

ACKNOWLEDGMENT

The authors thank the second adviser who is Dr Irek Ulidowski for his advice and those who participate testing part.

REFERENCES

Investigating the amigdala-thread effects on cognitive learning via graphical domains and Bayesian networks

Serpil Ustebay, Ahmet Orun, Huseyin Seker and Ahmet Sertbas

Abstract - In this paper we investigate the effect of “thread” factor on cognitive learning process by use of virtual graphical domains whose parameters can be chosen in a systematic flexibility. Nowadays, spread of graphical environments in parallel with the technological and communication tools make investigation of learning factors more easier which are in association with such graphical media. In presented study, learning process analyzed by the proposed test involving some subjective graphical elements (fear, object type, etc.) and their stimulant links with “amigdala” the brain fear centre are investigated by artificial intelligence techniques such as Bayesian networks, whose subtle effects are normally not visible to naked eye.

INTRODUCTION

Graphics are often used in the field of education. In recent years, researchers have made various investigations on the success of graphic media used in education. Even though the results demonstrate an increase in educational performance by using graphical environments, the researches still keep investigating how the cognitive factors may become more effective on learning process in a structured graphical environment. Studies [9,10] examine the impact of colors and emotional expression in a graphic environment on learning. The links between these factors and learning were detected in terms of memorizing the related information. In the study introduced by Mayer at al. [9], it was discovered that the emotional multimedia design principles improve the learning of students. But they did not give detailed information about which emotional factors (e.g., thread, pleasure, sadness) affect the learning process. In the other study [10], shape and color effects on multimedia learning material were investigated. Their results showed that these factors can stimulate positive emotions in learners and facilitate learning, but there is insufficient information about which emotion is more effective on learning. To fill the gap in above works, firstly the objective of our work is to investigate which graphical factors (particularly thread based) may have impact on learning process in association with brain fear unit amigdala. Secondly, we use an original tool called Bayesian networks to unveil the subtle links between the graphical factors and cognitive learning process.

In this study, three different models have been exploited to identify fear factor and fear acquisition. This models are generally complementary to each others.

According to General learning model [2], it is assumed that a fear can be acquired on the condition of negative information transmission, instructions and observation. Whereas non-associative model refers to innate fears and prepared learning model, which states that people who were exposed to fear may have better memory performance. This is because of primary role of amigdala (fear centre of brain) in the formation and storage of memories associated with fear. The study introduced by Vanessa and DeLoache [5] examined the impact of fear factor on attention in association with a graphical domain. In their study the participants were asked to find threat relevant and threat-irrelevant pictures on 3x3 matrix. Participants who had snake fear found threat relevant pictures more faster than threat-irrelevant ones. In another study [7], fMRI results were demonstrated that people without snake and spider fear were affected negatively by fear-related stimulis than the control subjects. Comparing these results with regards to fear learning models indicate that evolutionary life-threatening fears can be detected by brain relatively fast [3]. In such studies the common stimulis used are snakes, scorpions, spiders, etc. Whereas we used similar threat stimulis in our works such as dragon, shark, scorpion, etc. The other characteristics of the tests is animacy, as animacy words and graphics have mnemonic values and remembered better than inanimate ones[21-23]. Important role of animacy on survival instinct has a priority effect on memory and easy to recall [23].

Limbic system is a set of brain structures associated with emotions and drives[16]. One of the structures of limbic system is amigdala which is called fear centre and interacts with hippocampus which is brain's memory centre[14] because of storing emotional stimulis in the memory [13-15]. Researchers particularly focused on fear factors because of increasing amigdala activation in fear related circumstances. In one of studies [4], four different categories of pictures chosen from the International Affective Picture System (IAPS) to investigate impact of threat factors on amigdala, with the help of fMRI method, an activity of amigdala was detected with the same intensity level of threat. In another study [6] that adapted stroop test used with threat factor, the participants with snake fear were asked to find the colors of presented pictures. Participants who have stronger snake fear gave late answer for snake pictures compared to other pictures. The reason of delay was linked to the activation of emotional memory when threat stimuli is effective. Our study aims to investigate the threat affect on cognitive learning with an indirect link with amigdala activity.

For the data analysis to unveil invisible links between the graphical test attributes like fear, live objects, learning, etc. We used Bayesian networks tools. Bayesian networks (BN) are “directed acyclic graphs” that exhibit knowledge presentation and reasoning. In our work two kinds of (General) Bayesian network tools are used, one is for classification purposes that is called PowerPredictor™ and the other is used as an inference tool called
PowerConstructor™ [23]. Both packages utilize “Markov condition” by investigating the amount of information flow between two attributes, then connect those attributes in the event of certain amount of information flow which is above the specific threshold. The earlier use of Bayesian networks in “cognitive modeling” was firstly introduced by Lee [24]. The other study which exploited BN to identify user identity in video game domains was introduced [26], where the video game attributes and physiological body signals were used in a combined form. Limbic system is a set of brain structures associated with emotions and drives [16]. One of the structures of limbic system is amigdala which is called fear centre and interacts with hippocampus which is brain’s memory centre [14] because of storing emotional stimuli in the memory [13-15].

Researchers particularly focused on fear factors because of increasing amigdala activation in fear related circumstances. In one of studies [4], four different categories of pictures chosen from the International Affective Picture System (IAPS) to investigate impact of threat factors on amigdala, with the help of fMRI method, an activity of amigdala was detected with the same intensity level of threat. In another study [6] that adapted stroop test used with threat factor, the participants with snake fear were asked to find the colors of presented pictures. Participants who have stronger snake fear gave late answer for snake pictures compared to other pictures. The reason of delay was linked to the activation of emotional memory when threat stimuli is effective. Our study aims to investigate the threat affect on cognitive learning with an indirect link with amigdala activity.

For the data analysis to unveil invisible links between the graphical test attributes like fear, live objects, learning, etc. We used Bayesian networks tools. Bayesian networks (BN) are “directed acyclic graphs” that exhibit knowledge presentation and reasoning. In our work two kinds of (General) Bayesian network tools are used, one is for classification purposes that is called PowerPredictor™ and the other is used as an inference tool called PowerConstructor™ [23]. Both packages utilize “Markov condition” by investigating the amount of information flow between two attributes, then connect those attributes in the event of certain amount of information flow which is above the specific threshold. The earlier use of Bayesian networks in “cognitive modeling” was firstly introduced by Lee [24]. The other study which exploited BN to identify user identity in video game domains was introduced [26], where the video game attributes and physiological body signals were used in a combined form.

**EXPERIMENTS**

During the analysis, the following experiments were carried out and corresponding results obtained;

**Experiment 1**: In Experiment 1, the major target is to investigate the graphical factors whose attributes may have supportive effect on cognitive learning process. Based on the above literature review, we focused on threat factors and live objects initiated by the graphical domains. For the tests 70 healthy participants amongst the students from the University campus were used. For this study a graphical test was prepared (Figure 4) including 15 questions and each question has two answers. Test questions are independent of each others in subject but their sequential cause-effect structure is expected to support a progressive learning. Each participant was informed at the beginning of test that how test cycle will process. At the first stage each participant sees test question on the screen in a graphic mode with two answering choices. At the second stage the participant selects only one answer and then see the correct answer on the screen. By this stage each participant has an opportunity to compare his/her answer with the real one, so that he/she can make links between the graphical test attributes in mind, which meanwhile lead to cognitive learning. The procedure is repeated for all 15 questions.

**Data set specification**: Each answer of participant was recorded as 1 if it is true, otherwise 0. All the answers are formed in a vector called *User’s Graphic Response Vector* (Vg). Any transition from 0 to 1 at Vg is assumed that cognitive learning occurs. In any opposite transition (from 1 to 0) we assume reverse learning is in question. The main attributes associated with the test are; threat factor, live object, learning progress (lrn_progress) and learning backward (lrn_backw). The further explanation of each attribute (also shown in Table 1) are as follows;

- **Threat**: This attribute can be '1' or '0'. where “1” refers to test question containing a threat factor and “0” means question does not have a threat factor. Threat factor can be scorpion, dragon, shark, etc.
- **Live object**: An object which represents a creature like shark, gorilla, etc.
- **lrn_progress**: Progressive learning refers to transition from 0 to 1.
- **lrn_backw**: Contrary to lrn_progress, negative learning refers to transition from “1” to “0”.
- **Graph_no**: Indicates the number of test question.

The data sets formed after each test were used as input for Bayesian network utility to build the semantic networks, whose attributes and connections are displayed in a graphical form (Figures 1, 2 and 3).

<table>
<thead>
<tr>
<th>Question_No</th>
<th>Threat</th>
<th>Live object</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Having carried out the experiment, the following results were obtained.
Results 1: Bayesian Network method was used to define relationship between attributes the attributes “thread”, “live object” and “learning” (lrn_progress_d ) contained within the test. Figure 1 shows the results of automatically established links between these attributes. The graphical results show the direct relationships between the threat and learning and live object and learning.

The second network representing the test results of adult group is shown in Figure 3 where there is no direct link found between the threat and learning. However, a direct link is established between live object and learning factors. The findings appear to suggest that the threat factor is only indirectly effective on learning process but in association with the live object.

Experiment 2: Its aim is to compare between two age groups by investigating how the attributes thread and live objects are separately or jointly effective on each group. In the experiment “age of participant” is the main factor to focus on. Here the participants are divided into two different age groups. First group represents young participants whose mean age is 22. Meanwhile second group represents adult category whose mean age is 29. Both groups contain 20 person in total. The materials like Bayesian tools and graphical test used here were exactly the same ones with Experiment 1.

Results 2: There are two sorts of Bayesian network built and each corresponds to graphical test results. First one is shown at Figure 2 which represents the young group. As is seen there is a direct link between thread, live object and learning factors. It is shown that thread and live object factors are separately effective on learning process of the young age group.

CONCLUSIONS
The results of experiments show that there may be direct link between the cognitive learning process and thread effects in association with amigdala the brain fear centre, whose function is the formation and storage of memories associated with fear or other emotions. Threat factor is one of the emotions which affect people’s attention stimulating them to focus more quickly[5,6]. Our experiments have strengthened that during the cognitive learning process, especially increasing activity of amigdala under stress and

Figure 1.Bayesian network which corresponds to Experiment 1 results, where direct links between the attributes “thread”, “live object” and “learning” (lrn_progress_d ) which is class node are displayed. (Due to large volume of the network, only class node connections are shown here)

Figure 2. Bayesian Network which belongs to young age group of participants

Figure 3. Bayesian network which belongs to adult age group of participants

Figure 4. A sample test frame which contains “thread” factor (valued as thread =1 and live object =2 in Table 1) Each frame also contains a question which is not displayed here.
threat causes the release of stress hormones which perform transmission of information to the memory [19]. One of the results of this study showed amigdala’s positive effect on cognitive learning [20]. The experiments including ones within this work have also proven relationship between age and learning. This can be explained by late maturing of amigdala. During the maturing progress of amigdala, it provides better identification of treat and fear [17]. The other supportive work is about the age related differences of amigdala’s functions which may have a possible link with learning process [18]. We impose different threat factors in the test, not only live objects like shark, scorpion, etc. but also fear conditions like standing on a broken chair, fire damage, etc. These conditions may have different effects on young an adult participants. Young ones can perceive more threat than adults because of their different amigdala characteristics [17]. This was proven by our experiment whose results are shown in Figure 2.

REFERENCES

10. Steffi Heigid, Elizabeth O. Hayward, Bruce D. Homer,Enjoon Um
14. Dr. Timuçin Oral ITICÜ 15 Aralık 2009 Limbik Sistem ve Emosyon ders notları
15. Prof Dr Süheyla ÜNAL Limbik Sistem Ders Notları
18. Lau JV; Britton JC; Nelson EE; Angold A; Ernst M; Goldwin M; Grillon C; Leibenluft E; Lissek S; Norcross M; Shiffman N; Pine DS: Distinct neural signatures of threat learning in adolescents and adults. Proc Natl Acad Sci USA 2011; 108:4500–4505
Comparison of EEG Devices for Eye State Classification

Oliver Roesler*, Lucas Bader, Jan Forster, Yoshikatsu Hayashi, Stefan Heßler, David Suendermann-Oeft

Abstract—In this paper, we investigate whether the price of an EEG device is directly correlated with the quality of the obtained data when applied to a simple classification task. The data of three different devices (one medical and two consumer) was used to determine the eye state (open or closed). For classification, 83 machine learning algorithms were used on the raw EEG data. While the cheapest device performed extremely poor with only one classifier better than the majority vote the two other devices achieved high accuracy. The lowest error rate for a more expensive consumer EEG was 1.38% and produced by KStar. For the medical device the best performing classifier was IBk which achieved an error rate of 1.63%. Except for KStar, the classifiers achieved a lower error rate by the medical EEG measurement system than the consumer EEG system.

I. INTRODUCTION

There are many possible applications which could use brain activity measured by electroencephalography (EEG) as an input mode. For instance, the use of brain waves to control computer games [1], track emotions [2], provide handicapped people with an alternative input mode [3] and for different military scenarios. [4]

There are several important points that have to be taken into consideration when thinking about the use of EEG signals to control applications:

1) the accuracy for the detection of a certain state from which the control commands are generated
2) the speed of the detection algorithm
3) the cost of the EEG measurement device
4) the usability which includes preparation time and restrictions for the user during the use of the device.

One possible input for binary control tasks could be the state of the eyes that is whether they are open or closed. In a previous study [5], it was shown that the eye state can be predicted with a consumer EEG with high accuracy of 97.3%. The advantage of a consumer EEG system in comparison to a medical EEG would be the lower price and the higher usability. Most of the consumer EEGs can be easily set up and do not restrict the movement of the user by a wired connection to amplifiers.

Yet, although the results were promising, there were several problems. First of all, the study [5] was conducted with only one participant. Thus, it was not clear whether the results were statistically significant. Second, there was no comparison to the accuracy achieved when using a medically approved EEG. Despite of the fact that an error rate of 2.7% can still be too high for less fault-tolerant applications, the problem with the best classification algorithm is, as already pointed out in Roesler’s study, the slow classification speed. The slow classification is caused by the fact that KStar is an instance based learner [6]. This means that the algorithm classifies a new instance by comparing it to a database of previously classified instances which makes the classifier unusable for online classification and thus for the use in any real-time application.

Thus, even if a professional EEG device would only increase the performance of other classifiers, it could be a necessary investment when the goal is to use the EEG signal as an input signal for real-time applications.

The paper is structured as follows: Section II provides the detailed information about the conducted experiment, the EEG devices and the obtained corpora. In Section III the machine learning algorithms used for classification are described. Then, the classification results for each device are described. Finally, Section VII draws conclusions and outlines possibilities for future work.

II. MATERIALS AND METHODS

A. Stimuli and participants

The experiments were done in the same way as described in [5]. All experiments were conducted in a quiet room. During the experiment, the face of the participants was recorded. The experimental procedure was specified as follows:

1) After placing the electrodes on the scalp the participants were told to sit relaxed, face the camera and change the eye state at free will after clicking the start button.
2) The task was repeated one to two times after a resting period of one minute.

Additional constraints given to the participants were that the individual eye state intervals should vary in length and the duration of both eye states should be about the same when accumulated over the entire session. The participants were not aware of the fact that the first 20 seconds of the measurement were not recorded. This was done to prevent artifacts due to the clicking on the start button and initial movement to face the camera.

B. EEG Measurements

The duration of the measurement was 140 seconds. Yet, as described in the previous section the initial 20 seconds were discarded to prevent artifacts. Most of the eye states were automatically annotated during the measurement by the video

*Corresponding Author (e-mail: oliver@roesler.co.uk).
Oliver Roesler and David Suendermann-Oeft are with the Baden-Wuerttemberg Cooperative State University (DHBW), Stuttgart, Germany (e-mails: oliver@roesler.co.uk and david@suendermann.com).
Lucas Bader, Jan Forster and Stefan Heßler are with the IBM Deutschland MBS GmbH, Ehningen, Germany (e-mails: {lucas.bader, jan.forster, stefan.hessler}@de.ibm.com).
Yoshikatsu Hayashi is with the University of Reading, School of Systems Engineering, Reading, UK (e-mail: y.hayashi@reading.ac.uk)

David Suendermann-Oeft is now with ETS, San Francisco, USA
recording program. Only frames which the program could not classify due to bad lighting conditions where later annotated by hand. The used camera was a simple webcam.

C. EEG devices

Three different EEG devices were used for the measurements. Due to the high accuracy achieved with the Epoc in [5] the MindWave[1] from NeuroSky was selected to see whether a similar accuracy can be achieved with an even simpler and cheaper device. Secondly, the Epoc[2] from Emotiv was used to verify the generalizability of the results and to represent a more expensive consumer EEG. And finally, the BrainAmp Standard[3] from Brainproducts was investigated which is a medical EEG and needs specific training to be set up. Table I shows a comparison of some of the specifications of the individual devices.

<table>
<thead>
<tr>
<th>Device</th>
<th>available channels</th>
<th>used channels</th>
<th>sampling rate</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>MindWave</td>
<td>1</td>
<td>1</td>
<td>512Hz</td>
<td>80$</td>
</tr>
<tr>
<td>Epoc</td>
<td>14</td>
<td>14</td>
<td>128Hz</td>
<td>700$</td>
</tr>
<tr>
<td>BrainAmp</td>
<td>32</td>
<td>12</td>
<td>1kHz</td>
<td>60,000$</td>
</tr>
</tbody>
</table>

**Table I: EEG device comparison**

Due to the fact that the MindWave and the Epoc have different fixed electrode positions a direct comparison of the quality of a single electrode is not possible. Therefore, the difference in the number and position of the electrodes was not specifically assessed when evaluating the experimental performance. Thus, the number of electrodes of the BrainAmp Standard was constrained to 12 of the 32 possible electrodes to reduce the setup time and size of the corpora. The 12 electrodes were placed at the frontal, central and temporal positions of the 10-20 system. [10] This was done because most electrodes of the Epoc are frontal and the MindWave has only one frontal electrode. A detailed overview of the electrode positions is given in Figure 1.

D. Corpora

Six corpora were recorded for each device. The corpora were from up to four different probands (see Table III). All corpora consist of as many attributes as the number of electrodes of the corresponding EEG device plus the binary class attribute. How many instances correspond to one class varies between the corpora because the probands decided how long they opened or closed their eyes. The percentage of instances belonging to the eye open state varied between 18% and 63%. On average 44.7% of the instances in the MindWave, 39.3% of the instances in the Epoc and 44.4% of the instances in the BrainAmp Standard datasets belonged to the eye open state.

III. MACHINE LEARNING ALGORITHMS

For classifier testing, the Weka toolkit [11] version 3-6-11 was used. Ten-fold cross-validation was carried out for all 83 standard classifiers of the toolkit with default settings. Yet, only 50 were running without an error on the datasets from all devices. Some classifiers were not suitable for cross-validation while others had problem with negative attribute values of the BrainAmp Standard datasets. Although, it would have been possible to preprocess the data to avoid the errors. The goal was to evaluate the performance on the unprocessed raw data and the remaining 50 classifiers were still sufficient to provide a general picture about the quality of the obtained EEG data. All classifiers were applied to each corpus separately. The mean error rate was then calculated across all corpora of a device.

IV. CLASSIFICATION RESULTS

A. MindWave

For the MindWave the overall number of classifiers which made a majority vote was 20. 29 of the remaining 30 classifiers were performing worse than majority vote and only one classifier (Conjunctive Rule) achieved an accuracy of 43.52% which was slightly better than majority vote of 43.66%.

These results clearly show that the MindWave headset is not suitable for even a simple task as the classification of the eye state. However, this is not necessarily due to the fact that it has
only one sensor. The Epoc headset still gets less than half the error rate when the classifiers are only trained with the values of electrode F4. Thus, the bad accuracy is also due to the quality of the electrode. To sum up, the MindWave does not seem to be useful for serious EEG experiments or to control an application.

B. Epoc

The results for the Epoc in [5] were already quite promising. Yet, the results on the six corpora have shown an even lower error rate for most classifiers. KStar [6] is again the best performing classifier with a mean error rate of 1.38%. This is a relative reduction over the result reported in [5] for KStar with default settings of 56%. IBk [12] achieved with 2.66% the second lowest error rate. Thus, the results are consistent with [5] in that instance based learners achieved the lowest error rates.

Close to them is RandomForest [13, p.407] with an error rate of 3.83%. In contrast to the instance based learners, RandomForest could be suitable for online classification. All other classifiers which achieved an error rate below 10% were mostly decision tree algorithms like FT, J48 or REPTree. These algorithms could also be suitable for online classification if the error rate can be decreased with the help of parameter tuning.

The average error rate over all classifiers (excluding classifiers which only achieved majority vote) was 16.5%.

C. BrainAmp Standard

For all classifiers except one the BrainAmp Standard achieved a lower error rate than the Epoc headset. In contrast to the Epoc headset the instance based learner IBk achieved the lowest error rate with 1.63%, followed by an error rate of 1.72% achieved by KStar. As previously mentioned, instance based learners are in general not suitable for online classification due to their slow classification speed. KStar took nearly two weeks for ten-fold cross-validation on one dataset from the BrainAmp Standard on a system with Ubuntu 12.04.1 LTS, QEMU Virtual CPU version 0.15.1, quadro core with 2GHz each and 32GB RAM.

KStar and IBk were followed by FT [14] which is a decision tree algorithm and which was also the third best classifier for the Epoc device. However, the fourth place was taken by the MultilayerPerceptron [15] algorithm which achieved a more than five times lower error rate for the BrainAmp Standard. Also the second neural network algorithm, VotedPerceptron, performed nearly three times as good on the BrainAmp Standard data. This clearly shows that most of the algorithms which require extensive training but provide fast classification were performing better on the BrainAmp than the Epoc data.

The average error rate over all classifiers (excluding classifiers which only achieved majority vote) was 11.1% and therefore 5.4% lower than the average error rate of the Epoc headset. Which is a relative reduction of the error rate of 33%.

V. CONCLUSIONS AND FUTURE WORK

The results of the experiments with the three different EEG devices have shown that the MindWave equipped with only one sensor cannot be used for eye state prediction. Yet, the more expensive consumer EEG device the Epoch headset shows very high performance for eye state prediction. Nevertheless, the quality of the data of a professional EEG like the BrainAmp Standard is still higher than the Epoch indicated by the lower average error rate. This can make a significant difference when using the device for online classification since several fast classifiers had a much lower error rate on the data from the BrainAmp than for the Epoc data. Additionally, a professional device offers the
possibility to freely choose the number and positions of the used electrodes. This could also improve the performance due to the selection of electrodes close to more influential brain regions for the task of eye state prediction. Yet, the advantages of the Epoc in comparison to the BrainAmp Standard are the significantly lower price and the higher usability due to a wireless connection and a faster and easier setup.

Due to the fact that the Epoch and the BrainAmp Standard performed quite similar on the investigated task, further work will focus on more complicated tasks to see whether the difference in price will then be more obviously represented in the classification performance. Another step will also be to classify the eye state in real time and while the participant is carrying out other activities like moving around. This would show whether an EEG can be used successfully to determine the eye state in a real environment in which the participant will most of the time not be able to sit relaxed on a chair without moving. And further investigations will also explore whether some electrode positions are more important than others for eye state prediction in which case the number of sensors could be decreased which would simplify the set up and could lead to the development of a high quality EEG device with only a few electrodes and a wireless connection to a computer to evaluate the data in real-time.

VI. ACKNOWLEDGMENTS

Our gratitude goes to all involved participants.

REFERENCES

A First Step Towards Binaural Beat Classification Using Multiple EEG Devices

Jan Forster*, Lucas Bader, Stefan Heßler, Oliver Roesler and David Suendermann-Oeft

Abstract — This study analyzes the influence of binaural beats, an acoustic phenomenon which rises from two sonic waves of insignificantly different frequencies noticed as a single tone on each ear, on brain activity. Although this topic is not new, few scientific reports have been released. Therefore this study investigates if there is a measurable impact on the EEG activity for different frequency bands using a low budget, a mid-price ranged and a medical EEG device. Probands were exposed to a 12 minute 40 seconds counting task while listening to ambient music. During this task binaural beat samples were played covering different frequency bands, each followed by a pause of 10 seconds where no binaural beat sample was played. The probands were not informed about the presence of binaural beats, the only information given was to count the red shapes moving over the screen. The aim was to classify to which particular frequency the proband was exposed by analyzing their brain activity. The best result was achieved with the medical EEG by the FT classifier with a mean absolute error of 7.14% averaged over 4 datasets. In comparison the best result achieved on the NeuroSky Mindwave headset using the Naive Bayes classifier was a mean absolute error of 26.04%, while introductory experiments with the Emotiv EPOC headset showed a mean absolute error between 2.83% and 27.70%. Further experiments have to be conducted to validate the Emotiv EPOC classification performance. This shows that it is possible to classify the currently played binaural beats frequency rather accurate with appropriate hardware.

I. INTRODUCTION

In recent years the measurement of vital parameters such as skin temperature, blood pressure or more complex features like EEG activity has become a focus of the industry and made available to a wide range of consumers for example in commercial sports devices [1]. There are still a lot of areas where no research for the utility of EEG devices has been conducted. One of them being the influence of binaural beats on brain waves and brain activity [2], which is to be researched in this paper.

Binaural beats can have influence on the emotional states of a person. Nowadays devices for binaural beats are sold worldwide with advertised benefits like increased focus and concentration by listening to the respective binaural beats regularly1. Identifying binaural beat samples or frequencies that demonstrably enhance these emotional states can support people to deal with daily activities and improve health care processes [3,4].

Although there has been some research in the past [5,6,7], there was no recent publication which investigated the influence of binaural beats on EEG activity in a detailed manner for all frequency bands. Frequency bands are used to divide and analyze EEG activity, although the determination of the frequency bands is discussed in [8,7] compared the influence of beta and theta frequency bands on the mood. However, authors did not use the full frequency for analysis, but focused on two frequency bands. In [9] the influence of binaural beats to treat anxiety is researched, but the results are based on a self-report of the probands without the usage of machine learning methods. There are still a lot of areas where no research for the utility of EEG devices has been conducted. One of them being the influence of binaural beats on brain waves and brain activity, which is to be researched in this paper.

As no standardized test method for measuring EEG activity exists to the best of our knowledge, we want to propose a method and a set of binaural beat samples to set up a base for further experiments. The second objective of this work is to provide evidence using scientific methods how binaural beats can influence EEG activity using multiple EEG devices. Further we propose the basis for a classification of mental states to determine the specific influence of certain binaural beats. To achieve this, the influence of all frequency bands, each covered by a single binaural beats sample, is analyzed for a corpus, containing data of multiple probands to try to achieve universally valid results. If this study or future studies can demonstrate noticeable influence, there would be a wide range of application.

The paper is structured as follows: Section II gives details about the binaural beats used in this study and how the EEG measurement was done. In section III classifiers are discussed. An analysis of the results is done in section IV. The last section draws conclusions and outlines possible future work.

II. MATERIAL & METHODS

A. Probands

Five probands were performing the experiments in a quiet room which was illuminated by natural light. Probands were told to perform a counting task while their attention level was allegedly measured, in front of a laptop screen, where moving shapes in different colors could be seen. The actual
background of the experiment was not revealed. Probands were asked to count the number of red shapes displayed. The experiment software is our custom development and is publicly available\(^2\). During the experiment the proband listened to ambient music through phones without knowing that binaural beats will be played after a specific amount of time.

\[ B. \text{ Binaural Beats} \]

The experiment duration was 12 minutes and 40 seconds. At the beginning a 2 minute sample of ambient noise was played without binaural beats. This ambient noise sample was played during the whole time frame of the experiment. After 2 minutes, delta (0,1 - 4 Hz) binaural beats were played additionally. Followed by a 10 second pause, where only the ambient noise was played, then transitioning to theta (4 - 8 Hz) waves. This was repeated for alpha (8 - 13 Hz), beta (13 - 30 Hz) and gamma ( > 30 Hz) binaural beats, each separated by a 10 second pause to avoid artifacts. To minimize the impact of artifacts in the data and to get a larger universe each proband had to do the experiment two times, with a break of 5 minutes in between. The binaural beat samples are publicly available\(^3\).

\[ C. \text{ EEG Measurement} \]

To conduct the experiments, different EEG devices were used to investigate whether they cause performance differences. The devices that were used are NeuroSky Mindwave headset\(^4\), Emotiv EPOC headset\(^5\), as well as the medical EEG device Brain Products BrainAmp\(^6\). The two former are affordable, mobile consumer EEG devices and can be set up by naive users. The latter has to be set up by trained personnel. We measured raw data of the EEG activity at a 512 Hz sampling rate for the NeuroSky Mindwave headset, at a 128 Hz sampling rate for the Emotiv EPOC headset and at a 1024 Hz sampling rate for the BrainAmp EEG device. Positioning and the channel naming for the Emotiv EEG (green) & NeuroSky Mindwave (blue) sensors are depicted in figure 1.

We used the open-source software Weka\(^7\) for our classification experiments to create reproducible results for the scientific community.

\[ D. \text{ The Corpus} \]

The corpus for the Mindwave experiment consists of multiple datasets of altogether around 650,000 instances with 6 attributes (5 representing the frequency band data from delta to gamma and one representing the frequency class) in chronological order. The values are taken from raw data of a single electrode that are split up into the respective frequency band information based on fast fourier transformation.

\[ III. \text{ Algorithms} \]

The classification of our experimental data was carried out with 39 suitable classifiers from the Weka toolkit. We used classifiers from established categories including statistical-, function-, tree- and rule-based classification. The classifiers were used in default settings and with ten-fold cross-validation.

\[ IV. \text{ Results} \]

The results of the experiments averaged over all corpora are shown for the NeuroSky Mindwave and medical EEG in figure 2. The mean absolute error for the stratified results varied between 26% and 28% for the NeuroSky Mindwave. The best results were achieved by the standard classifier NaiveBayes, followed by tree- and rule-based classification techniques such as RandomTree or Ridor. The best result of all test sets was achieved by NaiveBayes with a mean absolute error of 26.1%.

The analysis of the BrainAmp medical EEG device resulted in a mean absolute error between 4.74% and 27.68%. The best results were achieved by the Ridor classifier considering the

---

\(^2\) https://github.com/lcsbdr/binaural_beat_classification
\(^3\) http://jetcityorange.com/binaural-beats/
\(^4\) http://store.neurosky.com/products/mindwave-1
\(^5\) http://www.emotiv.com/epoc.php
\(^6\) http://www.brainproducts.com/
\(^7\) http://www.cs.waikato.ac.nz/ml/weka/
best performing data set, whereas FT achieves a mean absolute error of 7.14% averaged over all data sets.

Figure 2. Performance of all classifiers with default settings.

Since the NeuroSky Mindwave EEG headset produces one-dimensional measures, the corpus is relatively sparse. NaiveBayes, which is the most tolerant tested classifier to both data sparseness and noise, was expected to perform better than others in this setup [10]. However, the differences are only marginal over all tested classifiers. The classification of data produced by the medical EEG has more dimensions and a higher frequency, and therefore was expected to provide better results. Indeed the best classification algorithm outperforms the Mindwave NaiveBayes classification by 19%, making it almost four times as efficient. Introductory experiments with the Emotiv EPOC headset were giving results in the range of the NeuroSky Mindwave headset, with few classifiers like IB1 [11] performing extraordinarily well, and even better than the best results of the medical EEG. Therefore, extensive research is needed for the Emotiv EPOC headset to determine whether it can provide the same results regularly.

V. CONCLUSION

We conducted extensive experiments using the medical EEG equipment, as well as the Mindwave EEG headset device. The results reveal that medical EEG equipment is likely to be suited to classify frequency bands according to influence of binaural beats. The Mindwave EEG headset device does not produce results precise enough to be considered for online classification applications. Further introductive experiments with the Emotiv EEG headset device hint that this device, combining mobility with relatively low classification error rates, might be suitable for such application scenarios.

This study shows that the single electrode of the NeuroSky Mindwave headset isn’t sufficient to demonstrate whether EEG activity is influenced by binaural beats and to classify brain activity according to binaural beat influence. Using the more complex EEG devices the influence of different EEG frequencies, which have been detected by numerous researchers in the past (1970-1980), summarized in [12], could be verified.

ACKNOWLEDGMENT

The authors thank the lab at the University of Reading to use the BrainAmp medical EEG with the special permission of Yoshikatsu Hayashi, as well as all probands.

REFERENCES


Multiclass Disease Identification Employing Functional Protein Microarrays
Patrick Kalmbach and David Suendermann-Oeft

Abstract — In this paper, we tackle a five-class problem for distinguishing Alzheimer’s Disease, Parkinson’s Disease, Breast Cancer, Multiple Sclerosis, and healthy control patients. For this purpose, we used the raw microarray data from a publicly available corpus. We tested performance of 25 base learners, investigated the influence of feature selection on classification performance, applied parameter tuning and boosted performance even further using ensemble learning. We found AdaBoostM1 with J48 (open-source implementation of C4.5) as base learner performing best, with a classification error rate of 8.8%.

INTRODUCTION

Diseases such as Alzheimer’s Disease, Parkinson’s Disease, Multiple Sclerosis, and Breast Cancer affect large parts of the population. Early detection of these diseases remains a problem despite huge efforts and successes. Often, a disease is diagnosed based on visible symptoms only, at which point the patient may face substantial damage already. Early detection and treatment, however, can help to slow down the progress of the disease, help mitigate symptoms and improve the overall live quality of patients.

It has been shown, that autoantibodies in human blood are reliable biomarkers for identifying these diseases [1-3]. By using protein microarrays, the authors were able to identify small sets of antibodies differentially expressed for patients with one of the respective diseases. This knowledge can be used to develop new drugs and therapeutic methods.

The selected antibodies allowed for a very accurate identification of the respective disease. However, they were not the only antibodies differentially expressed. Restriction to this small subset, information contained in the remaining antibodies cannot be used. Furthermore, in [1-3] RandomForest [4] was used as sole classifier to establish the predictive accuracy of the selected antibodies.

In this paper, we tackle a five-class problem distinguishing Alzheimer’s Disease, Parkinson’s Disease, Multiple Sclerosis, Breast Cancer and healthy controls. We present results of different base learners, investigate the influence of feature selection on classification performance and improve classification performance further utilizing ensemble learning.

To allow the scientific community to reproduce the experiments described in this work, we chose a publicly available corpus and used the open-source software kit Weka², which had previously been used successfully in context of gene expression microarrays [5]. Required changes to handle microarray data with Weka were implemented in Java. This code is also made publicly available³.

The remainder of this paper is structured as follows: In Section II, we briefly discuss the corpus and evaluation methodology used in our study. In section III, we report experimental results before drawing conclusions in Section IV.

MATERIALS AND METHODS

Materials and Methods

The corpus used in this work was first described in a study by Han et alters in [1] and is publicly available on the MIAME-compliant NCBI GEO database [6] under the accession number GSE29654. The corpus consists of microarrays of 159 patients. The detailed composition of the corpus is given in Table I.

<table>
<thead>
<tr>
<th>Table I. Detailed Demographic Information of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>XVI. Parkinson’s Disease</td>
</tr>
<tr>
<td>XVII. Alzheimer’s Disease</td>
</tr>
<tr>
<td>VIII. Multiple Sclerosis</td>
</tr>
<tr>
<td>XIX. Breast Cancer</td>
</tr>
<tr>
<td>XX. Older Controls</td>
</tr>
<tr>
<td>XLI. Younger Controls</td>
</tr>
<tr>
<td>Controls</td>
</tr>
</tbody>
</table>

The authors of [1] used ProtoArray v5.0 Human Protein Microarrays from Invitrogen to identify autoantibodies in human sera. Each microarray contained 9,486 unique human protein antigens to identify autoantibodies. Proteins are printed in duplicate providing 18,972 data points.

When analyzing the results with the freely available analysis software ProtoArray Prospector v5.2⁴ from Invitrogen, only half of the data points are used [7]. We decided to utilize all data points, therefore basically doubling the corpus to 318 instances.

To not miss any information, we did not apply any feature normalization method [8-9] as done in previous work. Instead, we used the raw signal derived from the difference of foreground and background intensity available from the aforementioned database.

Weka

Weka [10] is a popular open-source toolbox for data mining offering a large variety of state-of-the-art classifiers.

1 http://suendermann.com/diseaseClassification
2 http://www.cs.waikato.ac.nz/ml/weka

Patrick Kalmbach is with Hewlett Packard, Boeblingen, Germany (email: Patrick.Kalmbach@live.com), David Suendermann-Oeft is with ETS, San Francisco, USA. At the time this work was conducted, he was with DHBW Stuttgart, Germany. (e-mail: david@suendermann.com)

1 http://www.cs.waikato.ac.nz/ml/weka
As the corpus is rather small in size, we decided to run leave-one-(patient)-out cross-validation. In this technique, a model is trained on all instances except for those of a given patient. These instances are used for testing. This procedure is repeated for all patients, and the results are finally consolidated.

In our case, the training set for each fold consists of 316 instances and the test set of two instances from one patient. This separation by patient is crucial, since the results would else be overoptimistic. Using this method no microarray data from the individuals is in the training and test set at the same time.

**EXPERIMENTS**

**Default Options**

**Baseline**

First, we investigated how different classifiers behave on the problem at hand. To this end, we used 23 base learners provided in the Weka and applied leave-one-out cross validation, as described above. The results are depicted in Figure 1. All 9,486 autoantibodies were considered, and the default settings provided by Weka were used for each classifier.

We found that function- and tree-based learners performed best with a classification error rate of as low as 12%. Rule-based learners did not perform as well with the lowest classification error rate of 20%. Bayesian and lazy learning schemes performed worst with classification error rates between 30% and 63%.

The poor results for lazy learning schemes, is somewhat unexpected as those schemes often provide good performance. Other popular learners like Bayes [11], Support Vector Machines [12], C4.5 (J48) [13], Logistic Regression [14] or JRip [15] perform well for this problem.

**Feature Selection**

Inspired by previous work [1], we used attribute selection by information gain in order to investigate the influence of selecting different antibodies on the performance. We used a ranker discarding all features whose information gain is below a certain threshold. Results of these experiments are shown in Figure 2.

![Impact of Information Gain](image)

Figure 2. Dependency of classifier performance on feature selection by information gain of five exemplary base learners

The classification error of rule- and tree-based learners remains relatively unaffected by information gain between 0 and 0.5 after which the error rate start to increase. Function-based, lazy and Bayesian learners show a higher sensitivity towards feature selection. However, most of them have their minimal classification error between an information gain of 0 and 0.75.

These results are reasonable as rule- and tree-based learners of the use information gain themselves to decide when to split or create rules. The more features are taken away, however, the more this selection process is affected.

**Parameter Tuning**

**SimpleLogistic**

SimpleLogistic is a classifier for building linear logistic regression models. In order to fit the logistic models, SimpleLogistic uses LogitBoost [16] with simple regression functions as base learners. SimpleLogistic produces the best baseline with a classification error of less than 12%. Using a grid search, we tuned the beta value for weight trimming (W) for LogitBoost and the parameter for early stopping of LogitBoost based on heuristics (H). The tested values are W
where zero means no weight trimming and $H = 0, 25, ..., 100$ where zero means no heuristic stop. The results are depicted in Figure 3. The number of LogitBoost iterations is cross-validated. However, by setting the parameter $H$ to a value greater than zero, LogitBoost will terminate, if no new error minimum has been found in the last $H$ iterations.

We found that parameter $H$ did not have any effect at all on the performance of SimpleLogistic. In contrast, using weight trimming has a strong negative impact on the classification error. Assumptions made in [14] hold not true for this specific problem. In general, weight trimming can greatly improve computation performance without loss of accuracy. This is achieved by omitting attributes with very small weight.

**Ensemble Learning**

**Bagging** Bagging [17] is a popular ensemble learning scheme utilizing multiple versions of a single base learner combining them by majority vote [10]. The implementation of bagging we used in this work has two main parameters with potential impact on classification performance. The first, being the number of iterations ($I$) and the other the size of the bag in percent of the training set. We used again a grid search with $I = 10, 20, ..., 100$ and bag size of $P = 25, 50, ..., 100$. The results are depicted in Figure 4.

The minimal classification error with 11.94% was achieved with 100 iterations and a bag size of 50%.

**RandomForest** In previous studies [1-3], excellent results were achieved using RandomForest [4]. RandomForest is also an ensemble learning scheme using RandomTree as base learner. In our experiment, we grew an unlimited tree and changed the number of attributes for the random selection $K = 0, 10, ..., 100$. The results are depicted in Figure 5.

The minimal classification error we achieved was 16.04% with $I = 80$ and $K = 0$. RandomForest thus performed worse than bagging, boosting, and SimpleLogistic.

**AdaBoostM1** AdaBoostM1 [18] is an ensemble learning scheme utilizing boosting. Being related to bagging, boosting concentrates on subsequent iterations of previously incorrectly classified instances [10].

To get started, we used the default configuration provided by Weka. Here, DecisionStump was used as base learner and AdaBoostM1 run with ten iterations and weight trimming.

---

**Figure 3.** Classification Error Rate for SimpleLogistic depending on parameter $W$ and $H$. As $H$ has no influence on classification error rate, one line is visible only.

**Figure 4.** Classification error rate for boosting depending on the number of iterations $I$ and bag size $P$.

**Figure 5.** Classification error rate for RandomForest iterating over the number of RandomTrees $I$ and number of features to consider in random selection $K$. 

---

**Table:**

<table>
<thead>
<tr>
<th>SimpleLogistic</th>
<th>Bagging</th>
<th>RandomForest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Error Rate</td>
<td>Iterations</td>
<td>Number of RandomTrees</td>
</tr>
<tr>
<td>Weight Trim Beta</td>
<td>$P=25$</td>
<td>$K=0$</td>
</tr>
<tr>
<td>$P=50$</td>
<td>$K=10$</td>
<td></td>
</tr>
<tr>
<td>$P=75$</td>
<td>$K=20$</td>
<td></td>
</tr>
<tr>
<td>$P=100$</td>
<td>$K=30$</td>
<td></td>
</tr>
</tbody>
</table>

---
pruning of 100. This combination resulted in a classification error of 50%.

Next we applied J48 as base learner. J48 is the open-source implementation of the simple tree-based learner C4.5. Therefore, J48 is more likely to produce alternating results in presence of slight changes in the data set, as opposed to more complex schemes like LMT, being an essential property for applying Boosting.

In a first run we used AdaBoostM1 with different values for weight threshold and number of iterations. We observed a gradual decrease in classification error with increasing threshold and number of iterations.

Going from this observation, we evaluated AdaBoostM1 with J48 with higher values for iterations and weight threshold. Figure 6 shows the learning curves for this setup. Using this approach, we could achieve a classification error of 8.81% which equals 28 wrongly classified instances or 14 patients.

Figure 6. Learning curves for AdaBoostM1 with J48 as base learner

CONCLUSION AND FUTURE WORK

In this paper, we demonstrated how Alzheimer’s Disease, Parkinson’s Disease, Multiple Sclerosis, Breast Cancer and healthy controls can be distinguished with an accuracy of more than 91% using boosting and unprocessed data from functional protein microarrays. However, the corpus used in this study is relatively small. Therefore, in order to verify the presented findings, testing on a larger corpus would be preferable. In addition, other ensemble learning schemes such as Stacking could be evaluated in the future. Moreover, it would be interesting to see how feature normalization such as proCAT [8] or Robust Linear Regression [9] affects the outcome of this classification problem.

REFERENCES

Two-Stage Bioinformatics Approach for the Diagnosis of Hepatocellular Carcinoma and Discovery of its Bio-Network

Arinze Akutekwe and Huseyin Seker

Abstract— Machine learning and statistical techniques have been applied in identifying biomarkers and constructing predictive models for the early diagnosis of various diseases including Hepatocellular Carcinoma. These include the method of direct random walk for feature selection and classification using logistic regression. In this paper, we apply a two-stage approach for the discovery of novel bio-network in the diagnosis of the hepatocellular carcinoma. The results show that seven features selected by the Least Angle Shrinkage and Selection Operator (LASSO) using a 10-fold cross-validation yielded the best class discriminatory performance with the highest accuracy of 96.25%. The dynamic Bayesian Network modeling stratified the following biomarkers as associated with the disease; Transient receptor potential cation channel, subfamily V, member 3 gene might inhibit E2F transcription factor 4, p107/p130-binding and Superkiller viralicidic activity 2-like 2 that might also play inhibitory roles against protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte), with all having their cDNA sources from the liver.

INTRODUCTION

Hepatocellular Carcinoma (HCC) or primary liver cancer is currently ranked fifth among the most common cancer types in the world and is the third leading cause of cancer deaths [1]. There are an estimated 33,190 new cases and 23,000 people are expected to die from the disease in 2014 [2]. The most common risk factors for the disease are incidence of hepatitis B or C and cirrhosis, which can be caused by hepatitis C virus (HCV) and alcohol consumption. The liver is one of the largest organs in the body and performs many functions, which include filtering of harmful substances from the blood, making bile for digestion of fat that comes from food and storing glycogen sugar, which the body uses for energy. There are many tests and procedures for diagnosis of the disease, which include physical examination and history, liver function test, CT scan, Magnetic Resonance Imaging (MRI), ultrasound, serum tumour marker test and biopsy [3]. In line with technological advances in the post-genome era, the method of serum tumour marker has attracted significantly wide research attention. This is due to robust and reliable computational methods, which have been successfully applied to microarray data in the diagnosis of various diseases and exploration of significant pathway of disease mechanisms.

Researchers have proposed various computational approaches for diagnosing and detecting possible biomarkers for liver cancer from micro array data. The method of Particle Swarm Optimization (PSO) in tuning the parameters of Support Vector Machine (SVM) classifier was proposed to analyse the data [4]. The authors applied the f-score for feature selection and SVM for classification. A method of direct random walk for feature selection by topological importance was also proposed for similar task [5] where the feature selection method is combined with logistic regression classifier to evaluate the total recognition rate of selected features.

The proposed methods only analyse classification performance of feature subsets but are not able to model temporal relationships among selected biomarker features which enable better understanding of the relationships between potential biomarkers and hence efficient diagnosis of diseases. In this paper, we apply a two-stage bio-network discovery approach [7] to select high quality biomarkers and infer their temporal relationships for the diagnosis of the disease. At the first stage, feature selection is carried out using two methods with a 5 and 10-fold cross validation. At the second stage, Dynamic Bayesian Network is used to infer the temporal association of high quality selected biomarkers.

METHODS AND MATERIALS

A. Random Forest Recursive Feature Elimination (RF-RFE)

Random forest grows and aggregates many trees, which are trained on a sample of N cases drawn at random with replacement from the complete set of N cases [6]. For a given training set X consisting of N cases which belong to two classes and P features, a classification tree can be constructed by first selecting a feature x and a threshold t that splits X into two maximally distinct subsets, from all features of x and all possible values of t. The training set is then split in to two blocks X1 and X2 depending on whether or not x ≤ t. The process is repeated with each X1 and X2 using another (x, t) combination. The process then continues until no further splitting is possible. As the random feature selection is done, it also computes the ensemble of trees.

The feature to split in each node in the tree is selected as the best among a set of randomly selected features. After generating a large number of trees, the most popular class in the trees is selected [6]. Recursive elimination is carried out by successively eliminating the least important variable based on decreased classification accuracy. If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.

B. Least Angle Shrinkage and Selection Operator (LASSO)

The lasso penalty is an effective and successful technique that is capable of dealing with a data set that consists of large p (predictors) and small n, (observations)
problem usually encountered in high-dimensional feature selection. The method performs shrinkage and continuous subset selection for linear and logistic regression via an $L_1$-norm regularization penalty [7]. The lasso penalized regression includes $\lambda$ being the tuning constant that controls the strength of the penalty. Reducing the value of $\lambda$ and relaxing the shrinkage penalty results in more number of predictors being selected. To perform the best subset selection however, cross-validation is generally used to find an optimum value of $\lambda$.

iii. **Support Vector Machine Recursive Feature Elimination (SVM-RFE)**

The SVM-RFE was used for selecting predictors relevant to classification problem [8]. The method generates the ranking of features based on backward feature elimination by training a SVM. In this study, we used SVM with radial kernel function being one of the most popular kernels. For the kernel, $\gamma$ defaults to the inverse of dataset dimension and $C$ is the regularization parameter and controlling penalty for imperfect fit to training labels.

**B. Dynamic Bayesian Network (DBN)**

Bayesian Networks (BN) are directed acyclic graphs (DAG) with nodes that represent random variables [12].

Dynamic Bayesian Networks (DBNs) are Bayesian Networks that aid modeling of associations arising from temporal dynamics in time between features of interest which otherwise cannot be performed using static Bayesian Networks [10]. It is shown in [11] that in modeling a DBN, arcs defining dependence relationships among variables of successive time points can be represented when a vector auto-regressive process (VAR) model of order 1 (VAR(1)) is assumed.

DBNs are used to represent directed graphical stochastic models of dynamical systems that are problem specific. Variants of the Hidden Markov Models (HMMs), which are tools that can represent probability distributions over sequence of observations, can be considered to be DBNs [11]. HMMs, which can be seen as special cases of DBNs, are ubiquitous for modeling time series data where they are used to encode structures that are implied and not fully expressed in a DBN [12]. As the transition models in HMMs are assumed to be sparse, directed (usually cyclic) graph are used to encode their structure, whose nodes represent different states in the system [12]. This directed cyclic graphical representation generalizes a DBN representation which allows modeling of temporal feedback loops that are common in biological pathways, where parent gene inhibit or slow down the expression and chemical reaction of child genes [11].

Different shrinkage algorithms for learning and inference of DBN models for biological pathways using regularized estimators are studied in [12]. In this paper, we applied the G1DBN algorithm to infer temporal relationship of selected features.

**RESULTS AND DISCUSSION**

Publicly available liver cancer data set (GSE17856) was used in this study [14] to evaluate the proposed hybrid method and potentially identify an informative gene sub-set that is expected to help early diagnosis of the disease and understand the mechanism of disease. It comprises of 25073 microarray gene expression profiles of 43 tumour and 44 non-tumour liver tissues surgically resected from patients with HCV-associated hepatocellular carcinoma. These make up a total of 87 observations and 25073 features used in this study. We implemented our algorithm using R Language [15] on a Linux-based Intel® Xeon Workstation with 16GB RAM. The performance criteria of sensitivity, specificity, accuracy, standard deviation of mean accuracy (Std.Acc), false positive rate (type 1 error) and Matthew’s Correlation Coefficient (MCC), as shown in Table I, are used to rank the performance of the methods and selected features. Mathematical expressions of these measurements can be found in [16].

In order to select the best of biomarkers, 5-fold and 10-fold cross-validation were independently implemented for both the random forest and lasso feature selection methods. SVM was then used to evaluate the performance of various features selected by the feature selection methods with $C$ and $\gamma$ parameters set at $C=1$ and $\gamma = 0.001$. From Table I, 7 and 6 features were selected by the random forest method using the 5-fold and 10-fold cross-validation methods, respectively. In addition, 11 and 7 features were selected using the 5-fold and 10-fold cross-validation methods for the lasso. The SVM classifier was then used to ascertain the performance of the selected features. The results show that the features selected by the lasso using the 10-fold cross-validation have the best performance with the highest accuracy and sensitivity of 96.25% and 98.33%, respectively.

The G1DBN algorithm performs DBN modeling in two main steps. The first step infers a first order dependence score matrix (S1) which contains the score of each edge of the DBN. At the second step, the score matrix (S1), and edge selection threshold alpha1 obtained in the first step are used to infer the score of each edge of a DBN describing full order dependencies between successive variables. The smallest score refers to the most significant edge.

The result of the DBN modeling is shown in Figure 1. These consist of the features selected by lasso which have been chosen because they yielded higher. It is worth noting that the 7 features selected by the lasso using the 10-fold cross-validation are subsets of the 11 features selected using 5-fold cross-validation. From the figure, transient receptor potential cation channel, subfamily V, member 3 (TRPV3) which has its cDNA in the liver, might inhibit E2F transcription factor 4, p107/p130-binding (E2F4). The former belongs to a family of nonselective cation channels that function in a variety of processes, including temperature sensation and vasoregulation. The E2F4 plays a crucial role in the control of cell cycle and action of tumor suppressor proteins and is also a target of the transforming proteins of small DNA tumor viruses [13]. This gene has also been known to be highly associated with colorectal cancer [17].
The DBN model further reveals that the superkiller viralicidic activity 2-like 2 (SKIVL2) gene might be highly associated with protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte (PTPN4)). The protein encoded by the PTPN4 gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are also known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. These two genes also have their cDNA sources from the liver and may be highly associated with the disease [13]. Further investigation is therefore required for these genes.

Table II shows 10 key genes selected across all the three high accuracy methods which have their cDNA sources associated with the liver. These genes should be further investigated as they may further reveal new insights for biologist and scientists in the diagnosis and treatment of the disease as well as further help understand the mechanism of the disease.

**TABLE II DESCRIPTION OF THE BIOMARKERS IDENTIFIED**

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOT13</td>
<td>Acyl-CoA thioesterase 13</td>
</tr>
<tr>
<td>RPS6KA2</td>
<td>Ribosomal protein S6 kinase, 90kDa, polypeptide 2</td>
</tr>
<tr>
<td>SKIVL2</td>
<td>Superkiller viralicidic activity 2-like 2</td>
</tr>
<tr>
<td>E2F4</td>
<td>E2F transcription factor 4, p107/p130-binding</td>
</tr>
<tr>
<td>PTPN4</td>
<td>protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte)</td>
</tr>
<tr>
<td>SGMS2</td>
<td>sphingomyelin synthase 2</td>
</tr>
<tr>
<td>TRPV3</td>
<td>Transient receptor potential cation channel, subgroup V, member 3</td>
</tr>
<tr>
<td>MKS1</td>
<td>Meckel syndrome, type 1</td>
</tr>
</tbody>
</table>

**CONCLUSIONS AND FUTURE WORK**

In this paper, a two-stage bio-network discovery approach made up of two feature selection methods using the 5 and 10-fold cross-validation methods, and Dynamic Bayesian Network was successfully applied for the diagnosis of Hepatocellular Carcinoma (primary liver cancer).

The result obtained from comprehensive analysis showed that the 7 features selected by the LASSO using 10-fold cross-validation had the highest predictive accuracy of 96.25%. However, when 5-fold cross-validation was used, eleven features were selected by the LASSO yielding a predictive accuracy of 95.29%. It is worth noting that the 7 seven biomarkers found in the former were subsets of the 11 features selected by the later.

Dynamic Bayesian Network was used to model the temporal association of the selected features across two time points of t-1 and t. The results show that the stratified biomarkers (Transient receptor potential cation channel, subgroup V, member 3 gene might inhibit E2F transcription factor 4, p107/p130-binding and Superkiller viralicidic activity 2-like 2) might also play inhibitory roles against protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte), with all having their cDNA sources from the liver.

Having demonstrated the robustness of the approach taken, future works can implement other feature selection and classification methods where more parameters can be varied to determine the lowest error rate and best classification performance at the first stage of this study, which may help identify lower number of the biomarkers. Other regularized estimators such as the James-Stein Shrinkage can be applied in the second stage to infer temporal relationships and results compared. In addition, the disease would be studied exploring the biological databases to find out the relationship more about its bio-network and relationships with other genes.
REFERENCES


Interpreting the Prevalence of Regulatory SNPs in Cancers and Protein-Coding SNPs among Non-Cancer Diseases Using GWAS Association Studies

Zoya Khalid\(^1\), Osman Ugur Sezerman\(^1\)

**Abstract**— Biological mechanisms underlying diseases are quite challenging to understand, as there exists a complex relationship between human genetics and disease traits. Genome-wide association studies are potent methods in identifying single nucleotide polymorphisms (SNPs) which are linked with a large number of phenotypes. Even though GWAS list down the statistically significant SNPs which are found to be associated with phenotype of interest, still there is a need to look for the direct evidence regarding biological processes to better understand the disease development mechanism since it may differ in different complex diseases depending on the nature of the disease. From previous few years, there are above 11 million SNPs that have been labelled in databases specifically in dbSNP. Among them, the SNPs can be categorized as coding or non-coding SNPs depending on their location in the genome. Lethal changes in the coding parts of the genes might play role in development mechanism of complex diseases by directly affecting the functionality of the protein. Similarly those SNPs which are present at the regulatory regions splice regions, micro RNA binding sites and epigenetic sites might affect the level of gene expression and ultimately contributes in complex disease formation. We designed our study to understand the biological mechanism underlying complex diseases by performing statistical analysis on GWAS dataset. Seventeen different cancer types with the non-cancer diseases (autoimmune, Neurodegenerative and metabolic diseases) has been selected in order to understand the major factors involved in disease progression and the impact of SNPs in disease development. The statistical analysis includes the chi squared hypothesis testing with the null hypotheses \(H_0\): SNPs in coding and non-coding areas are not significantly different and with the alternate hypothesis as \(H_1\): They are significantly different. The results revealed that Complex diseases like Cancer are mostly caused by mutations occurring at non coding regulatory sites thus causing changes at expression levels of the genes involved, such as over expression of oncogenes and under expression of tumor suppressor genes as expected, whereas in other non-cancer diseases, mutations occurring at coding regions of the genes play more determinative role. These mutations change the functionality of the protein product thus having a direct impact on the autoimmune response. This study in future can be taken as a reference study for analysing coding and non-coding parts of the genome regarding divulging biological mechanisms involving complex diseases.

*Corresponding Author (e-mail: zoyakhalid@sabanciuniv.edu).
Author Two (e-mails: ugur@sabanciuniv.edu).

**INTRODUCTION**

The first drift of large scale genome wide association studies (GWAS) has contributed in understanding the mechanism underlying complex diseases. For some diseases including Breast cancer, Prostate, Asthma and type II diabetes there has been speedy growth of loci involved in predisposition. From the studies it has been found that GWAS studies are suitable for the identification of SNP variants with effects on phenotype.\(^1\) The major portion of medical research has inclined towards the protein-coding mutations owing to the fact that the mechanisms underlying the regulatory SNPs are quite complicated and still not fully understood. SNPs whether in the coding or non-coding areas may be detrimental and may contribute in the development of complex diseases. Genome Wide Association studies which includes hundreds and thousands of SNPs which are tested concurrently in large number of cases and control samples in order to associate them with the complex disease have developed the hunt for genetic basis of these diseases. The success of GWAS can be seen from the fact that it identified novel SNPs and have been hypothesized as valuable tools for finding complex disease genes with the help of association studies and, afterwards to be used as markers for further genetic analysis.\(^2, 3\)

Understanding a disease etiology is the challenging task for biologists as it is the first step for disease diagnosis and treatment. Among the complex diseases Asthma is one of the types. Asthma is an inflammatory disease of the airways which is characterized by intrusion and activation of inflammatory cells followed by structural changes. These changes are supposed to associate with the severity of asthma and moderately with the development of progressive lung function weakening. The principal mechanism including airway angiogenesis in asthma and its detailed clinical significance has not yet been fully revealed\(^4\). Asthma is a common, complex disease that is affecting more than 300 million people worldwide.\(^5\) Similarly rheumatoid arthritis is also a heterogeneous autoimmune disease followed by caustic inflammation particularly in joints. The pervasiveness in the common population is about 0.5% to 1%, and among them women are at more risk for developing the disease.\(^6,7\)

Among the cancer types, Prostate cancer is the second most common cancer among men population. The prostate cancer shares some common features with other cancer types like breast cancer and the colon.
cancer. GWAS studies have been contributing successfully in identifying common variants (SNPs) which are significantly associated with the Prostate cancer. [8] Similarly for the other cancer types GWAS appeared to be as a potent tool for identifying diseased loci in cancers [9].

In this study we selected seventeen different types of diseases including both cancer and non-cancer phenotypes namely Asthma, Arthritis, Parkinson Disease, Crohn’s Disease, Hypertension, Obesity, Schizophrenia, Alzheimer, Multiple Sclerosis and for cancer types Breast cancer, Prostate cancer, Pancreatic, Colorectal, Liver, Esophageal and Bladder. We compared the ratio of SNPs distribution among all the disease types selected depending on their location in the genome by applying chi-squared test statistics.

METHODOLOGY

The methodology has been designed in order to find the significant SNPs and to analyse the ratio of these top selected significant SNPs in coding and non-coding regions of genome among different diseases and to further distinguish the driver mutations from the neutral ones in the coding regions of cancer. The SNP’s were collected from GWAS. GWAS usually emphasizes on associations regarding single-nucleotide polymorphisms (SNPs) and complex diseases. We collected GWAS data for different cancer and non-cancer diseases. The significant SNP’s were downloaded from https://www.gwascentral.org/

The significant SNPs are functionalized into coding and non-coding genomic regions by using SNPnexus Server. It is a functional annotation tool to evaluate the likely significance of candidate variants which is then linked to the gene/protein isoforms that might be phenotypically important. It can be accessed through http://www.snp-nexus.org/

For statistical inference, Chi-squared Test analysis was performed to examine that whether the mutations observed in coding regions are significantly different from the non-coding ones.

The following parameters are followed.

- **Alpha value**: 0.05 that means 95% confidence interval.
- If the test statistics value is higher and the P-value is less than alpha (p< 0.05) the hypothesis is rejected that means two variables are statistically different.
- If the test statistics value is lower and the P-value is higher than alpha (p> 0.05) the hypothesis is accepted that means two variables are not statistically significant.

And we took hypothesis as follows

H₀: SNPs in coding and non-coding areas are not significantly different. (Null Hypothesis)

H₁: They are significantly different. (Alternate hypothesis)

RESULTS

Starting with the most significant SNPs obtained, polymorphisms belonging to selected cancer and non-cancer types have been analyzed. It has been observed that SNPs are occurring at different rates depending on their location in the genome i.e. coding and non-coding areas. Table 1 and 2 summarized the distribution of SNPs in coding and non-coding areas per disease type and Figure 1 further depicts the distribution of non-coding parts intro introns and downstream regions. The first part of our study is to analyze the ratio of SNPs among cancer and non-cancer types depending on their genome location. For this the chi squared test analysis was performed. The two hypotheses selected are mentioned in the methods part. From the results it has been observed that for cancer diseases the value of test statistic obtained was 185.0561 with the p-value 0.000045, while for non-cancer ones it was 30.69 with p-value 3.6315e-08. On the basis of these values obtained it has been clear that our null hypothesis has been rejected and the alternative hypothesis has been accepted. The alternative hypothesis states that the SNPs are occurring at significantly different rates in the coding and non-coding regions. The results are mentioned in Table 3. Furthermore the ratio of coding and non-coding SNPs among cancer and non-cancer diseases can also be graphically viewed in Figure 2. From the figure it has been visualized that in cancer like diseases the SNPs are more prevalent in non-coding parts in comparison to the coding ones, while the scenario is quite opposite in non-cancer diseases where the coding SNPs are in higher proportion than the non-coding ones.

DISCUSSION

From previous few years, there are above 11 million SNPs that have been labelled in databases specifically in dbSNP. Among them, the SNPs can be categorized into coding and non-coding SNP depending on their location in the genome. Lethal changes in the coding parts of the genes might play role in development mechanism of complex diseases by directly affecting the functionality of the protein. Similarly those SNPs which are present at the regulatory regions splice regions, micro RNA binding sites and epigenetic sites might affect the level of gene expression and ultimately contributes in complex disease formation. Biological mechanisms underlying diseases are quite challenging to understand, as there exists a complex relationship between human genetics and disease traits. A Mendalian disease can occur as a result of a single mutation on a gene that explains all the disease cases, however, complex diseases do not associate strongly with a single gene mutation, and rather they are caused by variant forms of several genes which can be the direct result of several single nucleotide polymorphisms
targeting these genes. Among various diseases, cancer is one of the multifactorial diseases because of the fact that they are likely associated with the effects of multiple genes in permutation with additional factors including lifestyle and environmental factors. To understand the biology behind complex disorders is very challenging, since the root cause for most of these disorders have not been identified so far. Genome Wide Association studies includes hundreds and thousands of SNPs which are tested concurrently in large number of cases and control samples in order to associate them with the complex disease have developed the hunt for genetic basis of these diseases. The success of GWAS can be seen from the fact that it identified novel common genetic risk factors involved with the significance of earlier recognized generic variants. Hence it is known that concentrating on few SNP and genes showing strong association with the disease is not sufficient to understand the underlying disease mechanism, because there exists a chance that those biologically important genetic variants that have a small disease risk are might get overlooked.

Bearing in mind all these factors we designed our study to understand the biological mechanism underlying complex diseases with GWAS studies. We analysed seventeen different types of cancer and non-cancer disease (autoimmune, neurodegenerative and metabolic). The top selected most significant SNPs which are ranked according to their p-values are filtered out. The threshold was set to p-value <0.05. We compared the frequency distribution for SNPs present in the coding and non-coding regions for all the disease types tested in this study. The analysis revealed that for cancer diseases more crucial part of research is the non-coding or the junk DNA. Mutation lying in this region is the regulatory mutation, which does not translate into protein but still has multiple impacts on disease. The regulatory mutations alter the ability of a transcription factor to bind to DNA which further affects the gene expression level and ultimately leading to disease. These regulatory mutations are likely the drug targets, so it’s certainly important to look for these mutations in order to fully understand the biological process behind this and furthermore to target correct drugs for treatment and therapies.\cite{10, 11}
### CONCLUSION

The analysis revealed that Complex diseases like Cancer are mostly caused by mutations occurring at non-coding regulatory sites thus causing changes at expression levels of the genes involved, such as overexpression of oncogenes and underexpression of tumor suppressor genes as expected. Whereas in other non-cancer diseases (autoimmune diseases, metabolic and Neurodegenerative diseases), mutations occurring at coding regions of the genes play a more determinative role. These mutations change the functionality of the protein product thus having a direct impact on the autoimmune response. This study in future can be taken as a reference study for analysing coding and non-coding parts of the genome regarding divulging biological mechanisms involving complex diseases.

### REFERENCES


**Abstract**—Drug-Drug Interaction (DDI) is one of leading health threats causing morbidity and mortality. Social network analysis provides a set of algorithms for link prediction to identify missing or spurious links. The link prediction methods used in social networks may be good candidates for prediction of DDIs. In this paper we exploited the social network similarity measures for link prediction on DDI networks. The performance of the methods applied on two different DDI networks, extracted from Drugbank and STITCH, was investigated.

**INTRODUCTION**

Drug-drug interaction (DDI) is one of leading health threats causing morbidity and mortality. In some cases, adverse effects or critical threats for patient who takes multiple medications may be observed due to DDIs. For example, 4% of the cancer patients were died because of complications originated from DDIs. The pharmaceutical companies and scientists have to pay much attention to detect and predict DDIs for both drug discovery and drug toxicology. Detection of drug interactions needs system-level analysis. Thus, network-based approaches can gain new insights to detection of beneficial and harmful drug combinations.

In drug-drug interaction network, a link (interaction) between two nodes (drugs) exists if the experiments or experts verify the interaction. The knowledge of the interaction network is usually not complete. Many drugs have been withdrawn from market because of adverse effects of drug interactions that is unknown before. The experiments needed to check all possible drug-drug interactions are costly. So focusing on the interactions that most likely exist can reduce these experimental costs dramatically. Social network analysis provides a set of algorithms for link prediction to identify missing or spurious links. The link prediction methods used in social networks may be good candidates for prediction of DDIs. In this paper we exploited the social network similarity measures for link prediction on DDI networks.

In literature, there are many recent works regarding DDI using information from biological and chemical networks [1, 2]. These approaches include prediction of DDIs from protein-protein networks [3], drug-target interaction network [4] and detection of Adverse Drug Reactions (ADRs) by integrating protein-protein networks and drug structures [5]. Hu et al. [6] presented the results regarding their works by applying network analysis methods based on some properties of 966-drug DDI network. In [7], the authors developed drug combination predictor using network features. Further, the analysis of DDI network in terms of functional context was made by the Lee et al. [8].

In this paper, by using two standard performance metrics [14] we explore the performance of the methods used in social network analysis for link prediction [11-17] on the drug-drug interaction networks. In section II the link prediction problem is introduced. We review link prediction methods and standard metrics for performance evaluation in Section III and Section IV respectively. The data sources are presented in Section V. Two last sections are results, discussion and conclusion.

**PROBLEM DESCRIPTION**

We consider $G= (V, E)$ is a graph $G$ with the node set $V$ and edge (link) set $E$; each edge is a link between pair vertices (nodes), $e=(x, y)$. For two distinct times $t<t'$, suppose $G = [t, t']$ be a subgraph of graph $G$ that it is constructed during time interval $[t, t']$. We will have two input graphs, **training graph.** subgraph of $G$ that is constructed during training interval $[t_0, t'_0]$ shown by $G[t_0, t'_0]$ and **test graph.** subgraph of $G$ that is constructed during test interval $[t_1, t'_1]$ assigned by $G[t_1, t'_1]$ where $t'_0<t_1$. We use two parameters $K_{\text{training}}$ and $K_{\text{test}}$ that denotes the number of interactions (links) for each node in training graph and test graph respectively. Also, **Core set** is used as the set of nodes which have at least $K_{\text{training}}$ interaction(s) in training graph and at least $K_{\text{test}}$ interaction(s) in test graph. $E_{\text{old}}$ and $E_{\text{new}}$ are the sets of links that are between core nodes.

**LINK PREDICTION METHODS**

In this section we present some basic methods to score and compute the probability of a link to appear in future network that does not exist in training graph. Each method computes score($x, y$) for unconnected pair of nodes $<x, y>$. Therefore, these methods can be calculated as similarity of nodes $x$ and $y$. Let $\Gamma(x)$ denotes the set of nodes that are adjacent to node $x$;

1. **Distance Graph [14]:** This method considers the length of shortest path between $x$ and $y$ as their similarity or probability to predict $<x, y>$ link.
2. **Common neighbor [14]:** In this method, the number of nodes that are in neighbor set of both nodes $x$ and $y$ is considered as similarity of $x, y$. Then, Score($x, y$) = $\left|\Gamma(x) \cap \Gamma(y)\right|$, where $\Gamma(x)$ is neighbors set of node $x$ and $\Gamma(y)$ is neighbors set of node $y$; the intersection of these two sets are their common neighbors. Since for any set $S$, $|S|$ denotes number of elements of the set $S$, and then $\left|\Gamma(x) \cap \Gamma(y)\right|$ shows the number of common neighbors.
3. Jaccard’s Coefficient [11] is a similarity measure that computes the probability of common features namely \( f \) that is randomly selected from features of \( x \) or \( y \) has. Since this paper just uses the “neighborhood “ feature, Jaccard’s coefficient for pair nodes \( x, y \) is:

\[
Score(x, y) = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{\mid \Gamma(x) \cup \Gamma(y) \mid}
\]

4. Adamic/Adar [13] have introduced a relative measure that computes similarity between two pages, as nodes of graph, by counting of common features by weighting rarer features more heavily. Since we are using just neighborhood feature, then, the measure is:

\[
Score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log \mid \Gamma(z) \mid}
\]

I. EVALUATION METRICS

To measure the performance of prediction methods, two standard measures are used: precision [18, 19] and area under the receiver operating characteristic curve (AUC) [20]. In fact, each link prediction method outputs an ordered list \( L_p \) that giving score (appearing probability) to all non-existing links between core nodes in training set (\( L_p = Core \times Core - E_{old} \)).

1. **Precision**: Precision metric is the ratio of right predicted link (according to test graph) to \( n \) where \( n = \mid E_{new} \mid \). That is, to compute this metric we choose \( n \) top scored predicted \((x, y)\) links from core nodes pairs and evaluates the accuracy of each one by considering test graph.

2. **AUC**: The AUC value is the probability that a randomly chosen \((x, y)\) link from test graph is given a higher score than a randomly chosen nonexistent link in union of test and training graphs. To compute AUC metric, we compare each element of \( L_p - E_{new} \) to all elements of \( E_{new} \). Suppose \( n_1 \) times correct links (elements of \( E_{new} \)) have higher score in \( L_p - E_{new} \) and \( n_2 \) times have same scores. Then, ACU value calculated as:

\[
AUC = \frac{n_1 + 0.5 \times n_2}{\mid L_p - E_{new} \mid \times \mid E_{new} \mid}
\]

DATA SOURCES

We collected two different sets of DDIs from STITCH [9] and Drugbank [10]. For the first dataset, Drugbank v2.0 is used as training graph. We used version 3.0 of Drugbank as test graph. Since test \((v_3)\) graph is overlapped to training graph \((v_2)\), we have used difference between them \((v_2-v_3)\) as test graph. For Drugbank dataset, the training graph has 841 nodes and 9047 edges and test graph has 1167 nodes and 22630 edges. For dataset obtained from STITCH, we take version 3.1 as training graph and difference between version 4.0 and version 3.1 as test graph. We consider only chemicals having drug property. The graph constructed using STITCH contains 6283 nodes and 166182 edges in the training dataset. The test graph of this dataset has 5779 nodes and 242870 edges.

RESULTS AND DISCUSSION

We tested prediction methods by changing \( K_{training} \) and \( K_{test} \) that are differing from 1 to 5 for all mentioned similarity measures over Drugbank v2.0 as training set and v3.0 as test. In these data sets, by increasing \( K_{training} \) values from 1 to 5, there is no any uniformity for each method that is, in some cases they are growing up but in some cases decreasing. However, when \( K \) parameter for both training and test gets bigger, precision increases but AUC value decreases. We interpret this as the nodes with more connections will be more likely to appear in the future graph. To apply mentioned methods on STITCH dataset we take 1, 2, 5, 10, 50 as values of \( K_{test} \) and \( K_{training} \). In this paper, because of page restriction, we just bring the table of results of best outperformed measure (graph distance measure) for Drugbank data set (Table I) and the table of best outperformed measure (Adamic/Adar) for STITCH dataset (Table II).

Tables I shows the performances of the outperformed measure (graph distance measure) for Drugbank dataset with both performance metrics Precision (showed in abbreviated for in Table as “Prec”) and AUC. In this table, the big amount of AUC performances are in the first column and maximum values are appears in first columns as well (this features is working for all measures). That means, in spite of Precision metric that says most of the predicted links would be in high interacted drugs, AUC metric says in output scoring list (\( L_p \)) we have best performance when we consider all nodes in test graph not just higher scores. By comparing results of Precision metric and AUC metric, we can say: these similarity methods giving a good scoring to all links but selecting first \( n \) high scored links for this dataset needs to revise. Moreover, in general case, when \( K_{test}=K_{training}=1 \) we can see maximum value for AUC belongs to graph distance (shortest paths) method where AUC=0.674. That means according to AUC metric, this method is the best method for general case. Also, the maximum value of AUC will choose graph distance method as the best similarity measure for this data set.

100
Table II shows that Precision performance gets significant values in STITCH dataset in compare with Drugbank dataset but it doesn’t have uniformity trend like Drugbank dataset by increasing \( K_{\text{test}} \) and \( K_{\text{training}} \). AUC metric doesn’t show uniformity trend by increasing \( K_{\text{training}} \) in this dataset, but we can see a decreasing trend when \( K_{\text{test}} \) is rising. Also, AUC displays the best performance for Adamic/Adar method and takes its maximum when \( K_{\text{training}}=50 \) and \( K_{\text{test}}=1 \). That gives two results: i) for prediction on STITCH dataset, can suggest to use Adamic/Adar method. ii) Maximum performances shows the best case for predicting is for hubs in training graph that have at least one interaction during test period.

For given two datasets, best performed methods are different. Graph distance method is the best for Drugbank dataset and Adamic/Adar measure performs better for STITCH dataset. This may be due to the characteristics of data sources which may give us clues about how they have been built and treat. Drugbank contains more experimental knowledge about DDIs whereas STITCH has been built using text mining, drug similarity and experimental knowledge to indicate a link between two drugs.

CONCLUSIONS

In this research we explored four network similarity measures, Common Neighbors, Graph distances, Adamic/Adar and Jaccard’s Coefficients on two drug datasets. We used two parameters \( K_{\text{training}} \) and \( K_{\text{test}} \). \( K_{\text{training}} \) is a bound to choose the nodes with at least \( K_{\text{training}} \) neighbors in training graph. In Drug graph this means these nodes have at least \( K_{\text{training}} \) interactions with other drugs. In precise speaking, when \( K_{\text{training}} \) is a big number, we interpret that the selected nodes are some special drugs that have more interactions with other drugs or some which there have been so many investigating on them to find their interactions with other. Then by rising \( K_{\text{training}} \), we are focusing on just more investigated and more interacted drugs. If performance for linked prediction becomes high for some \( K_{\text{training}} \) values, we can say that in this drug dataset, more interacted drugs will need to investigate more and vice versa.

According to our results, two different methods are outperformed depending on datasets. Graph distance scores best for Drugbank dataset and Adamic/Adar measure has better scoring for STITCH dataset. This is most probably comes from the different type of information used to construct these databases.
TABLE II. ADAMIC-ADAR MEASURE FOR STITCH DATASET

<table>
<thead>
<tr>
<th>K_{test}</th>
<th>K_{training}</th>
<th>Prec</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>0.558</td>
<td>0.905</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.413</td>
<td>0.908</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.385</td>
<td>0.9088</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.363</td>
<td>0.9089</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.358</td>
<td>0.909</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>0.253</td>
<td>0.811</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.341</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.313</td>
<td>0.8967</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.294</td>
<td>0.902</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.290</td>
<td>0.9034</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>0.492</td>
<td>0.866</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.333</td>
<td>0.881</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.305</td>
<td>0.893</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.285</td>
<td>0.895</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.281</td>
<td>0.8972</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>0.489</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.329</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.299</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.279</td>
<td>0.888</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.276</td>
<td>0.890</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>0.488</td>
<td>0.860</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.327</td>
<td>0.870</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.297</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.277</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.273</td>
<td>0.884</td>
</tr>
</tbody>
</table>

REFERENCES

Amino acid preferences at neddylation sites

Ahmet Sinan Yavuz, Namik Berk Sözer and Osman Uğur Sezerman*

Abstract— Neddylation is a dynamic post-translational modification in which NEDD8 proteins are covalently attached to the target site lysine residue. Neddylation may affect a target protein’s localization, binding partners and structure. Targets of this modification have commonly found in nucleus and the most well characterized target family is cullins, which is modulating ubiquitination and proteosomal degradation system in a cell. Disruptions in neddylation pathway implicated in various diseases such as Alzheimer’s, Parkinson’s and cancer. Therefore, understanding neddylation site recognition bears a huge importance in understanding the complete functional mechanism of this post-translational modification and revealing the mechanisms of associated diseases towards a cure. However, there is no study in literature investigating whether a common neddylation site motif exists or not. In this work, we have identified various amino acid preferences and hydrophobicity patterns seen in neddylation sites, differing from not neddylated lysine residues.

INTRODUCTION

NEDD8 is an ubiquitin-like modifier, which is encoded by NEDD8 gene in humans and Rub1 gene in S. cerevisiae. NEDD8 was initially identified as one of the ten neural precursor cell-expressed, developmentally down regulated genes (NEDD) and defect on NEDD8 pathway shown to be lethal in many organisms [1]. NEDD8 shares ~60% sequence identity with ubiquitin, and it is the most similar known ubiquitin-like protein (Ubl) [1].

Neddylation is the covalent attachment of NEDD8 proteins to the target sites. Similar to SUMO and other Ubl proteins, NEDD8 is synthesized in an immature form [2]. Cleavage of extra amino acids catalyzed by UCH-L3 enzyme located beyond Gly76 reveals the mature isopeptide linkage site, which will form a bond with target site’s lysine residue [2]. Following the maturation, NEDD8 can be activated to bind to an E1 enzyme (UBA3-APPBP1 heterodimer) consuming 1 ATP in the process (Figure 1). Afterwards, E1 bound NEDD8 is loaded on an E2 enzyme (UBC12) and from E2, with or without help of an E3 enzyme, it is transferred to the target site’s lysine residue [2]. Attached NEDD8 proteins then can be removed by NEDD8 isopeptidases, making neddylation a dynamic and reversible process.

As many other post-translational modifications, neddylation directly affects 3D surface of a target protein, which may alter binding partners of the substrate and/or stimulate a conformational change in the structure [2].

*Corresponding Author (e-mail: ugr@sabanciuniv.edu).

Ahmet Sinan Yavuz and Osman Uğur Sezerman are with Biological Sciences and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, Orhanlı, Tuzla, Istanbul, Turkey (email: asinanyavuz@sabanciuniv.edu and ugr@sabanciuniv.edu).

Namik Berk Sözer is with Department of Genetics and Bioengineering, Faculty of Engineering and Architecture, Yeditepe University, Istanbul, Turkey, (e-mail: namikberk.sozer@std.yeditepe.edu.tr).

Additionally, neddylation can encourage the recruitment of NEDD8-binding proteins, causing new protein complexes to occur [2]. All of these direct effects may also influence further changes such as in subcellular localization [2]. In conclusion, NEDD8 attachment to a substrate may significantly alter target protein’s lifespan, role, subcellular localization and structure.

Neddylated proteins can be found mostly in the nucleus and the most well-characterized targets are the cullin proteins [3], [4]. Cullin proteins are scaffold proteins of SCF-ubiquitin ligase complex, which controls the ubiquitination and proteosomal degradation system [4]. Neddylation of cullins increases the ubiquitination and proteosomal degradation of substrates. As main targets of cullins are cell-cycle regulation, transcriptional regulation and signal transduction proteins, neddylation plays a significant role in the maintenance of cell machinery [4]. Hence, disruptions in neddylation pathway has observed in many diseases, such as Alzheimer’s [5], Parkinson’s [6], and cancer [7].

Although neddylation plays significant roles in cells, target recognition and specificity is still unclear. There is no previously reported neddylation site motifs or published neddylation site prediction tools. Identification of neddylation target sites experimentally is also expensive and laboursome. Therefore, there is a need of identifying possible sequence properties of neddylation sites to aid in prediction of such sites.

In this short work, we aim to identify common seen amino acid preferences or hydrophobicity patterns seen in the neddylation sites.

METHODS

**Dataset**

We have searched PubMed with keywords “nedd8”, “neddylation”, “nedylation”, “rub1”, “rub2”, “rub3”, and “rubylation”, and manually collected 63 sites in 29
proteins from ~600 articles, published until July 1\(^{st}\), 2014. Among these sites, 6 were discarded due to neddylation was shown only in vitro, and 3 were discarded as neddylation was not reported in a single amino acid resolution. After this elimination, primary sequences of 28 proteins were retrieved from UniProt [8].

Redundancy elimination was performed with CD-HIT [9]. This program clusters sequence datasets and selects a representative sequence of each cluster having at least a given percent identity. We clustered sequences with 0.4 threshold, so that no two sequences sharing a sequence identity >40% left in the dataset. After such an elimination procedure, dataset was left with 22 proteins and 48 sites.

We prepared dataset for analysis by defining sequence windows as lysine residues flanked by 10 residues upstream and 10 residues downstream, forming a 21 amino acid long sequence segments. All sequence windows that contain experimentally identified neddylation sites were considered as the positive set and rest of the sequence windows was assumed as not neddylated and formed the negative set.

**Amino Acid Grouping and Hydrophobicity Scale**

In order to assess common biochemical properties in the sequence windows we have used both 20-letter amino acid alphabet and a 11-letter grouping of amino acids based on physicochemical properties, named as Sezerman grouping [10], [11] (Table 1).

<table>
<thead>
<tr>
<th>Groups</th>
<th>Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>IVLM</td>
</tr>
<tr>
<td>Q</td>
<td>RKH</td>
</tr>
<tr>
<td>C</td>
<td>DE</td>
</tr>
<tr>
<td>D</td>
<td>QN</td>
</tr>
<tr>
<td>E</td>
<td>ST</td>
</tr>
<tr>
<td>F</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>H</td>
<td>W</td>
</tr>
<tr>
<td>I</td>
<td>C</td>
</tr>
<tr>
<td>W</td>
<td>YF</td>
</tr>
<tr>
<td>K</td>
<td>P</td>
</tr>
</tbody>
</table>

Additionally, we have used Kyte & Doolittle [12] hydrophobicity scale to assess the hydrophathy difference between neddylated and not neddylated sequence windows.

**Statistical Testing**

In order to assess statistical difference in hydrophobicity values between known neddylated sequence windows and not neddylated sequence windows, two-tailed Mann-Whitney U tests were performed.

Additionally, amino acid profiles of neddylated and not neddylated sequence windows were compared with chi-square test of independence. Two strategies were employed to identify differences clearly. First one was implemented by creating 20x2 and 11x2 contingency tables for each position in the window, for normal amino acid and Sezerman grouping amino acid distributions, respectively. This approach aims to identify general differences in amino acid distributions. Second strategy was to identify whether particular amino acids are over or underrepresented in particular positions of the sequence windows. For this strategy, we have created twenty 2x2 contingency tables for normal amino acid representation and eleven 2x2 contingency tables for Sezerman grouping.

Benjamini- Hochberg [13] procedure has been applied for controlling false discovery rate at \(\alpha = 0.05\). All p-values have been adjusted according to this procedure.

All statistical tests were performed using R (version 3.1.0, The R Foundation for Statistical Computing, Vienna, Austria; http://www.r-project.org) and an in-house program written in Python 2.7.5 [14], with the SciPy [15] library (version 0.11.0).

**RESULTS**

**Sequence Logos**

In order to identify amino acid preferences visually, we have created sequence logos using WebLogo 3 [16] to represent probability of an amino acid to be present in a certain location in the sequence window (Figure 2).

Sequence logos identified a strong positively charged/polar amino acid preference difference in -3\(^{rd}\) position of the window (Figure 2a,b). Sezerman grouping results also supported this finding by more than 40% probability assigned to positively charged amino acids group (Q), and more than 20% probability assigned to polar amino acid groups (D, E) (Figure 2c,d). Similar difference can be observed in +8\(^{th}\) position, in which both Figure 2a-b comparison and Figure 2c-d comparison reveals a different preference of amino acids. Lastly, Figure 2a and 2b reveals additional differences between groups, such as overrepresentation of A at position -7, and overrepresentation of positively charged amino acids at position +3, however, commenting on these differences may require additional evidence, such as statistical testing.

**Statistical Testing**

We have performed statistical testing to identify differences in two aspects: overall amino acid composition, single amino acid over/underrepresentation in each position of sequence windows. Overall amino acid compositions showed no statistically significant difference occurs between positions (all p-values > 0.05).
On the other hand, single amino acid preference tests revealed several under/overrepresented amino acids. As it has been identified from sequence logos, positively charged amino acids (Sezerman grouping: Q) were significantly overrepresented in position -3 (44% of positive windows, while 15% of the negative windows), $\chi^2(1, N = 960) = 28.44, p < 0.001$. In this particular site, only arginine presence was statistically significant too, with arginine present in 25% of the positive sites and 4% of the negative sites, $\chi^2(1, N = 960) = 40.09, p < 0.001$. Charged and polar amino acids were found statistically significant in other two locations: +7 and +8. Histidine was overrepresented in position +7. 19% of the positive sites and only 2% of the negative sites have histidine in this position, $\chi^2(1, N = 960) = 40.75, p < 0.001$. Asparagine presence was found to be significant in position +8, where it is present in the 17% of the positive sites and 4% of the negative sites, $\chi^2(1, N = 960) = 16.23, p = 0.004$.

Apolar amino acids showed overrepresentation on various positions as well. Alanine was statistically significantly overrepresented with 25% of the positive sites, and 6% of the negative sites in position -7, $\chi^2(1, N = 960) = 23.44, p < 0.001$. Similarly, a methionine overrepresentation (19% of the positive sites, 2% of the negative sites) in -1$^\text{st}$ position was declared statistically significant, $\chi^2(1, N = 960) = 38.95, p < 0.001$. Valine was overrepresented in neddylated windows at position -4 as well. It was present at 25% of the positive sites and 7% of the negative sites, $\chi^2(1, N = 960) = 21.76, p < 0.001$. Lastly, isoleucine was overrepresented at position -5. It was present 29% of the positive sites, while only 6% of the negative sites, $\chi^2(1, N = 960) = 34.19, p < 0.001$. This overrepresentation can also easily seen from the sequence windows in Figure 2a-b.

It should be worth noting that frequency based statistical testing results reported in this section should be taken into consideration carefully, as some of the amino acids may be declared significant due to only dataset-specific frequency differences, and they may not imply anything on underlying biological principle.

Hydrophobicity

Hydrophobicity can differentiate target sites from non-target sites as it is an important effect in protein-protein binding [17]. Efficacy of hydrophobicity has shown in sumoylation site prediction previously [18]–[20]. Therefore, same principle may lead neddylation site recognition. In order to identify hydrophobicity differences, we have plotted boxplots of Kyte-Doolittle [12] hydrophobicities for each location and performed Mann-Whitney U tests to determine statistical significance (Figure 3). In Kyte-Doolittle [12] hydrophobicity scale, while hydrophobic residues have positive scores, hydrophilic residues have negative scores.

Supporting the sequence logos and statistical testing results, hydrophobicity plots has showed a statistically significant charged residue preference in -3$^\text{rd}$ position (Figure 3, p<0.001). Similarly, boxplots have showed a polar tendency in positions +3 and +8 (Figure 3, p<0.05). Although, they were not statistically significant, positions +4 and +7 also shows a polar tendency, as well.
**DISCUSSION**

On overall, we have showed that neddylation sites does not have significantly overrepresented “consensus” motif, as it was the case for sumoylation. However, this may only because of the limitations of the dataset. On the other hand, we have identified various significant amino acid preferences, especially charged amino acids in 3rd position. This fact may imply a significance of this position in neddylation site recognition by UBC12.

Identification of amino acid preferences may be the first step in decrypting neddylation site recognition, and accomplishing successful in silico identification of neddylation sites may open up various new application fields, such as studying ubiquitination and proteome degradation abnormalities and associated diseases. However, small size of experimentally identified neddylation sites seriously limits the in silico efforts to identify neddylation sites. Hence, with the ever-increasing amount of experimentally identified neddylation targets, we expect neddylation site identification methodologies will grow significantly.

In addition to obtaining primary sequences of new experimentally validated neddylation sites, we may need additional structural insights of site recognition, as neddylation site recognition may not only determined by primary sequence and hydrophobicity, but also conformational state, flexibility and subcellular localization. As most of the post-translational modifications are dynamic processes affected significantly by subcellular localization and conformational state, obtaining such information would be enlightening in understanding the actual mechanism of site recognition. However, it seems unlikely to have access to this information on neddylation sites, soon. Therefore, systematical identification of neddylation proteome still presents a great challenge.

**CONCLUSIONS**

Neddylation, a vital post-translational protein modification, plays significant roles in cellular machinery as it mainly functions as a regulator of ubiquitin-protein ligases and proteome degradation system. In this paper, we showed several amino acid preferences in neddylation targets via sequence logos, statistical testing and hydrophobicity scales. Future work lays in developing a neddylation site predictor for the use of research community that uses other possible properties that may affect neddylation site recognition as well.

**ACKNOWLEDGMENTS**

Ahmet Sinan Yavuz is supported by 2211 – National Scholarship Program for PhD Students of The Scientific and Technological Research Council of Turkey (TUBITAK).

**REFERENCES**


**In silico** determination of bioactive peptides in rubisc/o of invasive **Caulerpa racemosa**

Zeynep Agirbasli and Levent Cavas*

**Abstract**—Recently developed *in silico* tools give important clues about food quality. The present paper reports the bioactive peptides from rubisc/o of invasive *Caulerpa racemosa*. BIOPEP tool was used to estimate the bioactive peptides from rubisc/o of invasive seaweed *C. racemosa*. According to the results, antioxidant, antihypertensive, inhibitor of angiotensin-converting enzyme based peptides were detected. In conclusion, inasmuch as there is a direct correlation between *in silico* and *in vitro* studies, development of *in silico* tools for detection of biopeptides within foods is of great importance. Since *C. racemosa* contains very important bioactive peptides, it could be evaluated as an alternative food source.

**INTRODUCTION**

The correlation between health and vegetarian type nutrition has been proven by many reports from all around the world [1-2]. The most abundant protein in plants is ribulose bisphosphate carboxylase/oxygenase (rubisc/o) [3]. The green leaves of plants are consisted of mostly rubisc/o and it plays an important role at carbon-fixing reaction in the stage I of the dark reactions of photosynthesis [4]. Rubisc/o adds inorganic carbon (CO$_2$) to ribulose-1,5-biphosphate and then an unstable 6-carbon containing β-keto acid intermediate is formed and it later cleavages to 2 moles 3-phosphoglycerate [4].

The structure of rubisc/o is heterohexadecamer type and it comprise from eight large subunits and eight small subunits (Mw 50-55 kDa/12-18 kDa) [5].

Beside its function in photosynthesis, this enzyme is of great importance because of following properties: most of the autotrophic organisms, prokaryotes and eukaryotes contain rubisc/o and 30-50% of rubisc/o is consisted of soluble protein [5].

Very limited information is existed on the bioactive peptides from rubisc/o of vegetables [3]. After *in vivo* or *in vitro* proteolytic process, proteins split into small peptides and some of them are called as bioactive peptides [6]. Their lengths are mostly between 2-20 amino acid residues. In addition to their nutritional values, they have wide functional health effects on cardiovascular, nervous, gastrointestinal and immune system [6-7]. While most of them only show one function, some of them are multi-functional.

*In vitro* determination of bioactive peptides within foods are time consuming and also expensive. Newly developed bioinformatics tools provide a lot of contribution to our understanding on human metabolisms [8]. One of the developed *in silico* tools on detection of bioactive peptides are BIOPEP [9]. BIOPEP, a tool and database was developed by the Chair of Food Biochemistry of the University of Warminia and Mazury [10]. By using fasta format of protein sequences, it is very easy to estimate the bioactive peptides in foods by BIOPEP [10].

In this study, because of the minority of the studies in this area we aimed at detection of bioactive peptides in rubisc/o enzyme from invasive marine seaweed *Caulerpa racemosa*.

**MATERIALS AND METHODS**

The fasta format of rubisc/o of *Caulerpa racemosa*, ID W8E959, was provided from UniProtKB/Swiss-Prot database at the ExPASy bioinformatics resource portal [8].

**RESULTS**

**In silico** proteolysis of rubisc/o was performed and the results were obtained by using BIOPEP database [9]. In order to apply artificial cleavage, thermolysin, papain, bromelain is used with digestion enzymes chymotrypsin, trypsin, pepsin. As a result, Angiotensin converting enzyme (ACE) inhibitor, antioxidative, dipeptidyl peptidase (DPP)-IV inhibitor, regulating, antithrombotic, antiangiogenic and ubiquitin mediated proteolysis (UbMP) activating bioactive peptides were found through *in silico* enzymatic hydrolysis of proteins into peptides.

ACE, dipeptidyl carboxypeptidase, plays an important role in the rennin-angiotensin system (RAS) and it regulates blood pressure, homeostasis, fluid and salt balance. ACE activates angiotensin I (Ang I) hydrolyzed by rennin to a potential vasoconstrictor angiotensin II (Ang II) and plays a role at inactivation of vasodilator bradykinin [7]. ACE inhibitors prevent these effects by giving a reaction with the enzyme. Therefore, antihypertensive peptides show vasorelaxing effect on blood vessels [7].

Some of the bioactive peptides in rubisc/o of *Caulerpa racemosa* have an positive impact on health. They have antioxidative effects on oxidative compounds like reactive oxygen species (ROS) such as hydrogen peroxide and singlet oxygen and free radicals. These antioxidative peptides have many abilities such as radical scavenging, inhibition of lipid peroxidation and metal ion chelation properties [7-11]. Mechanism of action and

*Corresponding Author (e-mail: lcavas@deu.edu.tr).
Levent Cavas and Zeynep Agirbasli are with the Department of Biotechnology, Graduate School of Natural and Applied Sciences, Dokuz Eylül University, İzmir-Turkey.

**Figure 1. Fasta Format of Caulerpa racemosa**

**In silico** proteolysis of rubisc/o was performed and the results were obtained by using BIOPEP database [9]. In order to apply artificial cleavage, thermolysin, papain, bromelain is used with digestion enzymes chymotrypsin, trypsin, pepsin. As a result, Angiotensin converting enzyme (ACE) inhibitor, antioxidative, dipeptidyl peptidase (DPP)-IV inhibitor, regulating, antithrombotic, antiangiogenic and ubiquitin mediated proteolysis (UbMP) activating bioactive peptides were found through *in silico* enzymatic hydrolysis of proteins into peptides.

ACE, dipeptidyl carboxypeptidase, plays an important role in the rennin-angiotensin system (RAS) and it regulates blood pressure, homeostasis, fluid and salt balance. ACE activates angiotensin I (Ang I) hydrolyzed by rennin to a potential vasoconstrictor angiotensin II (Ang II) and plays a role at inactivation of vasodilator bradykinin [7]. ACE inhibitors prevent these effects by giving a reaction with the enzyme. Therefore, antihypertensive peptides show vasorelaxing effect on blood vessels [7].

Some of the bioactive peptides in rubisc/o of *Caulerpa racemosa* have an positive impact on health. They have antioxidative effects on oxidative compounds like reactive oxygen species (ROS) such as hydrogen peroxide and singlet oxygen and free radicals. These antioxidative peptides have many abilities such as radical scavenging, inhibition of lipid peroxidation and metal ion chelation properties [7-11]. Mechanism of action and...
efficiency of antioxidative peptides are based on their physicochemical properties. For example, lower molecular weight, a higher content of His and hydrophobic amino acids make them more effective [11].

Another bioactive peptides determined in rubisc/o are DPP-IV inhibitors. The number of the people affected from type2 diabetes has been increased and DPP-IV takes a place at the formation of this type of diabetes [12]. Inhibition of DPP-IV has a significant impact on this illness [13]. Incretin hormones such as glucose dependent insulninsotropic peptide (GIP) and glucagon-like peptide-1 (GLP-1) are inactivated by DPP-IV. DPP-IV causes to decrease insulin secretion after glucose intake [14]. Bioactive peptides such as Xaa-Pro, Pro-Xaa and Xaa-Ala inhibit this enzyme competitively [13].

Prolyl endopeptidase (PEP) inhibitor peptides are assigned as anti-ammesia. PEP is extensively found in many kinds such as bacteria, mammalian and human tissues and they show the biggest performance in skeletal muscle and in the brain, especially in the cortices [15].

This enzyme causes to degradation of bioactive peptides at the carboxyl side of Proline residues [16]. The formation of some disorders of nervous system is associated with this enzyme [15]. Beside Alzheimer’s disease, PEP also leads to Parkinson’s disease by mutation or escalation of the aggregation of a-synuclein protein [16]. Furthermore, PEP cleavages the proline-containing neuropeptides which is used at learning and memory process. In the case of mental diseases, such as PTSD, bipolar disorder, eating disorders and amnesia, PEP activity is detected at higher amounts [15-17].

The other important bioactive peptides found in rubisc/o of Caulerpa racemosa activate ubiquitin mediated proteolysis (UbMP). The ubiquitin-proteasome system is the major mechanism at regulatory and aged protein degradation. The ageing related researches indicate that as a result of the reduction in proteasome activity, failures at the protein degradation leads to neurodegenerative diseases such as Alzheimer’s and Parkinson’s disease [18]. In addition to ACE inhibitor peptides, antithrombotic peptides are also effective on the cardiovascular system. One of the important cardiovascular diseases is thrombosis and it is caused by two main routes: blood clotting and platelet aggregation [19].

While many other compounds act on platelet activation, mostly it occurs with the help of a protease thrombin which functions with fibrinogen. The activity of antithrombotic peptides inhibits fibrinogen binding [20]. Lastly, other founded bioactive peptides show regulating activity such as stomach mucosal membrane activity and ion flow regulating.

The frequency of the occurrence of these bioactive fragments in a protein chain (A) is determined as the ratio of the number of cryptic bioactive peptides within a protein sequence (a) to the number of amino acid residues in a protein (N) [3-10]. Also, after finding A values, the modified frequency of bioactive peptides, which are released after in silico proteolysis, is calculated as $A'$ for each enzyme. The only difference in the calculation is a value, it is changed to 'a' and it indicates total number of bioactive peptides [3].

\[ A = a/N \] (1)

According to this, frequency of occurrence of bioactive peptides in rubisc/o of Caulerpa racemosa is estimated (Table I).

<table>
<thead>
<tr>
<th>Activity</th>
<th>Frequency of Occurrence</th>
<th>A values</th>
<th>A' values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chymotrypsin</td>
<td>0.1906</td>
<td>0.0202</td>
<td></td>
</tr>
<tr>
<td>Trypsin</td>
<td>0.1143</td>
<td>0.0112</td>
<td></td>
</tr>
<tr>
<td>Pepsin</td>
<td>0.1345</td>
<td>0.0426</td>
<td></td>
</tr>
<tr>
<td>Bromelain</td>
<td>0.3655</td>
<td>0.0426</td>
<td></td>
</tr>
<tr>
<td>Thermolysin</td>
<td>0.2915</td>
<td>0.0471</td>
<td></td>
</tr>
<tr>
<td>Papain</td>
<td>0.4484</td>
<td>0.0583</td>
<td></td>
</tr>
</tbody>
</table>

Following this estimation, bioactivity of these peptides is calculated by Peptide Ranker (Table II) and it estimates their effect level between 0.1-1.0 points [21].

<table>
<thead>
<tr>
<th>Peptide Activity Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioactive Peptides</td>
</tr>
<tr>
<td>PWG</td>
</tr>
<tr>
<td>GF</td>
</tr>
<tr>
<td>FR</td>
</tr>
<tr>
<td>FG</td>
</tr>
<tr>
<td>MP</td>
</tr>
<tr>
<td>BG</td>
</tr>
<tr>
<td>PG</td>
</tr>
<tr>
<td>PL</td>
</tr>
<tr>
<td>GL</td>
</tr>
<tr>
<td>GR</td>
</tr>
<tr>
<td>LG</td>
</tr>
<tr>
<td>RL</td>
</tr>
<tr>
<td>JP</td>
</tr>
<tr>
<td>AG</td>
</tr>
<tr>
<td>LY</td>
</tr>
</tbody>
</table>

CONCLUSIONS

By using BIOPEP tool, five different bioactive peptide kinds were found in rubisc/o of Caulerpa racemosa. Fifty-three different bioactive peptides were recognized based on in silico analysis (Table III). There have been found as forty-three ACE inhibitor bioactive peptides while other peptides changed between antioxidative and other types.

In conclusion, in silico analysis of rubisc/o of C.racemosa for bioactive peptides reveals that biomass of this invasive alga could be evaluated in food industry as a natural and alternative bioactive food source. Similar investigation is also strongly recommended for other members of Caulerpa Genus.
REFERENCES


<table>
<thead>
<tr>
<th>Chymotrypsin</th>
<th>Position</th>
<th>Effect</th>
<th>Trypsin</th>
<th>Position</th>
<th>Effect</th>
<th>Pepsin</th>
<th>Position</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[266-267]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[457-458]</td>
<td></td>
</tr>
<tr>
<td>VAY</td>
<td>[231-233]</td>
<td>ACE inhibitor</td>
<td>QK</td>
<td>[440-441]</td>
<td>ACE inhibitor</td>
<td>GF</td>
<td>[256-257]</td>
<td>ACE inhibitor</td>
</tr>
<tr>
<td></td>
<td>[482-483]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[482-483]</td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td>[234-235]</td>
<td>ACE inhibitor</td>
<td>YK</td>
<td>[210-211]</td>
<td>ACE inhibitor</td>
<td>GL</td>
<td>[311-312]</td>
<td>ACE inhibitor</td>
</tr>
<tr>
<td>GF</td>
<td>[256-257]</td>
<td>antioxidantive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TY</td>
<td>[151-152]</td>
<td>antioxidantive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL</td>
<td>[311-312]</td>
<td>antioxidantive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Thermolysin</th>
<th>Position</th>
<th>Effect</th>
<th>Bromelain</th>
<th>Position</th>
<th>Effect</th>
<th>Papain</th>
<th>Position</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[527-528]</td>
<td></td>
<td></td>
<td></td>
<td>regulating*</td>
<td></td>
<td>[404-405]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[220-221]</td>
<td></td>
<td></td>
<td>[220-221]</td>
<td></td>
<td></td>
<td>[231-232]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[220-221]</td>
<td></td>
<td></td>
<td>[231-232]</td>
<td></td>
<td></td>
<td>[231-232]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[389-390]</td>
<td></td>
<td></td>
<td>[277-278]</td>
<td></td>
<td></td>
<td>[277-278]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[228-229]</td>
<td></td>
<td></td>
<td>[562-563]</td>
<td></td>
<td></td>
<td>[562-563]</td>
<td></td>
</tr>
<tr>
<td>LG</td>
<td>[396-397]</td>
<td>ACE inhibitor</td>
<td>IA</td>
<td>[228-229]</td>
<td>antioxidative</td>
<td>IA</td>
<td>[228-229]</td>
<td>antioxidative</td>
</tr>
<tr>
<td></td>
<td>[481-482]</td>
<td></td>
<td></td>
<td>[449-450]</td>
<td></td>
<td></td>
<td>[449-450]</td>
<td></td>
</tr>
<tr>
<td>LQ</td>
<td>[540-541]</td>
<td>ACE inhibitor</td>
<td>LG</td>
<td>[310-311]</td>
<td>ACE inhibitor</td>
<td>SG</td>
<td>[519-520]</td>
<td>ACE inhibitor</td>
</tr>
<tr>
<td>LN</td>
<td>[374-375]</td>
<td>ACE inhibitor</td>
<td>EG</td>
<td>[582-583]</td>
<td>ACE inhibitor</td>
<td>LG</td>
<td>[310-311]</td>
<td>ACE inhibitor</td>
</tr>
<tr>
<td></td>
<td>[417-418]</td>
<td></td>
<td></td>
<td>[366-367]</td>
<td></td>
<td></td>
<td>[366-367]</td>
<td></td>
</tr>
</tbody>
</table>

Regulating*: Stomach mucosal membrane activity regulating Ion Flow regulating
UBMP Activating**: Ubiquitin mediated proteinolysis activating
DPP-IV inhibitor***: Dipeptidyl peptidase-IV inhibitor
Regulating****: Ion Flow regulating
A possible potential hormonal marker for the disorders of bone density in children with type 1 diabetes mellitus

Hedef Dhafir EL-Yassin*, Dunia Adnan, Muneeb Ahmed Al-Zubaidy

Abstract—Preptin has recently been discovered as a peptide secreted with another peptide amylin along side with insulin. Both peptides are anabolic to osteoblasts. Preptin appears to act as a physiological amplifier of glucose mediated insulin secretion. However, its effects on the skeleton is unclear. Production of these peptides is abolished in patients with type 1 diabetes mellitus. As DXA imaging method in the management of children at risk of bone fragility is less certain and as preptin production is impaired in diabetic children, then the level of circulating preptin might contribute in the assessment of bone strength in children. Thirty diabetic (type1) children with a mean age range 6-12 years were selected and excluded any complicated cases. The control group consisted of apparently healthy number and age matched children who were recruited from staff relatives. Five milliliters of venous blood were collected from each child to provide 1.5 ml of serum after centrifugation. For the estimation of preptin and the rest of the biochemical parameters t-test was used to compare the mean values and SPSS version 17 was used for analyses. The results obtained showed that type1 diabetes mellitus has an indirect influence on bone formation through the suppression of preptin secretion. Deficiency of preptin causes increased osteoblast damage and apoptosis. At the same time the degradation of osteoblast leads to leakage of alkaline phosphatase enzyme in to the blood stream cuasing its elevation as was noticed in our results. The study also shows that there is no serum calcium variation due to the fact that excess calcium may be secreted in urine as there is always evidence of hypercalciuria associated with poor glucose control. Results also show a positive correlation between preptin and ALP in both groups patients and control with correlation coefficient of (0.18 and 0.34 respectively)

In conclusion: Results showed that preptin is highly affected (decreased) in type 1 diabetic children leading to a possible effect on bone formation and fragility. Thus preptin might be a potential marker for bone disorders especially in patient with destructed beta cells and type 1 diabetes mellitus.

INTRODUCTION

Diabetes mellitus is a common metabolic disease with increasing prevalence throughout the world. Frailty fractures owing to low bone strength in diabetics have become increasingly recognized as skeletal complications. Patients with type 1 diabetes mellitus (T1DM), which manifests at an adolescent or young adult age, have inadequate accrual of peak bone mass. Impaired bone formation has been proposed as a major contributing factor. (1) Type 1 diabetes is caused by a lesion in beta cells of the pancreas as a rule produced by an autoimmune mechanism. Beta cells secrete insulin thus in type 1 diabetes there is an absolute lack of insulin.

At least two other peptides, amylin and preptin, are cosecreted with insulin. Insulin and amylin are already known to have effects on bone cells in vitro and in vivo. Both are anabolic to osteoblasts. However preptin on the other hand is a recently discovered 34-amino acid peptide corresponding to Asp69-Leu102 of pro-IGF-IIE peptide (2). It appears to act as a physiological amplifier of glucose-mediated insulin secretion, but its effects on the skeleton are unclear.

Production of these peptides is abolished in patients with T1DM (3)

Figure 1: Impaired osteogenesis in T1DM. Pancreatic β-cell destruction in patients with T1DM prevents secretion of insulin, IAPP and preptin, thereby reducing their effects on the RUNX2 gene. This reduction decreases proliferation and differentiation of MSCs into osteoblasts and their resistance to apoptosis—preventing osteogenesis and bone mass accrual. Moreover, reduced insulin secretion in patients with T1DM prevents stimulation of osteoblasts to produce osteocalcin, which stimulates β-cell proliferation and acts on the testes to produce testosterone, a hormone that increases osteogenesis. Abbreviations: IAPP, islet amyloid peptide; MSC, mesenchymal stem cell; T1DM, type 1 diabetes mellitus (3)

DXA is the standard bone imaging method. In older adults, bone densitometry has been shown to predict fracture risk and reflect response to therapy. The role of densitometry in the management of children at risk of bone fragility is less certain. Ongoing studies are needed to help better define the indications and best methods for assessing bone strength in children and the clinical factors that contribute to fracture risk. (4).

This study is designed to investigate the possible effect of beta cells destruction and the secretion of preptin on osteogenesis. As an attempt to look for a suitable biochemical marker for bone disorders.
**FURTHER INFORMATION**

**Dataset collection:**
This study was conducted at Al-Mansour Teaching Hospital (Outpatient Clinic and Teaching Laboratories) during the period from November to July 2014.

Thirty children diagnosed with type 1 diabetes mellitus (according to WHO criteria) were enrolled in this study. Mean age ranged between 6-12 years. Mean disease duration was 3.6 years. Diabetic children with complications and other hormonal disturbances or celiac diseases were excluded.

Thirty other aged match children with apparently health status were also involved in the study and considered as a healthy control. They were recruited from staff relatives after approval.

The study was carried out with the approval of the medical ethical committee in the ministry of health and the parents for blood sampling.

**Blood samples**
Five milliliters of venous blood were collected between (9-12 A.M). The blood was allowed to clot at room temperature for 30 min. then centrifuged at (3000rpm) for 15 minutes to provide about 1.5mls of serum. The serum was then aspirated and divided into two aliquots in plain tubes one for preptin measurements and the other for the rest biochemical parameters.

Serum preptin levels were measured by enzyme-linked immunosorbent assay. Alkaline phosphatase and calcium were measured spectrophotometrically.

T-test was used to study the differences in means of the studied parameters in both groups. P value of less than 0.05 was considered significant. SPSS (Statistical Package for Social Sciences) version 17 was used for analysis.

**Results and discussion:**

The results showed that type 1 diabetes mellitus has an indirect influence on bone formation through the suppression of preptin secretion. Preptin promotes osteoblast survival that is it has an antiapoptotic effects on osteoblast. Thus its deficiency will lead to increased osteoblast damage and apoptosis.

Osteoblasts are one of the sources of alkaline phosphatase (ALP), an enzyme that has a role in the mineralisation of bone. Thus degradation of osteoblast will lead to the leak of ALP into the blood stream causing its elevation. This is also seen in our results (figure 2).

Maintenance of bone integrity depends on bone remodeling, the well-coordinated balance between bone formation by osteoblasts and bone resorption by osteoclasts. If for any reason there is uncoupling of the components of bone remodeling, such that bone resorption exceeds bone formation, bone loss occurs, leading to bone fragility, premature osteoporosis and fractures.(1)

The study also showed no serum calcium variations in both groups due to the fact that excess calcium may be secreted in urine as there is always evidence of hypercalciuria associated with poor glucose control, which confirmed that the presence of hypercalciuria was due to osmotic diuresis of glycosuria.(5)
CONCLUSIONS
This study was designed to investigate the possible effect of the destruction of beta cells and the secretion of preptin on osteogenesis. Results showed that preptin was highly affected (decreased) in type 1 diabetic children leading to a possible effect on bone formation and fragility. Thus preptin might be a potential marker for bone disorders especially in patient with destructed beta cells and type 1 diabetes mellitus.

ACKNOWLEDGMENT
This study was conducted at Al-Mansour Teaching Hospital (Outpatient Clinic and Teaching Laboratories). The authors thank members of staff of the mentioned departments for providing subjects and equipments to carry on in this study.

REFERENCES
Analyzing Relations Among Measurements of Diabetic Neuropathy

Olcay Kursun*, M. Muzaffer Ilhan, Ahmet Cinar, M. Erdem Isenkul, C. Okan Sakar, A. Esra Gursoy, Ertugrul Tasan, Oleg V. Favorov

Abstract— EMG is a standard method for the diagnosis of neuropathies seen in diabetic patients that can measure various conditions such as the nerve conduction velocity. However, as complementary methods physician examination and patient interviews/questionnaires (such as Michigan Neuropathy Scores - MNS) are also used in the field of neuropathic diagnostics. This study presents a preliminary work on EMG and MNS comparison as a part of a more detailed project aiming at analyzing data collected using multiple diagnostic tests (Blood tests and Cortical Metrics technology along with the aforementioned EMG and MNS scores) from 40 diabetic patients who appealed at the Department of Endocrinology in Medical Faculty Hospital of Bezmialem Foundation University. The preliminary analyses performed show that the EMG and MNS tests are not highly correlated and thus should be used in a complementary manner for the diagnosis and monitoring of diabetic neuropathies. Moreover, it can be concluded that the MNS-patient interview section produces higher correlations with EMG scores as compared to those produced by the MNS-physician examination section.

INTRODUCTION

Diabetes can cause several damages in the nervous system. Although glucose level can be kept under control using insulin or applying dietary cure in approximately half of diabetic patients, neuropathies can develop. Neuropathy induced disorders may increase over time and cause numbness, tingling, and loss of sensation especially in hands and feet [1-2]. Early diagnosis can prevent the disorders originated from diabetic neuropathy. A standard diagnostics of diabetic neuropathies is applying surficial Electromyography (EMG) to the patients in regular periods [2]. In surface EMG, with the help of electrodes placed on the skin, the nerve conduction velocity is measured by sending electrical pulses to the related nerve.

However, EMG is not necessarily the most practical diagnostic method [2] since it is inconvenient for patients and also requires expertise to be applied. Besides, the neuropathies that can be detected by EMG are limited to peripheral nervous system and cannot identify the abnormalities in the central nervous system (CNS). However, in the early stages of diabetes, cortex can be affected from the insulin function abnormalities. Therefore, the measurements aiming at analyzing neuropathies that affect cortex is expected to enable the detection of some early symptoms of diabetes [3]. Therefore, a noninvasive diagnostic device called Cortical Metrics (CM) [4-6] utilizing tactile stimulation that can detect and quantify systemic alterations in CNS and a set of questionnaires and physical examinations (as a package called MNS - Michigan Neuropathy Scores [1, 7]) are applied to the patients to help improve the neuropathy measurements traditionally obtained only by EMG.

Although the dataset collected within this study also contains blood test results of the subjects such as HbA1c and cholesterol, this study is limited to be a preliminary work to determine the relations between MNS and EMG tests only.

MATERIALS AND METHODS

The data were gathered within a period of 3 months (from January to April of 2013) from 40 diabetes patients (10 males and 30 females) with ages ranging from 32 to 71 (56.5±8.9) at the Department of Endocrinology in Medical Faculty Hospital of Bezmialem Foundation University. Firstly, blood samples are taken from the patients for the blood test. The diabetes level of the subject is then determined after the examination of these blood test results by the physician. Then, MNS test is applied to the subjects. After, five different protocols are applied to the subjects using CM. The last data collection step is the EMG test which is performed by a neurologist. All the subjects were informed about the tests and attended the tests voluntarily with the permission of ethical committee of Bezmialem University.

EMG dataset used in this study is composed of 10 EMG measurements. Six of these measurements are obtained only by applying surface EMG test to the subject. The motor latency, motor amplitude, motor conduction, sensory latency, sensory amplitude, and sensory conduction measurements are the numerical values provided by the EMG test. The physician evaluates the EMG test results and identifies the following values: binary polyneuropathy value indicating the existence or absence of polyneuropathy (by thresholding the velocity and the amplitude), binary carpal tunnel syndrome value indicating the existence or absence of carpal tunnel syndrome.

The rest of the measurements are the responses of the subject to a medical examination which is based on MNS questionnaire test performed by the physician. MNS is applied to find out the presence of diabetic neuropathy. The questionnaire consists of 15 “yes or no” questions based on
foot sensation including pain, numbness, and temperature sensitivity. A higher score indicates advanced neuropathic symptoms. The second part of the MNS is a brief physical examination that contains inspection of the feet for deformities, dry skin, hair, or nail abnormalities; callous or infection; semi-quantitative assessment of vibration sensation at the dorsum of the great toe; grading of ankle reflexes; and monofilament testing. Further analyses and evaluation are performed to the patients screening positive on the clinical portion of the MNS (greater than 2 points on a 10 point scale, diagnosed as neuropathic [7]).

**EXPERIMENTAL RESULTS**

Figure 1 shows the MNS scores; note that shown in blue, patients with polyneuropathy have relatively higher MNS scores for both the patient interview and the physician examination. The correlations between EMG and MNS scores are shown in Table I. Moreover, MNS patient questionnaire test scores are somewhat correlated with EMG scores are shown in Table I. Moreover, MNS patient examination. The correlations between EMG and MNS scores for both the patient interview and the physician examination that contains inspection of the feet for deformities, dry skin, hair, or nail abnormalities; callous or infection; semi-quantitative assessment of vibration sensation at the dorsum of the great toe; grading of ankle reflexes; and monofilament testing. Further analyses and evaluation are performed to the patients screening positive on the clinical portion of the MNS (greater than 2 points on a 10 point scale, diagnosed as neuropathic [7]).

**CONCLUSIONS**

This study is a preliminary work of a more detailed project on the diagnosis and monitoring of diabetic neuropathy. In this paper, we focused on the MNS and EMG tests, which are found to be not highly correlated and thus should be used in a complementary manner. It can also be concluded that the patient interview is more relevant to EMG scores as compared to physician examination section. As a part of future research, we aim to enlarge the dataset by collecting data from more subjects.

**ACKNOWLEDGMENTS**

This study was approved by the ethics committee of Medical Faculty of Bemzıalem University (approval date: 05/03/2014, no:5/23). The authors thank the Departments of Endocrinology and Neurology of Bemzıalem University for their help in collecting the data. The continuation of this analysis will involve using the Cortical Metrics device with the support of Mark Tommerdahl from UNC.

This work was supported by TUBITAK-1001 Research Project (Project ID: 114E178) and by Istanbul University BAP project 37275.

**REFERENCES**

Abstract— Cancer outcome prediction aims to forecast disease progression from transcriptome data of patients. Predictive signatures can be constructed by analyzing gene expression data. Integration of gene expression data and biological networks can improve the prediction power of such signatures. In this study, we evaluated a network-based approach using different interaction networks to predict survival time and tumor subtypes of pancreatic cancer patients. The Signal Transduction Score Flow (SITSFlow) algorithm identifies a predictive gene signature by incorporating network topology with given expression data. SITSFlow was applied to Transfac and HPRD interaction networks, as well as unified KEGG pathways. The HPRD interaction network can predict patient outcome classes with 80% accuracy; therefore it is more successful compared to Transfac and KEGG. As conclusion, integration of microarray and network data could improve the outcome prediction of malignant diseases.

I. NTRODUCTION

High-throughput experiments aim to explore predictive signatures, which would provide identification of clinical outcome of diseases. Experimental and computational noises in data as well as limited sample sizes reduce the predictive power and biological interpretability of such signature genes. Integration of network information together with gene expression data is able to provide more efficient signatures for outcome prediction in cancer studies.

In this study, we tested a network-based outcome prediction method using different interaction networks and pancreatic cancer data sets. The Signal Transduction Score Flow (SITSFlow) approach was previously developed to assess the biological activity of processes in KEGG pathways [1]. Therefore, we extended the SITSFlow algorithm to assess the prediction performance on interaction networks. SITSFlow was applied to Transfac regulatory network, HPRD protein-protein interaction network, as well as unified KEGG pathways. Our aim to predict survival time and tumor subtypes of pancreatic cancer patients. The experiments showed that physical protein-protein interaction networks (i.e. HPRD) build more predictive signatures and improve the accuracy on average 7-15% for different outcome classes.

METHOD

The Signal Transduction Score Flow (SITSFlow) algorithm was previously developed to evaluate the biological activity of the processes in KEGG pathways [1]. Main steps of the algorithm are given in the following:

1. Gene expression data is assigned as the initial score of each node in a pathway.
2. An activity score of a node is calculated by summing up the weights of all incoming edges.
3. The node scores are transferred to every path in the pathway.
4. After convergence of the algorithm, the final activity score of a node describes the biological response of this node to the given experimental data.
5. Top 10 nodes with highest activity scores are selected as signature genes.

The SITSFlow algorithm identifies a predictive signature by incorporating network information with given expression data (Fig. 1).

![SITSFlow algorithm overview.](image)

Figure 1. The algorithm was applied on two sets of pancreatic cancer patients:

1. Survival time data: It contains gene expression samples for 30 pancreatic cancer patients [2]. The outcome classes are defined as poor (n=15) or good (n=15) survival, which indicates the average survival time of a patient after the surgery.
2. Tumor subtype data: It has gene expression samples of three sets of patients: pancreatic cancer (n=10), chronic pancreatitis (n=12), normal tissue (n=5). The goal is to discriminate cancer and chronic pancreatitis patients by using gene expression data.
The SiTSFlow algorithm combines network topology and transcriptome data; hence we tested different kinds of biological networks for cancer outcome prediction. Transfac is a network of transcription factors and target genes [3]. It contains 2,400 genes and 5,700 interactions. HPRD covers the physical interactions of human proteins [4]. It contains 10,000 proteins and 40,000 interactions. The unified pathway is constructed by merging of 35 unique signaling pathways from KEGG Database [5]. It has 900 nodes and 1700 interactions.

Our goal is to predict the outcome class of a patient for Survival time and Tumor subtype data sets. The SiTSFlow algorithm selects the most representative 10 genes (i.e. features) using gene expression and network topology to classify patients into pre-defined classes. This signature is used to train a machine learning classifier – Support Vector Machine (SVM). The method classifies given samples into patient classes based on the expression levels of the selected genes. We used the SVM implementation as provided in the R package e1071. The expression of each gene was used as an independent feature to train the classifier.

RESULTS

The prediction accuracy of different biological networks and outcome classes are shown in Fig. 2. In the case of tumor subtype problem, all networks can predict much better than a random predictor (i.e. 50%). The prediction accuracy of HPRD and Transfac networks were quite similar, 84% and 80%, respectively. KEGG pathways can predict with 74% accuracy.

![Figure 2](image)

**Figure 2** Prediction accuracy of three networks (HPRD, KEGG, Transfac) for two cancer data sets (Tumor subtype, Survival time).

The accuracy of all three networks dropped in the case of survival time predictions. Similarly, the HPRD network provides higher prediction accuracy (80%) compared to Transfac (71%) and KEGG (60%). It is clear that outcome prediction problems have different levels of complexities for survival time and tumor subtype classes. The estimation of possible lifetime of a pancreatic cancer patient is more difficult task than tumor subtype prediction. The reason might be the strength of gene expression signal, which looks like much stronger in tumor subtype dataset.

When the effect of network type is analyzed, the most predictive signatures are obtained by the HPRD physical interactions. All three networks cover different kinds of human interaction data (e.g. physical, regulatory or signaling). HPRD contains more connections between human proteins compared to the Transfac network and KEGG pathways. As a conclusion, the SiTSFlow algorithm works more efficiently with a more complete protein-protein interaction network for outcome prediction problem.

![Figure 3](image)

**Figure 3** Expression profile of signature genes selected by Transfac.

We also analyzed gene expression profile of signatures obtained by different networks for various outcome classes. The gene expression of 10 signature genes is plotted in Fig 3. These genes were obtained by using the Transfac network and they can discriminate pancreatic cancer and chronic pancreatitis patients with 80% accuracy. If we use a statistical test to select signature genes (without using any network information), we would not choose most of these genes due to their similar expression profiles for two classes. Therefore, incorporation of network data to select signature genes helps to reduce experimental noise and statistical deficiency.

**Table 1: Biological Proof of Signature Genes**

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>PubMed Reference</th>
<th>GO Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MMP9</td>
<td>19629003</td>
<td>Apoptotic process, cell death</td>
</tr>
<tr>
<td>2. MYC</td>
<td>19022256</td>
<td>Apoptotic process, cell death</td>
</tr>
<tr>
<td>3. CDKN</td>
<td>11751405</td>
<td>Apoptotic process, cell death</td>
</tr>
<tr>
<td>4. HLA-DQB1</td>
<td>19022256</td>
<td>Apoptotic process, cell death</td>
</tr>
<tr>
<td></td>
<td>18153707</td>
<td>Immune response</td>
</tr>
</tbody>
</table>

The biological role of signature genes was validated by searching literature studies and applying GO enrichment. Five signature genes have similar expression profiles reported in previous studies as well (Table I). These genes are annotated with specific biological functions (e.g.
apoptosis, immune response), which are also related with cancer progression. These genes might be further investigated in a wet-lab to check their biomarker potential.

CONCLUSIONS

In this study, a network-based outcome prediction method was assessed on pancreatic cancer datasets using different interaction networks. The prediction performance highly depends on the quality and signal strength of transcriptome data. The interaction data provided by biological networks is another factor, which relatively influences the performance. The predicted signatures were also validated by literature search and functional annotation.

Our experiments proved that network information would decrease the noise in gene expression data analysis. Although such networks cover only 20% of estimated human interactome [6], integration of transcriptome and network data could help to improve cancer outcome prediction. The predicted signatures might be used as potential biomarkers in the future to forecast malignant disease progression.

ACKNOWLEDGMENT

We kindly acknowledge funding by DFG, BMWi (Hybris, GeneCloud).

REFERENCES

Chaos by Neural Networks: The Quasi-periodic Route
Marat Akhmet*, Mehmet Onur Fen and Ayşegül Kıvılcım

Abstract—In the present study, we investigate neural networks that possess chaos with infinitely many unstable quasi-periodic motions in the basis. Retarded shunting inhibitory cellular neural networks and Hopfield neural networks are utilized for the chaos generation and extension processes. The presence of chaotic motions as well as the stabilization of quasi-periodic solutions are demonstrated. The results may be useful in healthcare areas such as medical image analysis, sleep apnea detection, cancer treatment, speech/auditory signal recognition and processing, and in many prediction tasks such as protein secondary structure and protein solvent accessibility.

INTRODUCTION

Neural networks have many applications in various areas such as bioinformatics, healthcare and pattern recognition. The main feature of neural networks in bioinformatics is prediction [1]. Neural networks are preferable as a machine learning tool in two ways. One is the prediction performance is so efficient and fast. The other one is it provides high quality results. Moreover, neural networks have two different characteristics such as learning and testing. For this reason, they have many applications in the clinical diagnosis, image and signal analysis and drug development. Many studies have been conducted on the availability of neural networks to healthcare systems. A neural network model for the diagnosis of the breast cancer is considered in the study of Wu [2]. Moreover, the benefit of neural networks as decision making tools in the field of cancer is assessed in [3]. Especially in healthcare systems the speed and the accuracy of the tool that is used is so important. To illustrate, if the cancer is detected in early stages, the treatment and cure for the cancer become more efficient.

Neural networks are useful rather than many machine learning methods in two ways. One is that the prediction performance is very adequate and fast. The other one is that it provides high quality results. Neural networks are widely applied to the prediction of protein secondary structure which is a regular and repetitive spatially local structural pattern in protein structures. Neural networks have two different features such that they are learning and testing. Neural networks found huge applications in the clinical diagnosis, image and signal analysis and interpretations and drug development. Many studies have been conducted on the availability of neural networks to the healthcare systems. A neural network model which diagnose the breast cancer is presented by [1]. Moreover, the benefit of neural networks as decision making tools in the field of cancer is assessed in [2]. Especially in healthcare systems the speed and the accuracy of the tool that is used is so important. To illustrate, if the cancer is detected in early stages, the treatment and cure for the cancer become more efficient.

The implementation of chaos in neural networks is applicable for separating image segments, information processing, and synchronization of the networks. Moreover, synchronization is observed in the dynamics of coupled chaotic cellular neural networks (CNNs). The performance of CNNs on problems that have local minima in energy (cost) functions is increased by introducing chaos to the system, since chaotic behavior of CNNs can prevent the network from local minima and reach the global optimum. Moreover, chaotic dynamics in CNNs can be observed in the studies of chaotic communication and combinatorial optimization problems.

Quasi-periodic motion is the type of motion executed by a dynamical system containing a finite number (two or more) of incommensurable frequencies. Periodic and quasi-periodic solutions have many fundamental usages in biological and artificial systems, which are central pattern generators, establishing stability properties and bifurcations (leading to the discovery of periodic solutions). For example, the dynamics of the brain activity is investigated by Izhikevich [4] as a system of many coupled oscillators with different incommensurable periods.

Delayed neural networks have applications in many areas such as signal and image processing, associative memories, combinatorial optimization and automatic control. Because of the finite switching speed of the amplifiers, time delays occur during the hardware implementation of neural networks [5,6]. Therefore, it is of prime importance to study neural networks with time delays.

In the present paper, we demonstrate the existence of chaos in shunting inhibitory neural networks (SICNNs) [7] and Hopfield neural networks (HNNs) [8] with infinitely many quasi-periodic motions. Li-Yorke chaos

*Corresponding Author (e-mail: marat@metu.edu.tr)
Author One, Author Two and Author Three are with the Department of Mathematics, Middle East Technical University, Ankara, Turkey (e-mails: marat@metu.edu.tr, ofen@metu.edu.tr and kivilcim@metu.edu.tr).
Let us consider the retarded SICNN

\[
x'_{ij}(t) = -a_{ij}x_{ij}(t) - \sum_{c_{h} \in N_{i}(i,j)} c_{hi}^{ht} f(x_{hi}(t - \tau_{1}))x_{ij}(t) + L_{ij}(t) \tag{2.1}
\]

where \(i, j = 1, 2, 3,\)

\[
\begin{bmatrix}
    a_{11} & a_{12} & a_{13} \\
    a_{21} & a_{22} & a_{23} \\
    a_{31} & a_{32} & a_{33}
\end{bmatrix} = \begin{bmatrix}
    4.5 & 2.5 & 3.4 \\
    1.2 & 4.1 & 2.6 \\
    1.7 & 2.8 & 3.2
\end{bmatrix},
\]

\[
\begin{bmatrix}
    C_{11} & C_{12} & C_{13} \\
    C_{21} & C_{22} & C_{23} \\
    C_{31} & C_{32} & C_{33}
\end{bmatrix} = \begin{bmatrix}
    0.003 & 0.007 & 0.001 \\
    0.008 & 0.002 & 0.004 \\
    0.005 & 0.009 & 0.006
\end{bmatrix},
\]

\[f(s) = s^{3/2}\] and \(\tau_{1} = 0.3\).

In order to achieve chaos with infinitely many unstable quasi-periodic motions in the network (2.1), we take into account the external inputs \(L_{ij}(t)\) in the form of relay functions with chaotically changing switching moments. More precisely, we set \(L_{ij}(t) = R_{ij}(t, \theta, \mu)\), where

\[
R_{ij}(t, \theta, \mu) = 1.5 + 0.5|\sin(\sqrt{2} k) + 0.4 \sin(\pi k)|
\]

for \(\theta_{2k}(\mu) < t \leq \theta_{2k+1}(\mu)\) and

\[
R_{ij}(t, \theta, \mu) = 0.8 + 0.2|\sin(\sqrt{2} k) + 0.4 \sin(\pi k)|
\]

for \(\theta_{2k-1}(\mu) < t \leq \theta_{2k}(\mu)\). Here, \(\mu\) is a parameter and the sequence \(\theta = \{\theta_{k}(\mu)\}\) is defined by the formula \(\theta_{k}(\mu) = k + \zeta_{k}(\mu), k \in \mathbb{Z},\) where the sequence \(\{\zeta_{k}\},\) \(\zeta_{0} \in [0, 1],\) is generated through the logistic map

\[
\zeta_{k+1} = \mu \zeta_{k}(1 - \zeta_{k}) \tag{2.2}
\]

There exists a sequence \(\{\mu_{q}\}, q \in \mathbb{Z},\) of period-doubling bifurcation values of the logistic map (2.2), and we need those values of the parameter \(\mu\) between 3.57 and 4 such that the period-doubling cascade accumulates there to provide the chaotic structure [15]. The interval \([0, 1]\) is invariant under the iterations of the map for the values of the parameter between 1 and 4 [15]. Moreover, the sequence \(\theta = \{\theta_{k}(\mu)\}\) of switching moments are the same for each cell of the SICNN (2.1). It is worth noting that for any periodic sequence \(\theta_{k}(\mu)\) the function \(R_{ij}(t, \theta, \mu)\) is quasi-periodic due to the incommensurate periods of the sequences \(\{\sin(\sqrt{2} k)\}\) and \(\{\sin(\pi k)\}\). As \(\mu\) increases through the bifurcation values \(\mu_{q}, q \in \mathbb{N}\) the network (2.1) undergoes a bifurcation such that new stable quasi-periodic motions twice the number of previous existing ones appear. For a value of the parameter \(\mu\) such that the logistic map (2.2) is chaotic through period-doubling cascade, the network (2.1) is also chaotic because of the chaotic structure of the switching moments. Therefore, SICNN (2.1) possesses chaos with infinitely many unstable quasi-periodic motions for such values of the parameter. We call this chaotification mechanism as the quasiperiodic route.

In what follows, we take \(\mu = 3.8\) such that the map (2.2) is chaotic through period-doubling cascade [15, 18]. Define the constant function \(u(t) = \{u_{ij}(t)\}, i, j = 1, 2, 3,\) with \(u_{11}(t) = 0.985, u_{12}(t) = 0.612, u_{13}(t) = 0.481,\) \(u_{21}(t) = 1.149, u_{22}(t) = 0.415, u_{23}(t) = 0.593, u_{31}(t) = 0.838, u_{32}(t) = 0.559\) and \(u_{33}(t) = 0.504.\) We use the sequence \(\theta\) with \(\theta_{0} = 0.83,\) and represent in Figure 1 the \(x_{21}\) coordinate of the network (2.1) with initial data \(x(t) = u(t)\) for \(0.83 - \tau_{1} \leq t \leq 0.83.\) Figure 1 reveals that SICNN (2.1) is chaotic. Similar simulations can be obtained for the remaining coordinates.

![Figure 1. Chaotic behavior in the x21 coordinate of SICNN (2.1).](image-url)
extension of chaos

In this section, we will show that the generated chaos can be extended. For that purpose, we take into account the following Hopfield neural network with delay,

\[
y_i'(t) = y_i(t) - 0.003 \tanh(y_i(t - \tau_2)) + 0.004 \tanh(y_2(t - \tau_2)) + 0.007 \tanh(y_3(t - \tau_2)) + 1.3x_{13}(t) + y_2'(t) = -y_2(t) + 0.004 \tanh(y_2(t - \tau_2)) - 0.012 \tanh(y_2(t - \tau_2)) + 0.006 \tanh(y_3(t - \tau_2)) + 0.9x_{22}(t)
\]

\[
y_1'(t) = -y_1(t) + 0.002 \tanh(y_1(t - \tau_2)) - 0.005 \tanh(y_1(t - \tau_2) - 0.008 \tanh(y_1(t - \tau_2)) + 1.7x_{13}(t),
\]

where \( \tau_2 = 0.4 \).

HNN (3.3) is obtained by using solutions of (2.1) as external inputs. In the absence of the external inputs \( 1.3x_{13}(t), 0.9x_{22}(t) \) and \( 1.7x_{13}(t) \), the network (3.3) admits an asymptotically stable equilibrium point.

In HNN (3.3), we use the solution \( x(t) = \{x_i(t)\} \) of (2.1) with the same initial data and same value of the parameter \( \mu \) as in Figure 1, and we depict in Figure 2 the \( y_2 \) coordinate of (3.3). The initial data \( y_1(t) = v_1(t), y_3(t) = v_3(t) \) for \( 0.83 - \tau_2 \leq t \leq 0.83 \) is used, where \( v(t) = (v_1(t), v_2(t), v_3(t)) \) is a constant function defined as \( v_1(t) = 0.489, v_2(t) = 0.281, v_3(t) = 0.667 \). Figure 2 reveals that the chaos of (2.1) is extended by HNN (3.3).

IV. Stabilization of Quasi-Periodic Motions

The main source of chaotic motions in the coupled neural network (2.1)+(3.3) is the logistic map (2.2). It is possible to control the chaos of (2.1)+(3.3) by controlling the logistic map. For that purpose, let us use the Pyragas control method [16]-[18]. The aim in this method is to stabilize the unstable fixed point \( 1 - 1/\mu \) of the map (2.2) by using the delayed feedback algorithm \( \kappa_{k+1} = \mu \kappa_k (1 - \kappa_k) + \gamma (\kappa_k - \kappa_{k-1}) \), where \( \gamma \) is the feedback gain [16]-[18]. The fixed point becomes stable for \( \frac{\mu - 1}{2} < \gamma < 1 \). The optimal value of the feedback gain is \( \gamma_{op} = \mu - 2(\mu - 1)^{1/2} \) [17]. For \( \mu = 3.8 \), we have \( \gamma_{op} \approx 0.453 \).

We use the Pyragas control method around the fixed point 2.8/3.8 of the logistic map to stabilize the corresponding quasi-periodic solution of the coupled network (2.1)+(3.3). The simulation result is represented in Figure 3. The control is switched on at \( t = \theta_{50} \) and switched off at \( t = \theta_{150} \). The control becomes dominant approximately at \( t = 55 \) and prolongs till \( t = 200 \), after which the irregular behavior develops again. The \( x_{21} \) and \( y_2 \) coordinates of the networks (2.1) and (3.3) are depicted in Figure 3, where the stabilized quasi-periodic motion is observable for \( 55 < t < 200 \). The optimal value 0.453 of the feedback gain is used in the simulation. Figure 3 confirms the presence of quasi-periodic motions in the chaotic attractor of the coupled network (2.1)+(3.3). Moreover, it reveals that the Pyragas control method applied to the logistic map is appropriate to stabilize the quasi-periodic motions in the continuous-time dynamics of the retarded neural networks.

![Figure 2: Extension of chaos by HNN (3.3).](image)

![Figure 3: The unstable quasi-periodic solution of the coupled network (2.1)+(3.3).](image)

REFERENCES


121


Abstract—Parkinson’s disease is a chronic neurodegenerative disease caused by the reduction of the dopamine producing cells in the brain. The major symptoms of Parkinson’s disease are bradykinesia, rigidity, tremor, and postural instability. In this study, we propose a low cost system that diagnosis Parkinsonian tremor with machine learning where tremor data is acquired with Nintendo Wii Remote. Instead of calculating the frequency of the tremor, we extract features from the accelerometer data and learn them with Support Vector Machines. The method is evaluated on a dataset consisting 50 subjects where 30 of them have Parkinson disease and others are normal subjects. The experimental results of the system are promising.

Introduction

Parkinson’s disease is a chronic neurodegenerative disease caused by the reduction of the dopamine producing cells in the brain. Dopamine is a neurotransmitter involved in movement control [1] and lack of dopamine results in loss of motor and non-motor functions while increasing slowness and rigidity. It is estimated that over 1150 000 people aged over 65 in Europe are suffering from Parkinson’s disease and this number is increasing each year [2].

The major symptoms of Parkinson’s disease are bradykinesia, rigidity, tremor, and postural instability. Tremor is the unintentional oscillating motion that appears involuntary and it is the unilateral shaking of the foot, hand, chin, etc. It generally occurs at a frequency between 4 and 6 Hz [4]. Resting tremor is the type of the tremor that occurs at rest and dissolves with action and sleep. Bradykinesia is the reduction at the speed of movements that causes slow walking. Rigidity is a resistance to passive movement and means stiff or inflexible muscles. Postural instability is the loss of balance and postural reflexes.

The clinical diagnosis of Parkinson’s disease is heavily performed according to the Unified Parkinson Disease Rating Scale (UPDRS). UPDRS assesses 42 aspects of PD which is split into categories including mentation, behavior, activities of daily living, etc. Each aspect is evaluated and given a score from 0 to 5. The clinicians evaluate the major symptoms according to their background and diagnose the disease. This is subjective and only gives information about a limited time of the patient. Therefore, a system that gives quantitative information about the symptoms would help to evaluate the Parkinson’s disease.

In the literature, many sensor based systems are developed for the diagnosis and monitoring of Parkinson’s disease. Various instruments like accelerometer, EMG, and gyroscope are used for tremor detection [18], [19], [20]. The accelerometer is the most commonly used instrument because of providing reliable tremor indices [6].

Nintendo Wii Remote (Wiimote) is the controller of the Nintendo Wii console and it has a three axis ADXL330 accelerometer [7]. The accelerometer has a range of ± 3 G which is sufficient for tremor recording. Therefore, in this study we prefer to employ the accelerometer of the Wii Remote which is low-cost and easy-to-use for tremor data acquisition.

In this study, we propose a machine learning and Wiimote based system for diagnosing the Parkinson’s disease. In the method, after resting tremor of the subjects are recorded by the accelerometer of the Nintendo Wii, various features are extracted from the accelerometer values. These features are trained and tested with Support Vector Machines (SVM) and each subject is classified/diagnosed as having Parkinson’s disease or not.

The proposed system is evaluated and tested on a dataset which consists of the tremor data of 50 different subjects where 30 of them have Parkinson’s disease. The results are promising.

The rest of the paper is organized as follows: Section II includes the related work about the sensor based systems for Parkinson’s disease. Section III introduces the method. The experiments are presented in Section IV. Finally, Section V concludes the study.
**RELATED WORK**

In the last decade, various studies have been performed for evaluating motor symptoms in people with Parkinson’s [8], [9].

In [10], two 3D gyroscopes put on each wrist were used to record data of 10 different subjects and bradykinesia and tremor quantification were performed with a method based on spectral analysis.

Rigas et al. [11] detected tremor with an accelerometer based method and quantified tremor severity with 87% accuracy and discriminated tremor from other Parkinsonian motor symptoms during daily activities.

Patel et al. [13] put eight accelerometers to the subjects’ arms and legs and predicted the severity of tremor, bradykinesia, and dyskinesia with a dataset containing 12 subjects. They extracted features after high-pass filtering and the time series were segmented using a rectangular window and estimated the severity with SVM.

Geman and Zamfir [12] proposed a system based on wavelet analysis for identifying normal tremor from Parkinsonian tremor. They tested their system on a dataset containing 82 people where 52 of them with Parkinson’s disease. They also presented a tremor and gait system [16] that employs Wiimote and RGBD sensors for evaluating the motor symptoms of Parkinson’s disease.

Zengin et al. [14] presented a system for measuring the intensity of human essential tremor in a qualitative manner. The system is built on a ring which contains a three-axis accelerometer, a microcontroller, and a memory unit. In [15], a system for home based monitoring of subjects that have Parkinson’s disease is developed. In the system, a number of games are combined with a diary with the metrics gathered from Wiimote and the quantitative monitoring of the subjects are performed from this diary.

**METHOD**

In order to record the tremor using Wiimote, we develop a graphical user interface system with WiiLab library [17]. The Wiimote is connected to a computer with Bluetooth.

A subject holds the Wiimote in the resting position in his hand and the tremor data is acquired for 60 seconds. The sampling duration of the accelerometer is 100 msec and the system can analyze the tremor under 10 Hz. The data acquired for a subject has 600 sampled tremor values for each axe. The gravitational acceleration is removed from the data before processing.

After gathering the tremor data, the features are extracted for training. We use all of the data gathered from 3 axis (x, y, and z) and the tremor magnitude vector $T(t)$ which is calculated as

$$T(t) = \sqrt{\text{acc}_x(t)^2 + \text{acc}_y(t)^2 + \text{acc}_z(t)^2}$$  \hspace{1cm} (1)

where $\text{acc}_x(t)$, $\text{acc}_y(t)$, and $\text{acc}_z(t)$ are the accelerometer data at time $t$ in axis X, Y, and Z, respectively.

The features are extracted with a windowing technique. For each window, the average, minimum, maximum, and the standard deviation of the data of $\text{acc}_x$, $\text{acc}_y$, $\text{acc}_z$ and $T(t)$ are extracted. Also, Fast Fourier Transform is applied to each window and the power spectral energy is calculated and used as a feature.

The extracted features are trained and tested with SVM. We performed a leave-one-out approach; thus each subject is tested once where the other 49 subjects are used for training. Each subject is classified as having Parkinson’s disease or not with SVM.
EXPERIMENTS

We studied on 50 subjects where 30 of them have Parkinson’s disease (PD) and 20 of them are healthy. The demographic properties of the subjects are given in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>PD</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Age (average)</td>
<td>63±21</td>
<td>65±13</td>
</tr>
<tr>
<td>Gender(male/female)</td>
<td>21 male</td>
<td>10 male</td>
</tr>
<tr>
<td></td>
<td>9 female</td>
<td>10 female</td>
</tr>
</tbody>
</table>

After extracting features and classifying the subjects with leave-one-out approach by SVM, we measured three main metrics which are

\[
\text{Precision} = \frac{TP}{TP+FP} \quad (2) \\
\text{Recall} = \frac{TP}{TP+FN} \quad (3) \\
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)
\]

where \( TP \) is the number of true positives, \( FN \) is the number of true negatives, \( FN \) is the number of false negatives and \( TN \) is the number of true negatives. The Precision of our system in 0.93, the Recall is 0.93 and the Accuracy is 0.92. The accuracy of 0.92 means that 46 of the subjects are classified accurately by our system where 4 of the subjects are classified inaccurately.

<table>
<thead>
<tr>
<th></th>
<th>Parkinson</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parkinson</td>
<td>28 (TP)</td>
<td>2 (FN)</td>
</tr>
<tr>
<td>Healthy</td>
<td>2 (FN)</td>
<td>18 (TN)</td>
</tr>
</tbody>
</table>

The confusion matrix of the SVM is shown in Table 2. 28 of the 30 subjects that have Parkinson’s disease are diagnosed accurately; however 2 of them are diagnosed as normal although they have Parkinson’s disease. 18 of the healthy subjects are classified accurately. The Receiver Operating Characteristic (ROC) curve of the proposed system is shown in Figure 2.

The experimental results indicate that the proposed system diagnoses Parkinson’s disease with accuracy of 92% and it can be used for clinical assessment by the doctors. Since it is a low cost system, it can be used for Parkinson’s disease scanning in public places.

CONCLUSIONS

Various sensors are used for diagnosing, evaluating, and monitoring the motor symptoms of Parkinson’s disease. In this study, we proposed a low-cost Wiimote based system for diagnosing Parkinson’s disease. Instead of calculating the tremor frequency, we extracted simple features from the accelerometer data and learned them with SVM. The system is tested on a dataset containing 50 subjects and it has 92% accuracy. The experiments show that, the proposed system can be used in real clinical assessment for diagnosing Parkinson’s disease. For future work, we are planning to analyze depth data and video of gaits of the subjects and use them for computer aided diagnosis.

ACKNOWLEDGMENT

This study is supported by Istanbul Medeniyet University BAP project FBA-2013-360 entitled “Evaluation of Parkinson’s disease with computer vision methods”.

REFERENCES

On the comparison of validation techniques in protein classification

Çağın Kandemir-Çavaş* and Selen Yıldırım

Abstract—in this study, we applied fuzzy linear discriminant analysis in order to classify protein structure which is encoded as amino acid composition. After modeling the system, the test process is evaluated by an independent data set, re-substitution test and jackknife (leave-one-out) method. The classification accuracy rates are obtained as 100%, 94.74% and 87.72%, respectively. By the obtained results, we conclude that among these three cross-validation methods, independent data set method gives the best classification rate for the fuzzy discriminant analysis in classifying the intrinsically disorder protein structures.

INTRODUCTION

DNA sequence determines protein sequence, protein sequence determines protein structure and protein structure determines protein function. Protein function is strongly correlated with its structure. It is essential to predict structure of a protein in order to derive the understanding both of the functions of proteins, such as chemical activation of enzymes, and the general principles of protein structure and folding. Primary structure is composed by amino acid sequence, secondary structure is called by the assignment of helices and sheets, tertiary structure is the assembly and interactions of the helices and sheets, quaternary structure is the assembly of the proteins composed of more than one subunit [1-2].

Intrinsically disordered proteins occur when the deformations happen in the tertiary structure of a protein [3]. Disordered proteins play an important role in DNA/RNA/protein recognition, modulation of specificity/affinity of protein binding, molecular threading, activation by cleavage [4]. The disordered proteins can be experimentally identified by Electron microscopy, nuclear magnetic resonance (NMR) spectroscopy and protein crystallography [5-7]. However, all these methods are time-consuming and costly. Therefore, in literature, the use of advanced computational methods to predict protein structure increase due to its rapidity, low-cost solutions and high effective results [8-14].

The linear discriminant analysis (LDA) relies on several assumptions such as normality, homogeneity of variance-covariance matrix... etc. In addition, since ordinary discriminant analysis uses the Euclidean distance to compute the coefficients of discriminant function, overlapped data points must belong to the same classes. However, discriminant analysis becomes appropriate to manage with such overlapped data points via integration with “fuzzy” concept. Fuzzy discriminant analysis (FDA) aims at maximizing the ratio between class variance and within class variance, also tries to find the best discriminating function to represent the classes [15]. This method is used in Bioinformatics field [16-18].

In data mining studies, the performance of the model is evaluated by testing processes. Testing of the model is stated by the validation techniques which are re-substitution, independent data set and jackknife (leave-one-out) methods [19].

In this study, the performance of the FDA on the intrinsically disorder protein structure is evaluated by re-substitution, jackknife and independent data set method. Because of the importance of high prediction accuracy, this study can be a guide for many scientists.

MATERIALS AND METHODS

The Fasta format of 57 ordered and 57 disordered proteins were extracted from DisProt [20].

Since proteins are composed of 20 different amino acids with a variety of shapes, size and chemical properties [21], each sequence of both prokaryotic and eukaryotic proteins are encoded such a 1-by-20 vector by computing the frequency of each amino acid in the related protein sequence [22].

Hence, LDA takes into account all data has the equal importance, whereas fuzzy linear discriminant analysis (FDA) consists of the strategy in which different points have different significance. The general model of the FDA model is obtained by modifying the LDA in terms of identification of membership value of each data point to the related classes. The membership values of the data points are obtained by Fuzzy c-means (FCM) algorithms.

FLDA tries to maximize the following objective function,

\[
J(B) = \max_B \frac{B'S_BB}{B'S_BB}
\]  \hspace{1cm} (1)

where \(S_B\) and \(S_W\) denote the between- and within-class scatter matrix, respectively. And

\[
S_B = \sum_c \left( U_c (m_c - m)(m_c - m)^T \right) \] \hspace{1cm} (2)

\[
S_W = \sum_c \sum_j \left( x_{cj} - m \right) (x_{cj} - m)^T \] \hspace{1cm} (3)

*CorrespondingAuthor (e-mail: cagin.kandemir@deu.edu.tr).
Çağın Kandemir-Çavaş is with the Department of Computer Science, Dokuz Eylül University, İzmir-Turkey.
Selen Yıldırım is with the Department of Statistics, Graduate School of Natural and Applied Sciences, Dokuz Eylül University, İzmir-Turkey.
where $U_{cj}$ is the membership degree of data point $x_j$ to the class $c$; $m_j$ denotes the fuzzy cardinality of the class $c$; $m$ is the fuzzy mean vector of a class $c$; is the fuzzy mean vector of the whole data points that are calculated as follows:

$$U_c = \sum_j u_{cj} \quad \text{(4)}$$

$$m_c = \left( \sum_j u_{cj} x_j \right) / U_c \quad \text{(5)}$$

$$m = \left( \sum_j \sum_{c} u_{cj} x_j \right) / U_T \quad \text{(6)}$$

where

$$U_T = \sum_c U_c \quad \text{(7)}$$

As in LDA, the objective function can be maximized by obtaining $B$ matrix given by the eigenvectors $v$ which corresponds to nonzero eigenvalues $\lambda$ of $S_w^{-1}S_B$ matrix. The result is the solution of this generalized eigenvalues problem.

In statistical prediction methods, an independent dataset, jackknife test or re-substitution test which are the kinds of cross-validation techniques have been used by many researchers. In resubstitution method, the classification model is computed for all data and validity is tested on the same data by using the model. Jackknife is a cross-validation technique. A jackknife test is used for evaluation of the algorithm performance within each data point behave as a test data and the remaining acts as training data point. From the top to end of the data set, each data point orderly behaves as a test data. In independent data set method, some of the data is separated as testing and some as training. In this study 60% of the data is used as training, the remaining 40% is used as testing process of the cross-validation. In this study we evaluated the performance of fuzzy discriminant analysis by using these three cross-validation techniques.

### Table 1: Accuracy Rate of FDA for Resubstitution Method

<table>
<thead>
<tr>
<th>Fuzzy Discriminant Analysis</th>
<th>108</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td></td>
</tr>
<tr>
<td>False</td>
<td>6</td>
</tr>
<tr>
<td>Prediction Percentage</td>
<td>94.74%</td>
</tr>
</tbody>
</table>

### Table 2: Accuracy Rate of FDA for Jackknife Method

<table>
<thead>
<tr>
<th>Fuzzy Discriminant Analysis</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td></td>
</tr>
<tr>
<td>False</td>
<td>14</td>
</tr>
<tr>
<td>Prediction Percentage</td>
<td>87.7193%</td>
</tr>
</tbody>
</table>

### Table 3: Accuracy Rate of FDA for Independence Data Set Method

<table>
<thead>
<tr>
<th>Fuzzy Discriminant Analysis</th>
<th>114</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td></td>
</tr>
<tr>
<td>False</td>
<td>0</td>
</tr>
<tr>
<td>Prediction Percentage</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Conclusions

As seen from the Tables I-III, the classification accuracy rates of FDA with independent data set, resubstitution and jackknife are 100%, 94.74%, 87.72%, respectively. Therefore, after modeling the classifier for the specific bioinformatics problem, its validation method must be carefully selected, because of its impression on the performance of the classifier.

In this study, it is obvious that an independent data set used for FDA classifier gives the best accuracy rate for the identification of intrinsically disorder protein.

### References


Mercury Sensing Based On Genetically Encoded Fluorescent Protein and MerR

Canan Özyurt*, Serap Evran, Azmi Telefoncu

Abstract— In this study, we developed a genetically encoded based mercury sensor by using MerR protein and Yellow Fluorescent Protein (YFP). The detection limit of this sensor was found to be 10 nM. The selectivity of the sensor for certain metals such as Ni, Mn, Cu and Zn, was also investigated.

INTRODUCTION

Mercury (Hg) is a universal pollutant and is toxic to the human nervous and endocrine systems (1, 3). Mercury forms organic and inorganic compounds. Mercury salts are found in two ionic state, mercury (I) and mercury (II) (2). Mercury (II), or mercuric salts, is much more prevalent in the nature than mercury (I) or mercurous salts. When these salts dissolve in the water, the mercury is used by living organisms and exhibit toxicity.

Due to the toxicity of mercury on the living systems, sensitive and selective analysis of mercury is very important. Many studies have been carried out for the purpose of analysis of mercury. Strategies that have been used for in vitro mercury detection are voltammetric detection with a modified glassy carbon electrode (4), spectroscopic based methods such as solid phase extraction (SPE) coupled with FTIR spectroscopy (5) and cold-vapor atomic absorption spectroscopy (CVAAS) (6), a combination of gas chromatography and mass spectrometry (7), quantum dots-based techniques (8), luminescent-fluorescent probes (9, 10). Although these systems are appropriate for mercury ion detection, the sensitivity and selectivity of many of these methods are low. Furthermore, expensive equipment and relatively long experimental times are required for these methods. In order to overcome these limitations of other mercury assays, we constructed a genetically encoded fluorescent protein-based sensing system.

In this paper, we describe a genetically encoded fusion protein consisting of YFP and MerR protein, which is highly sensitive and selective for mercury ion. Three cysteine residues are conserved in all MerR proteins, which is associated with mercury ion binding (15). The MerR protein is an important member of a metalloregulatory proteins family and selectively binds metal ions such as Hg$^{2+}$, Pb$^{2+}$, Cd$^{2+}$, Zn$^{2+}$, Cu$^{2+}$ and Ag$^{+}$ with up to picomolar- femtomolar sensitivity (16,17).

Although MerR responses other metal ions, its selectivity for mercury ion is 100 times greater than that for other ions. It has been shown that a mercury sensor can be designed by utilising MerR protein (18, 19). Since the first report on genetically encoded fluorescent protein based sensors (13), they have become a groundbreaking technology in many studies including detection of small molecules and ions (11), monitoring redox dynamics in living cells (12), and determining enzyme activities (14). GFP (Green Fluorescent Protein) and its variants are essential parts of genetically encoded fluorescent protein sensors. YFP (Yellow Fluorescent Protein) is a widely used genetic variant of GFP. For generating MerR:YFP fusion protein, the gene encoding MerR from Thermus thermophilus attached to the N termini of the gene encoding YFP was amplified by using three different PCR reactions. In each reaction, specially designed primers were used. The gene fragments encoding MerR and YFP were connected by an overlap extension PCR-based approach (20). The PCR primers were designed to include GSSGG peptide linker in order to enhance flexibility and provide rational freedom between MerR and YFP (21, 22). The resulting gene fragment was cloned into a pET 21a(+) vector (Novagen) and the purified plasmids were transformed into E.coli T7-Rosetta cells (Novagen). The transformants were grown at 37 °C in Luria-Broth (LB) medium supplemented with 0.15 mg/ml ampicillin. After an OD600 of 0.4-0.6 was reached, the temperature was decreased to 25 °C, and protein expression was induced by 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Following incubation under dark conditions for 18 hours at 200 rpm, the cells were harvested by centrifugation (30 min, 4000 rpm, 4°C). The cells were then resuspended in 50 mM potassium phosphate buffer containing 500 mM NaCl and 5 mM imidazole (pH 7.5). The cell lysate was obtained by sonication, and the insoluble fraction was removed by centrifugation (30 min, 13 000 rpm, 4°C). MerR:YFP fusion protein with a His-tag at its C-terminal was purified by Ni-chelate affinity chromatography (IMAC) using a HisTrapFF crude column (5ml; GE Healthcare). Fluorescence emission spectra of the purified MerR:YFP biosensor was monitored at the excitation wavelength of 488 nm using a Cary Varian Eclipse spectrofluorimeter and fluorescence intensity was measured at the emission wavelength of 527 nm. The changes in fluorescence intensity were recorded at 25°C by addition of increasing concentrations of mercury and other metal ions including Ni$^{2+}$, Mn$^{2+}$, Cu$^{2+}$ and Zn$^{2+}$ independently.

*Corresponding Author (cananozyurt@hotmail.com).
Canan Özyurt, Serap Evran and Azmi Telefoncu are with Department of Biochemistry, Ege University, Izmir, Turkey, (e-mails: serap.evran@ege.edu.tr and azmi.telefoncu@ege.edu.tr).
MerR from *Thermus thermophilus* is a well-known protein, which binds the mercuric salts selectively and sensitively. We constructed a genetically encoded sensor for mercury detection. In the novel approach presented here, the conformational change upon Hg$^{2+}$ binding to MerR protein was monitored by the change in fluorescence intensity of YFP. The detection limit for mercury was determined to be 10 nM.

**CONCLUSIONS**

MerR from *Thermus thermophilus* is a well-known protein, which binds the mercuric salts selectively and sensitively. We constructed a genetically encoded sensor for mercury detection. In the novel approach presented here, the conformational change upon Hg$^{2+}$ binding to MerR protein was monitored by the change in fluorescence intensity of YFP. The detection limit for mercury was determined to be 10 nM.

**REFERENCES**

Abstract—An electrically small and low-profile meander-line implantable antenna is proposed for biomedical applications. The planar inverted-F antenna (PIFA) type design fed by a vertical probe and covers the 401–406 MHz Medical Device Radiocommunications Service and the 433/434 MHz industrial, scientific and medical (ISM) bands. Also, the proposed antenna exhibits uniform radiation patterns at each frequency band. The HFSS-simulated antenna performance is presented in the paper.

INTRODUCTION

Without restricting movement and treatment of patients, the importance of biomedical telemetry that provides control in distance by means of radio signals has been greater than ever. Thanks to biomedical telemetry, symptoms such as microwave monitoring, arrhythmia, sphygmo, body temperature, respiratory rate, blood sugar, oxygen content of blood can be achieved, and diagnosis and treatment of cancer can be done. The Medical Device Radiocommunications Service band (MedRadio, 401–406 MHz) has been recently allocated [1] to this purpose. 433–434 MHz ISM band can also be used for the applications. A typical biomedical telemetry system comprises biosensor appropriate for the particular signals to be monitored, battery-powered, a transmitter and an antenna. Among all the components necessary for implanted telemetry applications, the antenna plays a critical role in obtaining robust communication network and a significant miniaturization of the whole device. In fact, designing miniature implantable antenna would operate in a tissue is a challenging task. In order to obtain realistic antenna performances, the dielectric properties and the geometry of the skin need to be taken into consideration. The other main challenging factors for the design are high tissue conductivity, impedance matching, antenna size, low power requirements and biocompatibility. Several implantable antenna designs, with different characteristics and target applications, have been recently presented [2–9]. A compact antenna for 3.5–4.5 GHz band is proposed in [2], while a printed design for an ingestible application at 1.4 GHz is introduced in [3]. A flexible antenna design has been recently introduced for such applications [4]. Many research groups [5–7] have widely utilized planar inverted-F antennas (PIFAs) to design implantable antennas because of the simple structure and compact size of PIFAs. Cavity slot radiators without integration with active components are presented for the 2.4–2.5 GHz ISM band for on-body and in-body applications in [8] and [9], respectively. Also, a meander-line implantable antenna is proposed in [10].

In this paper, we introduce a compact meander-line implantable antenna for biomedical applications similar to the design in [10]. However, the proposed antenna in this paper has smaller structure and covers MICS and 430 MHz ISM bands as compared to the design in [10] covering MICS and 2.4 GHz ISM bands. The proposed antenna in a PIFA-type configuration provides a wideband performance at 400 MHz with 20% bandwidth. We note that the full-wave analysis of the proposed design has been carried out using Ansoft HFSS v.13, utilizing finite element method. In this document, the simulated return loss and radiation pattern at each respective frequency band are presented.

ANTENNA DESIGN

The proposed antenna configuration with its design parameters is depicted in Fig. 1. A series of parametric studies were carried out to achieve desired antenna performance, particularly for tuning the resonant frequencies and return loss characteristics. In this design process, substrate’s thickness and permittivity, dimensions of meander-lines, dimensions of slot in the ground plane, position of the shorting pin and feeding were varied.

* S. C. Başaran (e-mail: cbasaran@akdeniz.edu.tr).
M. Akdağ, Y.E. Yamaç, Yalcın Albayrak and S. Cumhur Başaran are with department of Electrical & Electronics Engineering, Akdeniz University, Antalya 07058, Turkey (e-mails: ak.metinakdag@hotmail.com, yunusemyaymac@gmail.com, yalbayrak@akdeniz.edu.tr, cbasaran@akdeniz.edu.tr).

132
As seen in the Figure 1(a), the main radiator of the design consists of five stage of meander-lines that covers an area of 17.5×18 mm$^2$ and placed on a full-grounded Rogers RO3210 substrate with 0.64 mm thickness and electrical properties of $\varepsilon_r=10.2$ and tan$\delta=0.003$. A thick slot (17×0.4 mm$^2$) is inserted into the ground plane to tuning the resonant frequency as shown in Fig 1(b). The antenna is fed by a vertical probe placed between the ground plane and one of the meander-lines. Also, a shorting pin whose position determined by optimization is directly connected between one of the meander-lines and the ground plane. This pin is used to achieve electrically small antenna structure like PIFA design. In addition, a superstrate which is same material of the substrate of the antenna is used to prevent from contacting human tissue and conducting plates of the antenna as shown in Fig 1(c). The design of the proposed antenna is simulated with Ansoft HFSS v.13, and the final dimensions of the antenna are recorded as follows (in millimeters): $W=17.5$, $L=18$, $w_1=1.7$, $w_2=1.35$, $w_3=6.35$, $S_L=17$, $S_w=0.4$, $h=0.64$.

The return loss and radiation pattern performances of the proposed design are shown in Fig. 2 and Fig. 3, respectively. As can be seen from the Fig. 2, the proposed antenna provides an impedance bandwidth of 80 MHz (360–440) covering MICS (401–406 MHz) and ISM (433–434 MHz) bands. Also, an omni-directional radiation performance can be achieved at the respective frequency bands.

**CONCLUSIONS**

In the paper, we have introduced an implantable antenna based on PIFA for biomedical telemetry. Main radiator of the antenna is five stage meander-lines fed by a vertical probe. A thin slot is inserted into the ground plane for tuning the desired frequency band. The antenna has a very small structure that covers MICS (401–406 MHz) and ISM(433–434 MHz) bands. Also, an omni-directional radiation performance can be achieved at the respective frequency bands.

**ACKNOWLEDGMENT**

This study was financially supported by Akdeniz University Scientific Research Projects Management Unit.
REFERENCES


Understanding Brain Metabolic Network Changes in case of Glioblastoma Using Computational Systems Biology Approaches

Emrah Özcan and Tunahan Çakır*

Abstract—The changes of brain metabolic network in case of glioblastoma were investigated using computational systems biology approaches. General cancer characteristics, such as Warburg Effect, on cell metabolism were successfully simulated using the brain metabolic model (IMS570).

INTRODUCTION

Computational systems biology methods have been very popular recently in order to interpret high-throughput omics data (metabolome, transcriptome, proteome etc.) obtained for the aim of elucidating complex mechanism of brain and how brain responds to perturbations such as diseases [1]. The use of computational systems biology methods to analyze the nature of the diseases and the corresponding changes in cellular networks (metabolic, transcriptional and signalling) is also called Systems Medicine [2].

Glioblastoma (Glioblastoma Multiforme, GBM) are the most common primary brain tumors in adults with a median survival of 12 to 14.6 months for patients who receive the latest surgical, radiation, and chemotherapy treatment [3]. In this study, effects of the glioblastoma (GBM) tumors on cell metabolism are investigated by using computational systems medicine approaches. To this aim, the brain metabolic model (iMS570) reconstructed by our group [6] was processed using transcriptomic data from Gene Expression Omnibus to obtain a glioblastoma-specific brain metabolic network.

Flux balance analysis (FBA) is a widely used approach for studying metabolic networks, especially the genome-scale metabolic network reconstructions that have been built in the past decade. These network reconstructions contain all of the known metabolic reactions in an organism and the genes that encode each enzyme. FBA calculates the flow of metabolites through this metabolic network, thereby making it possible to predict the growth rate of an organism or the rate of production of a biotechnologically important metabolite [4].

MATERIAL & METHODS

The constructed brain metabolic model (iMS570) possessing 630 metabolic reactions in and between astrocytes and neurons, which are controlled by 570 genes includes central metabolism (glycolysis, pentose phosphate pathway, TCA cycle), lipid metabolism, reactive oxygen species (ROS) detoxification, amino acid metabolism (synthesis and catabolism), the well-known glutamate-glutamine cycle, other coupling reactions between astrocytes and neurons, and neurotransmitter metabolism [5-6].

Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm uses quantitative gene expression data and one or more presupposed metabolic objectives to produce the context-specific reconstruction that is most consistent with the available data [7]. The brain metabolic model (iMS570) was integrated with transcriptomic data of the healthy brain and glioblastoma types (GSE7307 and GSE13041 respectively) from Gene Expression Omnibus [8] using GIMME algorithm. Flux Balance Analysis (FBA) was applied to the condition specific metabolic models representing healthy brain and glioblastoma types obtained by GIMME algorithm (fig. 1).

Figure 1. Our approach to derive condition specific metabolic models for three different glioblastoma types and for the healthy case by applying GIMME algorithm on the reconstructed brain model, iMS570.
All computational processes were done using the COnstraints Based Reconstruction and Analysis (COBRA) Toolbox [9] which is a MATLAB package used for simulating, analyzing and predicting a variety of metabolic phenotypes using genome scale models.

**Conclusions**

Flux distributions obtained using FBA of the healthy brain metabolic model and experimental results for healthy brain have consistencies which reveal the robustness of the model.

The best characterized metabolic phenotype observed in tumor cells is the Warburg Effect, which is a metabolic alteration from ATP generation through oxidative phosphorylation to ATP generation through glycolysis, even under normal oxygen concentrations. This metabolic alteration is called Aerobic Glycolysis which is characterized by high glucose uptake, low oxygen consumption and elevated production of lactate [10]. Lactate release fluxes of GBM models significantly increased, which are consistent with the Warburg effect.

**References**

Genetic Distance Measurements by Using rRNA and miRNA Sequences

Nazife Çevik, M. Erdem Isenkul, Olcay Kurşun

Abstract - Nucleotide sequences of genes are compared for estimating genetic similarity/distance. In order to make these comparisons, the gene to be compared must be common in both species. Ideally, 16S ribosome for prokaryotes and 18S ribosome for eukaryotes are used to obtain such a measure of genetic diversity. However, there are some other sources of genetic information such as miRNAs that can be important in estimating genetic similarities. This study proposes a miRNA based genetic similarity or divergence of a pair of species by using their common miRNA sequences in its calculation.

INTRODUCTION
Phylogenetic trees of species are built using their genetic code similarities. For the calculation of genetic distances, nucleotide sequences of genes are used. Genetic comparisons are realized through one genome instead of whole genome. This gene must be presented in both species to be compared. This gene is 16S ribosome for prokaryotes and 18S ribosome for eukaryotes.

Ribosomes are structures that gather up suitable smaller molecules to carry out proteins. Each ribosome contains protein and ribosomal RNA molecules. Eukaryotic ribosomes have two subunits which are 40S and 60S subunits. Here, S represents the Svedberg unit that is the measurement of sedimentation rate. 40S subunit contains 18S rRNA molecule and several proteins.

Gene sequences that build for 18S rRNA is similar through the organisms. 18S rRNA gene cover the necessary information to produce 18S rRNA molecule. Even though this gene can be preserved or unchanged over centuries, since it has still variations, scientists use these genes to compare sequences of species. By looking at sequence differences of 18S genes, one can conceive how closely related species are. Although 18S rRNA gene is frequently used for phylogenetic search and contains important biological informations, it has a long sequence. microRNA of species also do contain some important genetic information. MicroRNA (miRNA) is a type of single stranded RNA molecule that is approximately 21-23 nucleotides long(shorter than 18SrRNA sequence) and responsible from the arrangement of gene expressions [1-6]. miRNAs are non-coding RNAs, i.e. they are coded by genes that are not translated into proteins but is transcribed by DNA [7]. Primer transcripts known as pri-miRNAs are processed and first transformed into pre-miRNA hairpin loop and then functional miRNA. Researchers have reported in humans that miRNAs regulate many fundamental cellular functions, therefore the abnormal levels of miRNA levels in cells is linked to the development of cancer [8, 9].

As pointed out by study [10], to obtain the phylogenetetic tree based on 16S rRNA sequences, twenty one rRNA sequences were aligned and for each pair of sequences nucleotide differences were counted. By using Jukes-Cantor method, these sequence differences were corrected for multiple nucleotide substitutions. Then, sequence mutations of results were used to infer the phylogenetic tree. Study [11] focuses on calculating pairwise genetic distances for 16S rRNA genes. They filter target variable region within 16S rRNA gene and are interested in if these variable regions have a relationship with the genetic diversity.

In this paper, we intend to show that a different genetic distance measurement based on miRNAs can be applied to genomic sequences. We analyzed the genetic distances by using different combinations of miRNAs, obtained 18S rRNA distances, and then compared all these distances with each other.

MATERIALS AND METHODS
In order to study on miRNAs and 18S ribosomal RNA sequences, firstly, miRNA precursors used in the analysis belong to different species (a total of 9) were derived from mirBase which is a searchable database for all published miRNA sequences and occurrence [12-15]. Totally nine different species are selected for this study shown in Table 1.

TABLE I. Species used to obtain miRNA and 18S ribosomal RNA sequences.

<table>
<thead>
<tr>
<th>Species</th>
<th>Class</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bos Taurus</td>
<td>Mammalia</td>
<td>Artiodactyla</td>
</tr>
<tr>
<td>Cricetulus griseus</td>
<td>Mammalia</td>
<td>Rodentia</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>Mammalia</td>
<td>Primates</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>Mammalia</td>
<td>Artiodactyla</td>
</tr>
<tr>
<td>Equus caballus</td>
<td>Mammalia</td>
<td>Perissodactyla</td>
</tr>
<tr>
<td>Anolis carolinensis</td>
<td>Reptilia</td>
<td>Squamata</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>Mammalia</td>
<td>Carnivora</td>
</tr>
<tr>
<td>Monodelphis domestica</td>
<td>Mammalia</td>
<td>Didelphimorphia</td>
</tr>
<tr>
<td>Taeniopygia guttata</td>
<td>Aves</td>
<td>Passeriformes</td>
</tr>
</tbody>
</table>

*Olcay Kurşun: (e-mail: okursun@istanbul.edu.tr )
Nazife Çevik, M.Erdem Isenkul and Olcay Kurşun, Department of Computer Engineering, Istanbul University, Avcilar, Istanbul, 34320, Turkey (e-mail: ncevik@istanbul.edu.tr, eisenkul@istanbul.edu.tr, okursun@istanbul.edu.tr)
To study on miRNA sequences, these nine species are pair combined to obtain shared miRNAs between species. Shared miRNAs have the same name and function but have different sequences and belong to different (but closely related) species [16]. As seen in Figure 1, species have some number of miRNAs totally, however, when these species combined as pairs, it is seen that they have also some shared miRNAs. Example shown in Figure 1 demonstrates that while the species Canis familiaris and Equus caballus have 323 and 341 miRNAs respectively, they have 222 shared miRNAs which have the same name and function but different sequences.

**Figure 1. Example of Shared miRNAs of Canis familiaris and Equus caballus**

Features used for miRNAs are simply basecount and dimercount numbers derived from miRNA sequences of species. Basecount is the number of A, C, G and U bases of sequences while dimercount is the number of pair-wise combination of A, C, G and U bases.

Secondly, 18S rRNA sequences are obtained by an on-line resource and aligned ribosomal RNA sequence database named SILVA [17]. SILVA provides updated datasets of aligned small and large subunit ribosomal RNA sequences. Resulting 18S rRNA sequences belonging to studied species are used to compute genetic distance over 18S rRNAs between species.

**GENETIC DISTANCE MEASUREMENTS**

Genetic distance is the genetic divergence between species or populations that is measured by some numerical method [18]. In decades, various measures of genetic distance such as Mahalanobis’s $D^2$ statistic, Nei’s distance and Euclidean distance to gene frequency data. In this study we used Euclidean distances between A, AA and whole bases of basecount and dimercount of shared miRNA sequences respectively. Table 2 shows different genetic distance measurement correlations of miRNA and 18S rRNA sequences of the species. It is possible to see the comparison of these measurements to each other from the relevant table. Highest correlation of these measurements is belonging to all base distances of basecount and dimercount features. Shared miRNA genetic distance is also highly correlated to base A and AA distances. These correlations demonstrate that genetic distances of species can be measured from different perspectives as an alternative solution for classical measurement techniques. Figures 2 and 3 present dendograms obtained using the two distance methods, respectively.

**CONCLUSION**

Both rRNAs and miRNAs contain important information genetically. Although most of the studies use ribosomal RNAs to measure genetic distances, we can conclude that miRNA sequences can also be used to obtain some measure of genetic distances between species. The genetic distances obtained by rRNA and by miRNAs do not have high correspondence (only small correlations), thus, miRNAs could also support (as additional/complementary to rRNA) genetic divergence or similarity measure for calculating genetic distances.

**ACKNOWLEDGMENTS**

This work was supported, in part, by Istanbul University BAP project 37275.

**REFERENCES**


Assessing the bias correction methods due to exposure measurement error in the European carbon black study.
Gonca Mert, Roseanne McNamee and Evridiki Batistatou

Abstract—Bias due to exposure measurement error resulting from the collection of error prone observations is a widely encountered problem, particularly in the epidemiological studies. The error caused by the mismeasured exposures in the exposure-response relationship can be eliminated or reduced through the application of some bias correction techniques given replications of exposure measurements. The performances of the most commonly used techniques to deal with the measurement error problem, which are the naïve approach, regression calibration (RCAL) method, simulation extrapolation (SIMEX) method and instrumental variable (IV) method, are compared through at simulation study and in an application from occupational epidemiology. The data are from the European carbon black study which investigated effects of carbon black dust on lung function. The distribution of the true exposure measurements is highly skewed. We would like to know whether the having skewed data affects the performance of the bias correction methods. This paper analyses the carbon black data and simulates normally distributed true exposure variables and lognormally distributed true exposure variables to be able to make a decision in terms of the performance of bias correction strategies.

1. INTRODUCTION

In occupational or environmental studies, collecting data without error is a very common issue: particularly, in which the primary subject of interest is human. In some studies, exposure measurement error has a considerable effect on the predictor variable. Taking not into account of this error leads to attenuation, and it can cause inaccurate results of the study.

Random measurement error problem may lead to attenuation when the true exposure X has not been observed, and an error prone variable has been observed instead of the true exposure in the error model. The mismeasured variable can cause serious effects on the exposure-response association.

Attenuation due to the mismeasured exposures requires correction for measurement error to eliminate bias in order to keep the effectiveness of the study. There have been some adjustment techniques to eliminate or reduce bias due to the measurement error. However, we will focus on four of these methods in this paper. These are the naïve approach, the regression calibration method, the SIMEX method and the instrumental variable method [3].

These conventional methods are applied to a real study, which is the European respiratory carbon black study. The exposure measurement in this study is highly skewed. This brings about the question of whether or not performance of bias correction methods is affected by lognormal data. To be able to answer this question we have simulation studies in which the true exposure is simulated from the normally distributed data and lognormally distributed data. The goal of this study is to mimic observed skewed data to obtain unbiased estimates of exposure effects in the presence of a measurement error scenario.

The main concern of this paper is to estimate the regression coefficient (β) of the true exposure (X), where it is obtained only through the error-prone exposures (W) and response variable (Y) is regressed on the error-prone variables. β explains how large the effects of measurement error are on the true exposure on response variable.

The structure of this paper is as follows: Section 2 gives a brief overview of the bias corrections techniques which are motivated. Section 3 introduces the carbon black study. Section 4 explains lognormal distribution briefly. The next section describes the regression model and the measurement error model. The proposed methods have been evaluated and compared through the simulation study in Section 6. Finally, Section 7 discusses the findings and addresses to the conclusion.

2. BIAS CORRECTION METHODS IN THE PRESENCE OF EXPOSURE MEASUREMENT ERROR

Bias correction techniques to deal with the exposure measurement error have been described in this section. More detailed information with regard to bias correction methods can be found in [3].

2.1 Naïve Method

The response variable Y is regressed on the individual mismeasured variable W by using a generalised linear model as a naïve approach. The estimated regression coefficient is called naive slope $\beta_{\text{naive}}$. The naïve approach does not properly account for measurement error. It takes
into account only covariates in the model. Naïve approach ignores the measurement error.

### 2.2 RCAL

RCAL approach is conceptually straightforward to apply when needed to bias reduction. This method is mainly based upon the information about the measurement error variance regarding independent replicate measurement. It requires at least a subsample of replicated or validation data. The true exposure variable is approximately replaced by its expected value conditional on the observed error-prone measurements (E(X|Z;W)) [3]. Z is known the covariate measured without error like body mass index or age. The next step is making suitable analysis (a logistic regression analysis, for example) and estimating parameters. Estimating the calibration function affects the performance of this approach positively. The amount of measurement error can affect having a precise approximation [9].

### 2.3 SIMEX

SIMEX is a simulation-based method to reduce attenuation due to measurement. Under the SIMEX approach, there is no need to specify the true variable and no need additional data like validation data. However, SIMEX requires a good knowledge about the distribution of the observed data. If the measurement error is low, SIMEX is a good technique for adjusting bias due to measurement error. SIMEX approach only does partial adjustments and the computation time of the SIMEX longer than the RCAL method, it is computationally intensive. Nevertheless, this technique is easy to implement and robust. [10].

The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can also be determined experimentally via simulation [3].

### 2.4 Instrumental Variable (IV) Method

IV is a general method for estimating the effect of X on Y when there are unmeasured confounders in the presence of exposure measurement error. The key idea here is that the parameters in the regression model could be estimated even where we have a lack of information on measurement error variance if we find a suitable instrument. The mismeasured exposure (W) is regressed on the instrumental variable, say T, to estimate true exposure. In the next stage, the response variable is regressed on the estimated true exposure as well as on the other covariates if present [11].

Instrumental variable, say T must fulfil these 3 requirements:
- T is correlated with X
- T is independent of E
- T is independent of Y given X

Provided that all the above requirements are accomplished, IV can be used to address the endogeneity problem and provide a consistent estimate; otherwise, estimation of instrumental variable is biased [1,8,12].

### 3. MOTIVATING EXAMPLE

The motivation in this paper is a study conducted by the European Carbon Black Centre (ECBC) in 1987 was aimed to investigate the relationship between carbon black exposure and respiratory morbidity/mortality in the European carbon black industry [5]. The carbon black study comprises three cross-sectional phases covering around eight years in 7 European countries. The data from the second (1991-1992) and third phases (1994-1995) are used in this paper [6].

In both phases, there were various number exposure measurements which were taken from the each worker. FEV$_1$ (Forced Expiratory Volume in 1 second) was measured showing the performance of the workers’ lung function as well as the information of the age-related variable, height and cumulative smoking situation.

### 4. LOGNORMAL DISTRIBUTION

Assume that the random variable X has lognormal distribution, then ln X , say U has normal distribution with the parameters $\mu$ and $\sigma$, mean and standard deviation, respectively.

$$U = \ln X \sim N(\mu, \sigma^2)$$

The mean of lognormal distribution is defined as follows:

$$E(X) = \exp(\mu + \frac{\sigma^2}{2})$$

and the variance of lognormal distribution is defined as follows:

$$Var(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$$

### 5. REGRESSION MODEL AND MEASUREMENT ERROR MODEL

Consider a simple linear regression model with only one exposure variable, say X and a continuous response variable, say Y

$$Y = \alpha + \beta X + \delta, \quad (1)$$
where $\text{cov}(\delta, X) = 0$ and the variance of $X$ is $V(X)$. Assume that there are continuous exposure variable, $X$ is not observed due to measurement error. Instead, an error prone exposure following the classical error model is obtained.

$$W = X + E$$

(2)

where the measurement error $E$ has normal distribution: $E \sim N(0, V(E))$ and $E$ is independent of $\delta$ and $X$. Assuming nondifferential measurement error $E$ implies that the error prone $W$ has no new information apart from $X$ [2].

The regression coefficient, say $\hat{\beta}_1$ which is obtained by regressing $Y$ on a single $W$ from the ordinary least squares satisfies the following equation

$$E(\hat{\beta}_1) = \lambda \beta$$

(3)

where $\lambda$ is defined as reliability coefficient of $W$ and found by

$$\lambda = \frac{V(X)}{V(X) + V(E)}$$

(4)

Reliability always takes the values between zero (means having a substantial measurement error) and one (zero attenuation) [7].

6.1 Simulation methods

The data are simulated from the normal distribution and lognormal distribution to compare the performance of the bias correction methods in the STATA software program. In all the simulations, we used the regression equation

$$Y = 5.14 - 0.1X + \delta,$$

(5)

where $\delta \sim N(0, 0.15)$ and $X \sim N(5, 1)$, intercept is the mean of the response values in the carbon black study data. The true value of the slope of the regression model is -0.1.

A similar approach is used for lognormal distribution based on the Equation (5), where $\delta \sim N(0, 0.15)$ and $U = \ln X \sim N(1.11, 1)$. Here, the mean and the variance of $X$ are calculated as 5 and 43 from the theory explained in Section 4 ($X \sim N(5, 43)$).

Assume that the true values are not obtained; instead, error prone exposures are taken from the each subject. In both parts of the simulation, the data are simulated with the reliabilities 0.2, 0.5 and 0.8, representing severe, moderate and small measurement error, respectively.

We also assumed that all the subjects have two exposure measurements, say $W_1$ and $W_2$ to maintain the effectiveness of the study.

In the naïve approach, since it ignores the measurement error, the mean of two exposures, say $W = \frac{W_1 + W_2}{2}$ was considered as the one covariate and taken into account for the each subject.

In this case, the reliability of two mean exposures, say $\lambda_2$ is calculated as from the Spearman-Brown Prophesy formula. $\lambda_2$ is higher than the reliability itself [4].

$$\lambda_2 = \frac{2\lambda}{1 + \lambda}$$

In the naïve approach, the reliabilities of two mean exposures are calculated as 0.33, 0.67 and 0.89 as low, moderate and high reliabilities, respectively.

In the IV method, $W_1$ is used the individual exposure, and $W_2$ is used its instrument.

In the RCAL and the SIMEX methods, both two exposures are used to estimate the coefficient of the true exposure.

Table 1. Distributions of the $E$s, $W$s and $\ln W$s with the different reliabilities.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$E$</th>
<th>$W$</th>
<th>$\ln W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>N(0.4)</td>
<td>N(5.5)</td>
<td>N(1.11,5)</td>
</tr>
<tr>
<td>0.5</td>
<td>N(0.1)</td>
<td>N(5.2)</td>
<td>N(1.11,2)</td>
</tr>
<tr>
<td>0.8</td>
<td>N(0,0.25)</td>
<td>N(5,1.25)</td>
<td>N(1.11,1.25)</td>
</tr>
</tbody>
</table>

The above table shows that 1000 subjects are simulated from these distributions.

6.2 Simulation Results

Table 2. Performances of the bias correction methods with the various reliabilities based on the true value $\beta = -0.1$ for the normally distributed data

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Naïve method</th>
<th>RCAL</th>
<th>SIMEX</th>
<th>IV method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>-0.0331 (0.0073)</td>
<td>-0.1023 (0.0271)</td>
<td>-0.0538 (0.0119)</td>
<td>-0.1019 (0.0337)</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.0665 (0.0101)</td>
<td>-0.0999 (0.0155)</td>
<td>-0.0897 (0.0138)</td>
<td>-0.1003 (0.0176)</td>
</tr>
<tr>
<td>0.8</td>
<td>-0.0889 (0.0115)</td>
<td>-0.0998 (0.0130)</td>
<td>-0.0992 (0.0130)</td>
<td>-0.0997 (0.0138)</td>
</tr>
</tbody>
</table>
In the above table, the means of the estimates of $\hat{\beta}$s and their SDs in the parenthesis are presented with the different reliabilities. Here, the exposure measurements are simulated from the normal distribution, across 1000 simulations. Obviously, the naïve approach has the worst performance since it ignores the measurement error. However, as the reliability increases, the estimates are getting closer to the true value. Secondly, it can be easily said that the RCAL has acceptable error under the low reliability and the high reliability, and performs better under the moderate reliability. The estimate of the SIMEX method is substantially biased when the reliability is 0.2. In spite of the fact that as reliability increases, its performance increases, the RCAL method still gives better estimates. In terms of the SDs, SIMEX performs better than RCAL. Finally, the IV method performs well when measurement error is moderate. However, when the measurement error is severe or low, the performances have acceptable error.

Table 3. Performances of the bias correction methods with the various reliabilities based on the true value $\beta = -0.1$ for the lognormally distributed data

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Naïve method</th>
<th>RCAL</th>
<th>SIMEX</th>
<th>IV method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>-0.0015</td>
<td>-0.0155</td>
<td>-0.0026</td>
<td>-0.0172</td>
</tr>
<tr>
<td></td>
<td>(0.0002)</td>
<td>(0.4184)</td>
<td>(0.0014)</td>
<td>(283.209)</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.0293</td>
<td>-0.0726</td>
<td>-0.0450</td>
<td>-0.0721</td>
</tr>
<tr>
<td></td>
<td>(0.0018)</td>
<td>(0.0190)</td>
<td>(0.0082)</td>
<td>(0.0129)</td>
</tr>
<tr>
<td>0.8</td>
<td>-0.0733</td>
<td>-0.0907</td>
<td>-0.0881</td>
<td>-0.0904</td>
</tr>
<tr>
<td></td>
<td>(0.0026)</td>
<td>(0.0083)</td>
<td>(0.0076)</td>
<td>(0.0044)</td>
</tr>
</tbody>
</table>

The results of Table 3 show that first of all, all the mean estimates of the naïve method have poor performance. Secondly, under the low reliability, $\lambda = 0.2$ none of the techniques is performed well. However, in this table the median values of the RCAL and the IV methods are replaced instead of the mean estimates as the mean estimates are found positive. It is probably because some extreme values are generated. Next, under the moderate reliability, $\lambda = 0.5$ the RCAL and the IV nearly perform equally, whilst the others have substantial bias. When the reliability is high, $\lambda = 0.8$ the performances of the RCAL and the IV methods are close to each other. Generally, it can be said from Table 3 that as reliability increases, the bias correction methods perform better. However, all of the mean estimates are substantially biased, even under the highest reliability.

7. DISCUSSION

In this paper, we investigate if the performance of bias correction depends on the shape of data or not. Our aim was to mimic observed skewed data to obtain unbiased estimates of exposure effects in the presence of a measurement error scenario.

It can be easily understood that naively regressing the response of interest on a single exposure gives poor results of the mean estimates of the true exposure analysis in the exposure mismeasurement scenario. For low and moderate measurement error, the SIMEX performances are also poor both for the normally and lognormally distributed simulation data. In both parts of the simulation the bias of the SIMEX method is severe. Nor the naïve approach neither the SIMEX approach is recommended in the case of exposure mismeasurement when the reliability low or moderate.

As shown from our simulation study, for the normally distributed data the mean estimates of the true exposure of the RCAL and the IV methods are pretty close to each other.

They both perform well with slight bias. However, it is not the same for the lognormally distributed data despite the fact that there is a slight difference between their performances. The amount bias is considerable.

We assume that there is one mismeasured exposure which leads to attenuation of the exposure-response relationship. The data which are generated from the normal distribution gives better results than that of the lognormal distribution. Nonetheless, when the other covariates are added and investigated again, the results may change. This is an important issue for future research.
REFERENCES


Identification of Pathways from Proteomic Analysis with DAVID and PANOGA

İlknur Melis Durasi* & Osman Ugur Sezerman

**Abstract**—Proteomic analysis of protein expression can reveal the role of proteins in complex biological systems. These biological systems include the variants of complex diseases. Understanding the mechanism of multifactorial diseases is important to develop new applications for reducing the risk of them. It is known that one of the possible reasons of complex diseases is, the changes in the functionally important biological pathways. DAVID and PANOGA are two free, open-source software tools for converting given protein gi number to gene symbol and to determine functionally important pathways through the identification of genes within these pathways. Together, they enable the scientists to identify genes from a protein gi id list to be used by PANOGA for pathway analysis. Thus, they make possible to find the disease related genes that are not included as a target in the proteomics analysis but located within the affected pathway. This protocol describes in detail how to use DAVID-Gene Id Conversion tool and PANOGA to perform such analysis.

**INTRODUCTION**

With recent developments, now it is not informative enough to use isolated genes as single variants to explain the mechanism of complex diseases. That is why, we integrate the proteomics analysis to PANOGA, to map the genes to a protein-protein interaction (PPI) network and determine the connected subnetworks. Then, KEGG pathways included in the subnetworks are determined and ranked according to the significance scores. PANOGA is designed to detect functionally important pathways through the identification of genes within these pathways. Those genes are the targets of single nucleotide polymorphisms (SNPs), in other words results of GWAS analysis [2]. In our study instead of using SNP targeted genes, we modify the procedure in a way that, it becomes possible to use gi numbers to identify the disease related active pathways based on proteomics analysis.

In this protocol, starting with a list of protein gi numbers, we propose a pipeline to determine disease related pathways through the mapping of protein gi numbers to gene symbols.

It is known that, mapping gene IDs to the relevant biological annotation is very important for any type of functional analysis. There are number of different type of functional annotations from public databases and it is not very easy to maintain consistency among the annotations for the reliability of latter analysis as not only one type of annotation is used for each kind of biological analysis. That is why in our procedure, we try to come up with a practical solution to a similar kind of problem. In our case, because PANOGA uses only genes represented with their official HUGO symbols, we automatically send the protein gi numbers information to DAVID to map them to official gene symbols.

DAVID has an integrated data-mining environment to analyze gene lists obtained from the biological analysis, proteomics analysis in our case. Although there are several different tools under DAVID environment, in our procedure we use “Gene ID Conversion Tool”.

**OVERVIEW OF THE PROTOCOL**

Data files: The protocol starts with a proteomic analysis dataset which contains protein information with gi numbers and their p-values assigned indicating the significance of each in a tab-limited text file, as detailed in “input1.txt” in Figure 1. In addition to sample proteomics data file, protocol uses another type of data file which is called “humanPPI.sif”. It contains a protein-protein interaction (PPI) network in a sif file format (“input2.sif”, Figure 1), which is used by Cytoscape as a network template.

DAVID: DAVID (the Database for Annotation, Visualization and Integrated Discovery) Bioinformatics Resources [6, 7], is able to work with any type of gene list coming from any genomic platform and software.

There are several similar tools analyzing gene lists. The common feature about these publicly available tools is, they all map the genes in the gene lists to the associated biological annotation. One property that makes DAVID different from other tools is the “DAVID Knowledgebase” [1]. It is used for the collection and integration of different type of gene identifiers. It provides the users to map their gene IDs to the entire database. If a significant portion higher than 20% of input gene IDs

*Corresponding Author (e-mail: melisdurasi@sabanciuniv.edu).
İlknur Melis Durasi, Sabanci University, Biological Sciences and Bioengineering, Tuzla, Istanbul, (e-mail: melisdurasi@sabanciuniv.edu).
Osman Ugur Sezerman, Sabanci University, Biological Sciences and Bioengineering, Tuzla, Istanbul, (e-mail: ugur@sabanciuniv.edu).
cannot be mapped to a DAVID ID, then a designed module of it “DAVID Gene ID Conversion Tool” [1,6,7] is activated to help the process of mapping those unmapped IDs. This special property of DAVID is used in our protocol.

fromGItoGeneSymbol.pl: We wrote a perl script which automatically parse the input data with gi numbers and p-value information to get the gi numbers in the appropriate format DAVID uses. It sends the gi numbers to DAVID. It asks for the species and assigns the name of the species according to the user’s indication. DAVID converts the gi numbers to official gene symbols and our script takes those results and match them to the corresponding p-values.

Cytoscape: Cytoscape is an open source software platform for integrating, analyzing and visualizing the complex network data [4, 5]. Cytoscape version 2.6.3 should be installed on a computer [4, 5]. It is possible to install newer versions but the modules jActiveModules and ClueGO plugins that PANOGA protocol uses work with Cytoscape version 2.6.3.

External plugins: PANOGA uses two plugins: (i) jActiveModules plugin [4], (ii) ClueGO [3] plugin 60 of Cytoscape 68. By jActiveModules plugin, Cytoscape determines the subnetworks and by ClueGO plugin Cytoscape does functional enrichment of those subnetworks identified by jActiveModules plugin. Plugins are available for free.

PANOGA: PANOGA uses the subnetworks identified by the jActiveModules plugin of Cytoscape. By running java executable programs for the following steps of PANOGA, functionally important pathways are identified (Figure 2). Executable jar files can be downloaded from the: (please e-mail the corresponding author).

![Figure 1](image1.png)

**Figure 1** An overview of the protocol.

![Figure 2](image2.png)

**Figure 2** Outline of PANOGA protocol

**PROCEDURE**

1. **Install PANOGA**
   - Download the PANOGA files.

2. **Preprocess the data file**
   - Choose a disease name for your project according to the disease of interest (e.g. Alzheimer_disease). It is important not to put space character while assigning the disease name. It will cause some problems.
   - Create a folder under following paths below named with your chosen disease name:
     - i. PANOGA_protocol/data/
     - ii. PANOGA_protocol/out/
     - iii. PANOGA_protocol/ClueGO/data/
     - iv. PANOGA_protocol/ClueGO/out

3. **Install Cytoscape and its plugins**
   - Install Cytoscape version 2.6.3 by following its installation manual. Although Cytoscape has newer versions, for our protocol we use version 2.6.3 as it verifies plugins jActiveModules and ClueGO.
   - jActiveModules and ClueGO version 1.4 plugins should be installed to Cytoscape. Installation should be done into Cytoscape_v2.6.3/plugins/.
     - i. Copy jActiveModules plugin from: PANOGA_protocol/EXTERNAL_TOOLS/jActiveModules.jar
     - Save under: Cytoscape_v2.6.3/plugins/.
ii. Copy .cluegoplugin from:
PANOHA_protocol/ClueGO/
Save under: Cytoscape_v2.6.3/plugins/.
Do not forget to get ClueGO licence from its website
(http://wwwwww.ici.upmc.fr/cluego/cluegoLice
ncc.shtml).
Save .lf file under
home/.cluegoplugin/Licence/.l/.
Save .lcf file under
home/cluegoplugin/Licence/.lc/.

Prepare the Gene Attributes data
1) PANOGA needs the attributes file which should be in .pvals format to be used in identifying the subnetworks by jActiveModules. Gene attributes file has a P-value for each gene. In order to prepare the following formatted file (.pvals) from the raw data of proteomic analysis, run
“fromGItoGeneSymbol.pl”.
> fromGItoGeneSymbol.pl

This run creates the gene attributes file
(yourdiseasename_CytoscapeInputAttributes.pvals)
under PANOGA_procedure/data/yourdiseasename/

Obtain network data
Choose which protein-protein interaction (PPI) dataset to use as the initial network. You can use the default human PPI network downloaded within the PANOGA installation package. It is available in:
PANOHA_protocol/data/humanPPI.sif

Additionally, you can work with a different human PPI network. Cytoscape does not only accept PPI networks in .sif format, but also different file formats (e.g. .gml, .xgmml, .xls, .SBML, BioPAX, PSI-MI) can be uploaded.

Load network data
1) Start Cytoscape.
2) Load network data. We recommend users to use default human PPI network that was already downloaded with the PANOGA installation process. Import human PPI network following these commands:
File ➔ Import ➔ Network(Multiple File Types)
As long as the official HUGO gene symbols are used in the network data, users are free to upload any human PPI network.
Choose the default human PPI network which is at:
PANOHA_procedure/data/humanPPI.sif

Import gene attributes
1) By this step values to genes (nodes) are assigned. The values for the genes are stored in the gene attributes file (yourdiseasename_CytoscapeInputAttributes.pvals). To assign the values to nodes use following commands:
File ➔ Import ➔ Attributes from Table
Choose the gene attributes file which is at:
PANOHA_procedure/data/yourdiseasename/.

Identify subnetworks
1) Start jActiveModules plugin from:
Cytoscape ➔ Plugins ➔ jActiveModules
2) In the Expression Attributes for Analysis panel, select “column 2” which contains the values of the nodes.
3) In the General Parameters panel, set “Number of Modules” parameter to 1000, “Overlap Threshold” parameter which defines the maximum percent of overlap between any two identified subnetworks, to 0.5.
4) Click “Find Modules” and identification of active subnetworks starts.
5) Save the result as text file into:
PANOHA_procedure/data/yourdiseasename/yourdiseasename_jactivemodules_output.txt

Parse jActiveModules output
1) Create a text file containing the gene names that is included in Gene attributes file under:
PANOHA_protocol/data/yourdiseasename
2) Run “parsejactivemodulesoutput.jar”.
>java –jar parsejactivemodulesoutput.jar yourdiseasename
time

This command will create files with gene symbols for each subnetwork with scores higher than 3. These files are saved under:
PANOHA_protocol/ClueGO/data/yourdiseasename
The number of subnetworks (numofsubnetworks) identified is printed on the screen. It will be used in the next steps of PANOGA.

Functional Enrichment of subnetworks
1) Identify KEGG pathways. At first open the clueGO.props file which is under:
PANOHA_protocol/data/yourdiseasename
To use KEGG pathways, in the clueGO.props file look for
SelectedOntologySources=KEGG_14.03.2012” under the “Select Ontologies”.
2) For performing functional enrichment for each of the identified subnetworks run the following command:
Combine functional enrichment results
1) Run “combinesubnetworkpathways.jar”.
   \texttt{java -jar combinesubnetworkpathways.jar diseaseasename numofsubnetworks}
   As a result of this step KEGG pathways in the subnetworks with their significance scores and gene lists are created. The results file are saved under:
   PANOGA_protocol\(\rightarrow\)out\(\rightarrow\)yourdiseasename\(\rightarrow\)your diseaseasename_subnetwork_genes_pathways.txt

TROUBLESHOOTING
Out-of-memory errors
During the subnetwork identification step of PANOGA which is performed by jActiveModules plugin of Cytoscape, there might be memory problems for big networks containing more than 10,000 nodes. In such cases, Cytoscape displays an error message. You can increase the memory for the Cytoscape:
- open Cytoscape.vmoptions file (under .cytoscape folder)
- Change the following lines
  -Xmx20G
  -Xss30M

or you can delete the unnecessary networks.

For further information about increasing the memory space:
http://www.cytoscape.org/cgi-bin/moin.cgi/How_to_increase_memory_for_Cytoscape.

ANTICIPATED RESULTS
PANOGA creates pathway and gene tables as shown in Table 1 and Table 2.
In Table 1, each row shows the identified pathway and its features (Table 1).
In Table 2, each row shows each gene’s participation in the subnetworks.

| TABLE 1 | AN EXAMPLE FOR PATHWAY-BASED REPRESENTATION OF PANOGA RESULTS, FOCUSING ON SUBNETWORK GENES |
|-------------------|-------------------------------------------------|-------------------|-------------------|
| KEGG ID | KEGG Term | Rank | Term Pvalue Corrected | Pathway Associated Genes Found in Subnetworks | Pathway Associated Genes Not Found in Subnetworks |
|-------------------|-------------------------------------------------|-------------------|-------------------|
| KEGG04 512 | ECM-receptor interaction | 1 | 8.76E-21 | COLA4A2; COLA4A1; ... | HMMR; HSPG2; IBDP; ... |
| KEGG04 630 | Jak-STAT signaling pathway | 2 | 1.01E-19 | IL6*; IL2RB; OSMR; ... | CREBBP; CRLF2; CSF2; ... |
| KEGG04 610 | Complement and coagulation cascades | 3 | 2.42E-19 | PLAT; KNG1; F11; ... | AZM; BDKRB1; BDKRB2; ... |

<p>| TABLE 2 | AN EXAMPLE FOR SUBNETWORK GENES BASED REPRESENTATION OF PANOGA RESULTS |
|-------------------|-------------------|-------------------|-------------------|</p>
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Times Found in Subnetwork</th>
<th># of Associated Pathways</th>
<th>Associated Pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAF6</td>
<td>1</td>
<td>2</td>
<td>[Leishmaniasis, Toxoplasmosis]</td>
</tr>
<tr>
<td>MCM7</td>
<td>56</td>
<td>1</td>
<td>[DNA replication]</td>
</tr>
<tr>
<td>POLA1</td>
<td>2</td>
<td>1</td>
<td>[DNA replication]</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>1</td>
<td>2</td>
<td>[Prostate cancer, Colorectal cancer]</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENT
We are grateful to Burcu Bakir Güngör for being helpful in technical issues related to PANOGA.

REFERENCES
Abstract We aim to study effect of neurodegeneration on cerebral cortex of rat pups caused by prenatal and postnatal ethanol exposure with modified liquid diet, to illuminate protective effects of betaine and omega-3 supplementation. When ethanol is consumed during prenatal and postnatal periods, it may result in Fetal Alcohol Syndrome (FAS) in the offspring. Rats were divided into control, ethanol, ethanol+betaine, ethanol+omega-3, ethanol+omega-3+betaine groups. The effect of betaine and omega-3 in response to ethanol-induced changes on cerebral cortex, by biochemical analyses cytochrome c, caspase-3, cathepsin B, cathepsin L levels and histological methods were evaluated.

Caspase-3, cathepsin B, cytochrome c levels in ethanol group were significantly higher than control group. Caspase-3 level was decreased in ethanol+betaine, ethanol+omega-3, ethanol+omega-3+betaine groups compared to ethanol group. Also cathepsin B in ethanol+betaine group was found to be significantly higher than control. Cathepsin B in ethanol+omega-3+betaine group was decreased compared to ethanol, ethanol+betaine groups. Cathepsin L was found not statistically significant. We found similar results in histological parameters. As a result of histological examination of tissue damage, hemorrhage, PMNL, microglia in control, ethanol+betaine, ethanol+omega-3, ethanol+omega-3+betaine groups were decreased compared to ethanol group. Congestion, necrosis ethanol+omega-3, control groups were decreased compared to ethanol group.

In conclusion, we found that pre and postnatal alcohol-exposure-related neurodegeneration of newborns, the protective effect of betaine in these regions and the benefits of omega-3 supplementation that is required for normal brain development was investigated.

INTRODUCTION

Alcohol is known to impede the growth of the central nervous system and to induce neurodegeneration. Alcohol exposure can have severe detrimental effects on brain function [1].

It is well known that alcohol abuse during pregnancy induces a wide spectrum of major and minor anomalies, which in their extreme manifestation constitutes the fetal alcohol syndrome (FAS). Exposure to alcohol during pregnancy may lead to birth defects such as microcephalia, anencephalia and myeloshyzis. Fetal alcohol spectrum disorders are composed of 4 diagnostic categories: (1)fetal alcohol syndrome; (2) partial fetal alcohol syndrome; (3)alcohol-related neurodevelopmental disorders; and (4)alcohol-related birth defects. Recent school-based prevalence studies using active case ascertainment strategies followed by comprehensive multidisciplinary evaluations have identified fetal alcohol spectrum disorder prevalence rate of 1% to 2% [2].

In this study, in rat cerebral cortex, prenatal and postnatal alcohol-exposure-related neurodegeneration of newborns, the protective effect of betaine in these regions and the benefits of omega-3 supplementation that is required for normal brain development was investigated.

MATERIALS AND METHODS

Animals and Laboratory Conditions: A large number of adult male and female Spraque-Dawley rats of the same age, which were 150-250 grams at the beginning of the experiments, obtained from TICAM (Medical and Surgical Experimental Research Centre, Eskisehir) and were used for breeding in the present study. All animal procedures were approved by the Animal Care Committee. Female rats were individually housed in transparent plexiglas cages. For mating, a male rat, picked at random, was placed into a female's cage. Rats were allowed to mate nightly until a vaginal plug was observed on the following morning. The females were checked each morning for the presence of a vaginal plug. The presence of a vaginal plug was evidence of a successful fertilization and the day that a positive plug was defined as E0. Embryonic (prenatal,gestational) (E = day of conception; E0 = the day a vagina plug is found in mated females) and postnatal (lactation) (P = days postnatal; P0 = the first 24 h after birth) offspring were produced from timed pregnancies. Usually the control groups gave birth after E20-22, but the ethanol treatment delayed birth by 1–5 days to E21 or E25. The animal model of prenatal...
(gestational=embryonic) and postnatal (lactation) ethanol exposure was made by modifying Ertem et al. method [3]. Rats were divided into 5 different groups; control, ethanol, and treatment groups: ethanol+omega-3, ethanol+betaine, ethanol+omega-3+ betaine. A total of 65 pups (n=13 for each group, biochemical experiment n=8, histology n=5) at P 21 (postnatal day 21) were used for the study. At the beginning of the study, rats were given the MLD without ethanol for 3 days (E0-E3). Omega-3 and/or betaine administered before ethanol administration (E3-E5). The omega-3 and betaine supplementanation groups had additional doses of 30 mg /100g/day and 1% w/v betaine (E3-P21), respectively [4,5]. Then, liquid diet with 2.4% ethanol was administered for 3 days (E6-E8). The ethanol concentration was increased to 4.8% for the following 3 days (E9-E11) and finally to 7.2% for another days (E11-P21).

Surgical Procedure: Pups were killed by decapitation on day P21. Their cerebral cortex were removed surgically. Tissues used for biochemical studies were frozen in liquid nitrogen and kept at -80°C until they were tested.

Cytochrome c Measurement: For the measurement of cytochrome c release, mitochondrial and cytosolic fractions were separated by the method of Soeda et al [6]. Cytochrome c levels were detected both in cytosolic and mitochondrial fractions with a commercial kit (Rat / Mouse Cytochrome c Immunoassay, R & D Systems MCTCO).

Caspase-3 Enzyme Activity: Caspase-3 activity was determined according to the method of Zovein et al [7]. The method is based on the measurement of pNA (p-nitroanilin) absorbance at 405 nm formed by hydrolysis of caspase-3 with commercially available kits (Caspase Assay Kit Colorometric 3, Sigma). Caspase-3 enzyme activity was calculated to receive µmol of pNA released per min per mg of protein.

Cathepsin B and L Activities: Cytosolic and lysosomal fractions were obtained from cerebral cortex homogenate [8]. Cathepsin activity was expressed as units/mg protein. The ratio of the cathepsin activities measured separately from cytosolic and lysosomal fractions showed the amount of lysosomal integrity.

Protein Assay: The protein concentration of homogenates gathered from brain tissues were determined using the Bradford procedure [9].

Histological Analysis: All of the specimens were fixed in neutral buffered formalin, the tissues were embedded in paraffin and serial sections of 4 um were prepared for each of the paraffin blocks. On average, 50 sections were collected per rat and sections were stained for histological analyses with H&E.

RESULTS

Figure 1 graphically illustrates caspase-3, cytochrome c, cathepsin B and L. When the caspase-3 activities (umol pNA/minute/mg protein) were assessed statistically, it was seen that the caspase-3 activity of the ethanol group (0.81±0.13) were statistically significant (p<0.001) compared to control group (0.41±0.19). The results for the ethanol+omega-3 group (0.43±0.21) and ethanol+betaine group (0.49±0.04) and ethanol+betaine+omega-3 group (0.49±0.17) were significant compared to the ethanol group (p>0.001). The change of cytochrome c release (cytosolic/mitochondrional cytochrome c) with respect to ethanol group (0.17 ± 0.09) was significantly higher than the control group (0.08 ± 0.04) (p<0.05). Cytochrome c release level for ethanol+omega-3 group (0.11 ±0.03), ethanol+betaine group (0.14±0.05) and ethanol+betaine+omega-3 group (0.10 ± 0.03) were slightly lower than ethanol group but the results were not statistically significant (p>0.05). The cathepsin B levels (cytosolic/lysosomal cathepsin B activity) of ethanol (2.41± 0.86) and ethanol+ betaine groups (2.13 ± 1.04) were higher than control (0.72±0.40) (p<0.001). The cathepsin B level ethanol+betaine+omega-3 group (0.72±0.33) was higher than ethanol and ethanol+betaine groups (p<0.001, p<0.001 respectively). When the cathepsin L levels (cytosolic/lysosomal cathepsin L activity) were assessed statistically, there were not a significant difference between control (0.68±0.12), ethanol (0.68±0.17), ethanol+omega-3 (0.72±0.14), ethanol+betaine (0.58±0.09) and ethanol+betaine+omega-3 groups (0.7±0.13) (p> 0.05).

Figure 1: Distribution of caspase-3, cytochrome c, cathepsin B and L activities in rat pup groups.

a: p<0.001 different from control group, b: p<0.001 different from ethanol group, c: p<0.001 different from betaine group, d: p<0.05 different from control group
When H & E stained brain cortex was examined, the normal tissue appearance was observed in the control group (Figure 2a). In the ethanol group, there was congestion, hemorrhage and necrosis and microglia/macrophage/MNL infiltration (Figure 2b). Histological examination of tissue damage, hemorrhage, PMNL, microglia in ethanol+betaine, ethanol+omega-3, ethanol+omega-3+betaine groups were decreased compared to ethanol group. In addition congestion, necrosis ethanol+omega-3 and control groups were decreased compared to ethanol group.

**DISCUSSION**

Alcohol consumption during pregnancy can cause neurodegeneration [10]. Current evidence indicates chemicals often trigger different cell death in a dose-dependent manner, including autophagy or apoptosis after lower-dose and necrosis at higher-dose exposure [14].

In this study, it was shown that cytochrome c secretion, caspase-3 activity, cathepsin b levels increased on cerebral cortex of rat pups from female rats that had administered ethanol in pre and postnatal periods. We found that caspase-3, cathepsin B, cytochrome c levels in ethanol group were significantly higher than control. Cathepsin B in ethanol+betaine group was found to be significantly higher than control. Cathepsin B in ethanol+omega-3+betaine group was decreased compared to ethanol, ethanol+betaine groups. Cathepsin L was found not statistically significant. Mitochondrial damage was observed on the cell death in ethanol-induced organ toxicities [10,11]. Kim et al. showed that betaine supplementation is protective against necrotic damage [12]. When there is a reduced level of docosahexaenoic acid (DHA) in the brain, many dramatic changes in brain function have been reported including changes in size of neurons, changes in learning and memory. DHA plays an important role in the structure and function of brain cellular membranes [15]. Similar to earlier studies [13], in this study it was histologically observed that ethanol caused significant damage on brain cortex tissue. Especially on neurons, the damage was evident in terms of congestion and necrosis. It is a remarkable finding that the necrosis caused by ethanol group was significantly reduced by ethanol+omega-3 administration.

As a result, it can be concluded that betaine and omega-3 both alone or together may inhibit neurodegeneration on offspring with fetal alcohol syndrome from the rats that were administered alcohol during pre and postnatal period. This study on brain tissue of newborn rat pups make the way for reversing the damage on children with FAS with betaine and omega-3 supplementation.

**REFERENCES**


Descriptive consideration of serum irisin levels various factors: obesity, type 2 diabetes mellitus, pre-diabetic status, gender, and athletics

Basil O. Saleh, Maysaa J. Majeed* and Ghassan M. Oreaby

Abstract—Recently obesity, has become the main risk for a number of conditions and metabolic syndrome including type 2 diabetes (T2DM). The aims of this study were to assess whether irisin can bring out the missing circle between increasing physical activity and weight reduction. 108 subjects were chosen to matched with 76 diabetic patients in age and BMI, both groups were divided into subgroups according to their BMI. Twenty two healthy athlete men were also involved in the study. The mean±SD value of serum irisin concentrations of obese healthy individuals was the highest concentration (p= 0.015) in comparison with normal and overweight healthy individuals. The more recent observation which describe irisin as adipokine factor (willing myokine source), irisin levels were examine pursuant to gender, obese women group have had the highest level of serum irisin than other women studied groups (p=0.001), while in healthy men serum irisin did not differ significantly. Careful examination were searched for irisin as a myokine after adjusting age and BMI for gender depending on that women had significantly lower serum irisin (p=0.006) inspite of significantly higher fat percentage (p=0.001) when compared with men. significantly lower serum irisin concentration in healthy men with normal physical activity (p=0.01) than healthy athletes men pointed that irisin as myokine potentionally. Diabetic patients showed significant lower serum irisin level than healthy subjects

Introduction

The rate of obesity is a peril evolution worldwide ‘lifestyle environmental factors and low physical activities may play a dominant role in obesity development. Central obesity, the main risk factor for a number of conditions and metabolic syndrome including type 2 diabetes mellitus (T2DM) [1]. Chronic exercise training induces adaptive structural and metabolic changes in skeletal muscle including a change in the type of muscle fiber, mitochondrial biogenesis, and angiogenesis. Forceful progress is developing the molecular regulatory circuitry in the exercise-induced changes in muscle structure and function, the expression of a gene transcriptional co-regulator called peroxisome proliferator-activated receptor γ (PPARγ), co-activator-1α (PGC-1α) is induced in muscle in response to exercise in rodents and humans. PGC-1α over expression in muscle, cause changes in whole body fat tissue depots. Profiling of muscle genes activated by PGC-1α identified a factor called fibronectin type III domain containing 5 (FNDC5) with predicted structural features of type I membrane protein that could be proteolytically cleaved to release a smaller protein into the bloodstream, a secreted protein product of FNDC5, named irisin was found [2].

Further Information

Study designed: This human study model was carried out at Biochemistry Department, College of Medicine, University of Baghdad, Al-Yarmok Teaching Hospital, Obesity Center at Al-Kindy Medical College, and Athletic College, University of Baghdad, Iraq, during the period from January 2013 to July 2013.

Investigations included serum measurements of irisin, insulin and leptin using ELISA.

Patients: One hundred T2DM patients were encountered, but only seventy six were included in the current study to be age and body mass index (BMI) matched with healthy individuals group. Thirty-four of these patients were overweight (15 women and 19 men), their BMI were > 25 -29.9 kg/m², with age range of 35 -67 year, forty-two were obese (21 women and 21 men), their BMI > 30 kg/m², and their age range was 38-62 year. The criteria for diagnosis and inclusion of these patients with type 2 DM was depended on careful history and/or the measurements of serum glucose and HbA1c.

Circulating irisin level was significantly higher in healthy obese subjects (49.58 ±7.55 ng/ml) compared with healthy normal weight subjects (42.8 ±14.3 ng/ml, P=0.009) and healthy overweight subjects (45.97 ±6.02 ng/ml, P=0.03). (Table 1)

ANOVA and t-test revealed, NS no significant differences in fasting glucose and insulin among the groups, ▲; t-test showed significant difference between the obese and normal weight groups (P=0.001), ▲▲; t-test revealed significant difference between overweight and normal weight groups (P=0.009), ●; t-test showed significant difference between obese and normal weight groups (P=0.001).

*Corresponding Author Maysaa J. Majeed Biochemistry Department, College of Medicine, University of Baghdad, Iraq. Email. msysjalal@yahoo.com
Despite that the mean of fat percentage value of women (39.39 ± 8.55 %, P=0.001) was significantly higher than that of men (28.72 ± 7.59 %), their serum irisin levels was significantly decreased (43.22 ± 10.11 ng/ml vs. 48.86 ± 10.61 ng/ml, respectively; p=0.006), while other related metabolic factors like serum insulin and leptin showed non-significant differences between men and women as shown in Table (2).

The above finding of the present study confirmed those found by Stengel et al. [3], Huh et al.[4] whose results indicated that morbidly healthy obese individual had higher circulating levels of irisin than normal weight, their conclusion depend on the fact that decrease in irisin levels after weight loss was explained by the decrease in muscle mass in normal healthy subjects suggesting that the positive association of irisin with BMI is at least in part explained by muscle mass. In the authors opinion, this could indicate that increased in circulating irisin is an adaptive response needed to compensate for the decreasing insulin sensitivity and disturbances in metabolize associated obesity.

The data presented in Table 2 may answer the above suggestion related to gender-irisin association which revealed that women serum irisin level was affected by the changes in obesity marker, the BMI values with higher serum irisin was higher in obese and overweight healthy women than in normal weight healthy ones. These discrepancies could reflect on gender-differences, and also suggest a dissociation between FNDC5 expression and circulating irisin under some condition[2].

CONCLUSIONS

Irisin myokine in facing irisin adipokine is powerful; that women had lower Irisin level even with higher fat percentage, in addition athletics healthy men had higher serum Irisin than the same gender with normal activity.

ACKNOWLEDGMENT

The authors thank all staff of obesity Center, Al-Kandy Medical College, Baghdad University, Teaching Laboratories, Baghdad Teaching Hospital, Diabetes Center.

REFERENCES

Abstract—Building feature vectors is of key importance for deciphering the patterns to the classifier algorithms for machine learning problems. Here, we present the YUFES web server that provides encoding feature vectors for the proteins given with respect to several feature encoding techniques within a few seconds with a user-friendly interface. The web server, which is designed to cater to a range of bioinformatics user expertise, fills the gap building feature vectors as online. The YUFES web server supports seven feature encoding techniques; some uses parameters, such as BLOSUM and PAM substitution matrices, calculated individually. The YUFES web server is freely available at http://yufes.yalova.edu.tr.

Keywords: Protein web server, Feature encoding, Peptide classification, Feature extraction

INTRODUCTION

Increasingly, web applications provide new insights to deploy software applications. It is obvious that conveying desktop applications to the web platforms is the trend segments of the software market. Web applications are built with a number of new languages, technologies, and programming models, and are used to implement highly interactive applications that have very high quality requirements [1]. In the field of bioinformatics, we present the first web based tool termed YUFES that builds feature vectors for machine learning classification and regression problems.

Today, as a consequence of having identification of the encoded proteome, understanding the function of proteins at the cellular level has become one of the most important objectives in biological sciences [2]. To recognize the role of the residues in a protein sequences can lead to remarkable inferences and predictions in the field of molecular biological problems. In this assay, researchers aim to decipher the patterns of peptides with the machine learning techniques. Feature encoding that is building feature vectors from peptide samples is considered to be of paramount importance for an effective recognition of peptide patterns.

The aim of feature encoding, which is a commonly used technique applied before classification, is defining a mapping from the original representation space into a new space where the classes are more easily separable. This will reduce the classifier complexity, increasing in most cases classifier accuracy [3]. To identify the as much as patterns to the classifier algorithms, amino acid sequences and their physicochemical, functional and evolutionary attributes can be used in the structure of feature vectors. YUFES web server implements and monitors the several feature encoding methods according to given protein sequences.

This paper is organized as follows. In “Methods”, feature encoding methods implemented by YUFES are explained. In “Developed Web Server”, developed encoding web server is explained. Finally, “Conclusion” concludes the paper.

METHODS

19. Feature Encoding Techniques

YUFES web server implements seven feature encoding techniques for protein sequences: Orthonormal encoding (OE), OE with BLOSUM and PAM substitution, Frequency-based Encoding (FE), composition moment vector (CMV), BLOMAP and OETMAP encoding techniques.

The first implemented technique of YUFES is Orthonormal encoding (OE) which is a common encoding method. According to OE, each amino acid symbol $p_i$ in a peptide is replaced by an orthonormal vector, $d_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{i20})$ where $\delta_{ij}$ is the Kronecker delta symbol. Then, each $P_i$ is represented by a 20-bit vector, 19 bits are set to zero and 1 bit is set to one based on alphabetic order of amino acids. Each $d_i$ vector is orthogonal to the rest of $d_j$ vectors and $P_i$ can be any one of the twenty amino acids available in human body [4]. Each octapeptide thereby is represented by a vector of 160 bits.

The second and third feature encoding techniques are combining of OE with BLOSUM and PAM substitution matrices, respectively. Amino acids of homologous sequences which are frequently substituted by each other over time are regarded as similar and the relationships are portrayed by substitution matrices, like the BLOSUM [5] and the PAM matrices [6]. Both BLOSUM and PAM matrices are used combining OE method where vector $d_i$ is multiplied by the diagonal entries of related matrix (i.e. row $i$ and column $j$) by [7].

---

*Corresponding Author (e-mail: murat.gok@yalova.edu.tr).
Author Two is both with The Conference and The Bioinformatics Course, Kusadasi, Turkey, (e-mail: murat.gok@yalova.edu.tr).
The fourth feature encoding method implemented by YUFES is the Frequency-based Encoding (FE) scheme. In this method, weight of each $P_i$ in a peptide is determined and then combined by OE. In this way, vector $d_i$ is multiplied by the weight of $P_i$ [8].

The fifth implemented technique is composition moment vector (CMV) encoding. CMV uses an approach to obtain information from primary sequence based on a composition moment vector, which is a measure that includes information about both composition of a given primary sequence and the position of amino acids in the sequence [9].

The sixth YUFES implementation is BLOMAP encoding scheme which utilizes a non-linear projection method to recognize the similarity information in the BLOSUM62 matrix [10].

The last technique implemented by YUFES is OETMAP which consists of a conjunction of OE and Taylor’s venn diagram [11].

**DEVELOPED WEB SERVER**

In order to build feature vectors, the YUFES accepts protein sequences either entering directly the primary structure into a textbox or uploading the FASTA format, which is a text-based format for representing protein sequences, to the server at once. Views from a typical YUFES screen shots are shown in Figure 1. The top section of Figure 1 contains input protein sequences into a text box via keyboard.

Protein sequences can be encoded according to selected method shown in the top section of Figure 2 by clicking “Encode!” button. This action returns encoded protein sequences to the user as a CSV file. Built feature vector can be downloaded from the link given below the system message menu shown in Figure 3.

**CONCLUSIONS**

YUFES is the first web based tool which builds feature vectors according to selected encoding techniques. The focus of YUFES web server has been on speed and minimizing response time. Most vector building implementations take 4–5 s against oligo-protein sequences in desktop applications. Our aim is to drive implementation times down even further, such that typical implementation times are in the 1-2 s range. We also plan to add new feature encoding techniques that combine physicochemical properties of amino acids, in
the near future. Also our long-term aim is to conduct peptide classification with YUFES according to given sample peptide sequences and selected learning algorithms such as k-NN, Bayes to the YUFES.

ACKNOWLEDGMENT

This work was supported by Yalova University projects (Grant 2014 / YL / 037) and (Grant 2012 / 046).

REFERENCES

Abstract— In this paper, a software program named GeneSeq, that will help diagnosing errors has been generated by DNA analysis device and computing a pairwise global alignment of the raw forward and reverse sequences. It is available on the Web page http://www.dnanlyse.com.

INTRODUCTION

The reason why this study has been made is that ABI 310 DNA analysis machine used in Ege University, Faculty of Medicine, Department of Pediatrics fails to generate evenly-spaced peaks. There are many reasons for this machine to produce incorrect results such as messy samples or dirty parts of machine. Thus, the chromatograms (generally known as trace files) can often be ambiguous and many programs can predict incorrect results. As this causes the miss-assembly of contigs, manual intervention is required to correct the results. In order to avoid manual intervention, we have developed GeneSeq DNA sequence analyzer program.

There are many trace file formats such as SCF [1], ZTR, ABI. These files contain the trace amplitudes, the base calls, their confidence values, and textual data about the particular sequencing experiment such as its chemistry machine type and operating conditions, etc. An extensive list of the common textual data associated with individual trace files is given at National Center for Biotechnology Information trace repository (http://www.ncbi.nlm.nih.gov/Traces).

Through the use of publicly available library of NET Bio (http://www.biocodeplex.com), GeneSeq can read the files of these formats and display their contents. The program can be used to locate and mark the extent of low quality data at each end of the sequence and the positions of vector sequences. It also enables the user to edit the base calls, though any edits made are normally saved to the reading’s experiment file [2].

A typical display of GeneSeq is shown in Figure 1. It includes several functions such as limiting the confidence values, importing DNA sequences from excel files, opening trace file formats, saving edited base calls, reversing sequencing reads, converting sequences Fasta to Raw and aligning the forward and reverse sequences.

MATERIALS AND METHOD

Accurate base-call quality scores are essential for producing finished sequences. Calculating accurate quality scores is a complex problem, due in part to variations in sequencing machines and techniques [9]. However, all modern capillary sequencing instruments use base-calling software that calculates quality scores, so virtually all recent chromatogram trace files include them.

Consequently, trace files processed with GeneSeq are expected to include quality scores, and GeneSeq represents bases which have low quality by Ns depending on parameter of quality level determined by user. Moreover, GeneSeq can trim regions of low quality from the beginning and the end of the sequences. Because low quality means that the observed read sequence is possibly wrong, and the wrong sequence may lead to a wrong alignment. The user can also directly edit the sequence from within the viewing window. Editing includes unlimited undo and redo operations functionality.

The program runs on PCs using the Windows operating system and has several useful functions including a base-calling feature. GeneSeq can compare the trace patterns of four bases automatically and reports the results of comparison, therefore users check the base-calling result. Finally, users can decide whether reads of machine is confident or not.
Base-calling errors are the cause of several problems in contig assembly. GeneSeq doesn’t need other base-calling programs to validate data. GeneSeq has a connection to NCBI web site, so users can use search tool in the program to get data from NCBI without using any web browser. The parameter of limit of quality value can be changed by user and sequencing can be reverse using reverse button. In addition to these features, GeneSeq reports the results of analysis. Alternatively, a NCBI nucleotide database BLAST search may be initiated with the edited sequence directly in GeneSeq. Sequences from the NCBI nucleotide database can be imported directly into software.

Researchers need to generate a consensus sequence from matching forward and reverse sequencing reads. Consequently, it is required to compute a pairwise global alignment of the raw forward and reverse sequences. Geneseq can be used to align the sequences.

CONCLUSIONS

In the literature, we can see many chromatograms trace editors like Trev [10], ChroView [11], TraceEdit[12]. SeqTrace offers several advantages in comparison with these earlier programs. First, GeneSeq is free and open-source software. It is available on the Web page http://www.dnanlyse.com. Second, GeneSeq is a self-contained application that does not require the installation of any additional bioinformatics software packages. Finally, GeneSeq is a special software so that the program is designed to contribute for researchers in dna analysing.

For researchers that need to derive finished sequences from many chromatogram files. GeneSeq can result in considerable time savings through Needleman–Wunsch alignment algorithm in comparison with most other free programs. These capabilities, along with a user-friendly and intuitive interface, make GeneSeq an efficient tool for generating high-quality DNA sequences from chromatogram trace files.

This software has been used for investigating genetic diseases, which is led by Prof. Dr. Afig Berdeli in Ege University, Faculty of Medicine, Department of Pediatrics and it has provided good results.

ACKNOWLEDGMENT

I would like to express my very great appreciation to Prof. Dr. Afig Berdeli for his valuable and constructive suggestions during the planning and development of this research work.

REFERENCES

Abstract—It is well-known that diffuse large B-cell lymphoma (DLBCL) is the most common form of human lymphoma. Clinical presentation, prognosis, and response to current therapies vary among patients with DLBCL. The standard treatment of advanced DLBCL is the combination of chemotherapy and immunotherapy. The most common chemotherapy regimen for advanced DLBCL is called R-CHOP, which includes rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone. We studied 28 patients with DLBCL who were treated with the R-CHOP regimen. We analyzed UGT1A1 and CYP3A4 gene polymorphisms and their association with clinical and outcome variables. DNA was isolated from formalin-fixed paraffin-embedded tumor lymphoid tissue blocks. UGT1A1*28 and CCYP3A4*1B polymorphisms were determined by pyrosequencing method. According to our results, these polymorphisms do not affect the efficacy of R-CHOP.

INTRODUCTION

DLBCL, representing 25% to 35% of new cases annually, is the most common Non-Hodgkin Lymphoma histology. This disorder is composed of a clinically and pathologically heterogeneous group of lymphoproliferative malignancies. Patients with DLBCL show variations in clinical presentation, prognosis, and response to current therapies. DLBCL, unlike indolent lymphomas, is an aggressive lymphoma that needs treatment. If it is not treated, survival may be measured in weeks to months. Patients often have a rapidly enlarging mass in a lymphatic region. Although uncommon, the presence of extranodal involvement or associated constitutional symptoms indicates a more aggressive phenotype. At diagnosis, only 20% to 30% of patients have evidence of DLBCL in the marrow [1].

Although response rates to standard therapy with R-CHOP range from 80% to 90%, overall survival rates range from 30% to 50% over 5 years for all patients with DLBCL [2]. This indicates a clinical spectrum of sensitivity to the standard treatment of R-CHOP.

Variations exist in the efficacy and toxicity associated with R-CHOP. Pharmacogenetic factors influence the variation in toxicity and response to R-CHOP. Better understanding of these factors gives the ability to individualize treatment.

We report here our results on a 28 DLBCL patients. Polymorphic variations of several drug metabolism enzymes were assessed for potential link with DLBCL patient prognosis. Recent data in the literature show that polymorphisms of both UGT1A1 and CYP3A4 enzymes might play a role in hematologic malignancies. Accordingly, the choice of enzyme gene polymorphisms was guided by the literature [3,4].

Further Information

The present study consisted of 28 patients with a median age of 60 years, ranging from 19 to 84 years. All patients were treated with R-CHOP. DNA was isolated from formalin-fixed paraffin-embedded tumor lymphoid tissue blocks. UGT1A1 promoter region contains between 5 and 8 TA nucleotide repeats in TATA region. UGT1A1*28 means that 7-TA repeat were present in the promoter region. Of the 28 patients studied for UGT1A1 polymorphism, 10 had heterozygote 7-TA repeat (UGT1A1*28 allele), 6 homozygous for UGT1A1*28 allele. And 12 patients carried wild type TA in promoter region (Table 1.).

CYP3A4 is one of the most important and most abundant CYP isozymes in humans and it is involved in the metabolism of over 50% of all drugs. Of the 28 patients studied for CYP3A4*1B polymorphism, 25 patient carried wild type and 3 had heterozygous allele (Table 1.).

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>ALLELE FREQUENCY OF CYP3A4<em>1B AND UGT1A1</em>28 GENE POLYMORPHISMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP3A4*1B</td>
<td>UGT1A1*28</td>
</tr>
<tr>
<td>WT</td>
<td>25</td>
</tr>
<tr>
<td>HT</td>
<td>3</td>
</tr>
<tr>
<td>MT</td>
<td>0</td>
</tr>
</tbody>
</table>

where WT, wild type; HT, heterozygote; MT, homozygote.
Figure 1: Homozygous mutant sample for UGT1A1*28 polymorphism.

CONCLUSIONS

Results of this study indicate that there is no relationship between allele and therapy with R-CHOP. However, data in this study were limited. With larger data, it may be possible to establish such a relationship.

REFERENCES


A Comparative Study on Diabetes Disease Diagnosis Using Data Mining Classification Techniques

Zeynep CEYLAN*, Zehra KARHAN, S. Ümit OKTAY FIRAT, Burhan ERGEN

Abstract—The health care sector is one of the world’s largest and fast growing area. This causes increase in amount of raw data stored in databases. Data mining is an important tool to extract hidden and useful knowledge from raw data. Therefore, application of data mining techniques on health sector is crucial. The aim of this study is to do the performance analysis on diabetes disease, which is one of the most important disease in medical field, using K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) classification techniques. Suitable technique was chosen based on highest classification accuracy measurement. As a result, the experimental results show that SVM classifier is more successfully than KNN for diagnosing diabetes disease.

INTRODUCTION

The healthcare industry is one of the world’s largest and fast growing area [1]. Because of increase in electronic health records, institutions are accumulating big amounts of patient data. However, the generated data by healthcare transactions are too complex and voluminous to be processed and cannot be analyzed by traditional methods. Therefore, data mining has become increasingly essential to transform these mounds of data into useful information for decision making in healthcare [2].

One of the important challenges in healthcare area is the discovering of useful knowledge from medical diagnosis [3]. Analyzing the parameters for diagnosing is difficult and time consuming for physician. Therefore, machine learning and data mining techniques have been developed to design automatic diagnosis systems for diabetes. Thus, the use of classification techniques in medical diagnosis is increasing gradually [4-5].

In this paper, a comparative study for Pima Diabetes Disease diagnosis was carried out. Because, diabetes is a serious health problem in both industrial and developing countries and it is rising dramatically [6-7]. Therefore, diabetes disease diagnosis through clear explanation of the diabetes data is a significant classification problem [8]. For this purpose, K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) classifiers were used, in this study. The results of the study were compared on same UCI machine learning database.

The rest of this study is organized as follows. Section 2 describes our dataset and gives a short description of both KNN and SVM classifiers. Section 3 describes our experimental results. Section 4 is a discussion and conclusion.

MATERIAL & METHODOLOGY

In this study, the dataset was obtained from the UCI Repository of Machine Learning Databases [9]. The dataset consists of patients who are women at least 21 years old Pima Indian heritage. It includes 768 samples, where each sample has 9 features which are detailed in Table I.

### TABLE I. DATASET DESCRIPTION [9]

<table>
<thead>
<tr>
<th>Data Description</th>
<th>Attribute Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of times pregnant</td>
<td>Numeric</td>
</tr>
<tr>
<td>2</td>
<td>Plasma glucose concentration a 2 hours in an oral glucose tolerance test</td>
<td>Numeric</td>
</tr>
<tr>
<td>3</td>
<td>Diastolic blood pressure (mm Hg)</td>
<td>Numeric</td>
</tr>
<tr>
<td>4</td>
<td>Triceps skin fold thickness (mm)</td>
<td>Numeric</td>
</tr>
<tr>
<td>5</td>
<td>2-Hour serum insulin (mu U/ml)</td>
<td>Numeric</td>
</tr>
<tr>
<td>6</td>
<td>Body mass index (weight in kg/height in m²)</td>
<td>Numeric</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes pedigree function</td>
<td>Numeric</td>
</tr>
<tr>
<td>8</td>
<td>Age (years)</td>
<td>Numeric</td>
</tr>
<tr>
<td>9</td>
<td>Class variable (0 or 1)</td>
<td>Nominal (Binary)</td>
</tr>
</tbody>
</table>

The class value ‘1’ means a positive test for diabetes (patient), and ‘0’ is a negative test for diabetes (healthy). There are 268 (34.9%) cases in class ‘1’ and 500 (65.1%) cases in class ‘0’. A brief statistical analysis is given in Table II [10].
This paper reports a comparative study of two machine learning methods, KNN and SVM. Experimental evaluation was conducted on the diabetes dataset to see the difference between two classification methods. Fig.1 shows the general flowchart of study.

**SUPPORT VECTOR MACHINE CLASSIFICATION**

Support Vector Machine (SVM) classification is a supervised learning technique, which was first introduced by Vapnik [11]. SVM is one of the most powerful algorithms for classification problems and commonly used to real-world applications such as text categorization, handwritten digit recognition, and financial forecasting, etc. SVM is based on the structural risk minimization principle and is used for classification problems to accomplish high generalization performance [12-14]. The aim of SVM classification is to generate a model, based on the training data to predict class labels of the test data accurately.

Fig. 2 shows SVM model representation. The dataset is grouped into two different classes which are illustrated as squares and circles. Furthermore, there are many hyper planes (H₁, H₂, H₃, H₄) which separate the datasets linearly [15]. SVM have to find the optimal hyper plane, which maximizes the distance (margin) between the different set of classes and minimize the risk of misclassifying examples of the test dataset, so that SVM has good generalization ability [16].

**K-NEAREST NEIGHBOR CLASSIFICATION**

K-Nearest Neighbor (KNN) is one of the widespread machine learning algorithms. It has been used in many applications in the field of data mining, statistical pattern recognition and etc. [17-18]. It classifies objects, based on closest training examples in the feature vector. Training data is used to form reference space. By the aid of this algorithm, the distance of sample from the all neighbors (reference samples) can be calculated [19].

K-nearest neighbor classification algorithm can be run in these steps [20]:
1. D is the set of training samples (yᵢ), and k is number of nearest neighbors.
2. Calculate d(xᵢ, yⱼ) using Euclidean distance for every sample yⱼ of D and for each test sample xᵢ do
3. To test sample xᵢ, choose the k close set training samples yⱼ (neighbors).
4. Classify the sample xᵢ based on majority class among its nearest neighbors.
5. End.

**RESULTS**

For the analysis, the data is firstly divided into training data and testing data. The training set is used to build the classifier and test set used to validate it. In this study the percentages used for training and testing data are 75% and 25%, respectively.

The accuracy measure, which is used widely to compare the performance of classifiers, was obtained. Furthermore, Sensitivity (true positive rate) and Specificity (true negative rate) statistical measurements that describe how well the classifier discriminates between a case with positive and with negative class were also adopted. We used following equation to measure the accuracy Eq. 1, specificity Eq. 2, and sensitivity Eq. 3 [21].
\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \\
\text{Specificity} = \frac{FP}{FP + TN} \\
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

where \(TP\) (True Positive) is the number of diabetes women detected as patient, \(FP\) (True Positive) is the number of healthy women detected as patient, \(TN\) (True Negative) is the number of healthy women detected as healthy and \(FN\) (True Positive) is the number of patients detected as healthy.

Table III shows the performance of KNN and SVM classifiers on diabetes datasets. The highest classification accuracy is achieved with SVM classifier by 82.29%.

<table>
<thead>
<tr>
<th>Classification Technique</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Nearest Neighbor (KNN)</td>
<td>72.17%</td>
<td>83.87%</td>
<td>50.23%</td>
</tr>
<tr>
<td>Support Vector Machines (SVM)</td>
<td>82.29%</td>
<td>88.80%</td>
<td>70.15%</td>
</tr>
</tbody>
</table>

As shown in Fig. 3, SVM constructed a hyper plane in N-dimensional space for classification the dataset. It separated diabetes dataset such that the distance between the classes is maximum.

**DISCUSSION & CONCLUSION**

The experimental results have shown that both classification techniques give different results on same dataset. Therefore, the classification technique which has shown the highest accuracy rate and lowest error rate over a dataset has been selected as the best classification technique. Results demonstrated that the SVM classifier is faster and significantly more reliable than KNN on diabetes dataset. This technique can be applied on medical datasets to help physicians to make more accurate decisions about diabetes disease.

**REFERENCES**


Interprofessional Education Model for the Health and Life Science Faculties
Ali Yildirim and Sezer Domac

Abstract—Interprofessional education (IPE) aspires to prepare health and social care students to work collaboratively to deliver a patient focus care when they qualify. IPE is now delivered in pre-registration education programmes and must be developed further to meet the needs of the faculties’ curriculums. IPE model is based on the assumption that students will gain professional competencies, for example, knowledge, skills and attitudes at pre-registration level. IPE model is based on team working and a collaborative approach and accommodates the varying degrees of knowledge, skill and competency within the health and social care teams. Applying knowledge in social care situations using triggers such as problem solving that enables students to face a range of situations and understand their findings in the new context of new situations. As a result, pre-qualified students can gain insight into future practice whilst qualified learners are able to look afresh at their current practice. IPE model offers links to future continuing professional development for the postqualification period.

INTRODUCTION

With the increasing number of older and disabled people and the advancement in health and social care live longer, the demographic needs of the population is changing. Consequently, the complexity around health and social care delivery is growing. The need for coordination and integration of health and social care through a multidisciplinary approach has become essential. It is now extensively distinguished that meeting the needs of service users/patients involves expertise from more than one profession and person-centred service can only be achieved with interprofessional collaboration and effective teamwork. This collaborative partnership concept is underpinned in recent UK government policies and legislation and is further elaborated upon through the concept of interprofessionalism. Terminology to define interprofessional collaboration and IPE is problematic. It has been suggested that different forms of interactions need different words and clear conceptualisations. For example, consider, ‘multidisciplinary,’ ‘interdisciplinary,’ ‘crossdisciplinary,’ ‘teamwork,’ ‘partnership,’ ‘collaborative relationships’, ‘coordination’, ‘integration’, ‘interprofessionality’, ‘interprofessional practice’, all terms, which differentiate and overlap (Leathard, 1994, 2003; Reeves et al., 2010). These terms are used in many health’s and social care contexts and are often used to express the coming together of a wider range of health and social care practitioners (Leathard, 2003, p.5). The prefixes of ‘multi’ and ‘inter’ are often used interchangeably with the suffixes of professional and disciplinary (Oandasan & Reeves, 2005, p.23). This has been described as ‘a terminological quagmire’ within which ‘any grouping of terms is debatable’ (Leathard, 2003, p.5). Barr (2010) defined IPE as: the application of principles of adult learning to interactive group-based learning, which relates collaborative learning to collaborative practice. IPE has a long history and over the last decade has become established as a necessary teaching method within health and social care curricula.

Research continues to advance our understanding of the mechanisms of these learning events; helped by appreciating the theoretical underpinnings that explain and shape IPE.1970’s: Global interest and concerns about delivery of health care, role of interprofessional teams, reducing errors; Britain and Canada lead the way. 1980’s: Centre for the Advancement of Interprofessional Education (CAIPE) established in Great Britain 1990’s: Canadian Interprofessional Health Collaborative (CIHC). IOM report To Err is Human (1999) evokes new interest in improving patient safety.2000’s: Institutions undertake IPE initiatives, conferences proliferate, new organizations form First Collaborating Across Borders conference held on the University of Minnesota-Twin Cities campus, Minneapolis, sponsored with the Canadian Collaborative WHO Study Group on Inter professional Education & Collaborative Practice In 2007 After that World health organization (WHO) published a report about the interprofessional education and collaboration. Interprofessional working has been related to a lot of outcomes, covered patient safety and case managing, ideal using of the talent of each health care professionals and anticipation of superior health services (Berridge 2010; Reeves 2010; Suter 2012; Zwarenstein 2000).

The concepts of IPE different from multi-professional education, in multi professional education members have different background and they have learned basic things step by step, however it is not necessary to communicate to the each other or share some knowledge and
experience. On the other hand, IPE have many good points such as, members of IPE share their ideas and knowledge from the others and also they need to communicate professionally and closely (Finch 2000). Overall, IPE is the growing model of new ages education system. Under the shadows of all this, the history of IPE shows the long journey of this new education philosophy. Interprofessional approaches to health and social care are linked to improved clinical services and enhanced problem solving (Mitchell et al., 2010). In addition, IPE target the patient center health and social care. Health profession students educated separately; unprofessional socialization can lead to different approaches and philosophies of care trained to think as well as communicate differently. Patient safety literature emphasizes the importance of ‘interprofessional’ or ‘team approaches’ that is the key strategy to mitigate health and social care human resource challenges facing the health and social care system. In other words, in the health and social care fields, practitioners need to move beyond individual skills and tasks to work collaboratively to high standards to cater for the increasing demands on services. Hence in a competitive economy where the government is seeking to secure and stabilize health and social care, finding more generic skills can drive interprofessional collaboration forward.

**FINDINGS AND DISCUSSION**

There are two types of benefits in IPE; one of them is educational benefit such as, students interactively learn with and from each other. IPE educators also have chance to reflect and monitor the students’ progress from different professional perspectives and develop/adjust programs that will suit the needs of the curriculum. The other important benefit of IPE is about the enhancement of productivity of the professionals and collaborative practice in workplace. This can be positively effect staff’s confidence and improve the patient’s safety (WHO, 2010) Furthermore improved networking with the other professionals as it aims to encourage professionals to work together and learn together. IPE fosters respect for the integrity and allows contribution of other professionals towards the patient care and safety. In addition, it develops an open and effective communication. Furthermore, it increases the level of understanding of each professional role and dispels stereotypes about each other. Delaney at al (2013) have advocated for the need for team-based, collaborative care delivery as the hallmark for delivering high quality patient centered care.

In IPE there is a need to provide evidence of the benefits of this new pedagogy if it is to become accepted into professional education (Barr, 2002). Reeves et al., (2010) carried out an update of a Cochrane systematic review and stated more thorough mixed method studies of IPE are needed to offer a greater clarity of IPE and its effects on professional practice and service users’ care. There are only a few studies, which examine over a certain time after post-qualification, evaluating the effectiveness and competencies related to IPE. Long-term studies should also cover learning routes and the transfer of learning into sustained post-qualifying practice. Service users’ and careers’ views on the nature and effectiveness of IPE could and should be explored. The cultural aspects of how learners perceive and learn from IPE are an area that also requires further examination. The ideal IPE model is related to interprofessional learning outcomes and activities, which means that team based learning activities shows the quality of IPE models. Within these activities, Students have opportunities to shows their role and learn the others roles as well as they need to communicate each other’s learn together and work together. In addition learning is interactive, experiential and reflective process. The key stage of the learning activities should be invented to dispute the stereotype and breakdown hierarchies (Thistlethwaite et al, 2007). The positive outcomes of IPE might be include the such things positive reaction to the learning, new knowledge, positive attitude, working ability with other professions.(Freeth et al 2005)And also some other parameters affect the IPE outcomes, which are communication, delivery of the care and collaborative working. The potential of IPE learning outcomes is related to professionals roles, occurrence, worth’s and professionalisms. IPE can improve the communication skills of the student and professionals and also change their attitude of the leadership in the team (Thistlethwaite et al., 2007). Freeth et al in 2005 mentioned that the effectiveness of IPE are in five mechanisms. The first is clinical effectiveness and others are evidence-based practice, cost effective, professionalism and ethical practice, clinical governance.

**CONCLUSIONS**

IPE shows that the professionals who work and learn together can increase the outcomes for the patients. However the working and learning together has some barriers such as, working culture, training, and qualification. Furthermore there are some advantages of IPE, for example, patient centre perspective, working together technics, communicate the each other and learn about the others roles in the team. On the other hand IPE has some disadvantages that are related to the organising and coordinating the IPE programmes between different professionals. This includes managing the time and culminating in a fear of loss of individual professional identity. Indeed, in the new century policy designers and curriculum makers should need to change the traditional
education model to the IPE model. This model is based on the observation-based learning and linked to the psychoanalytical and learning theories and aims to create more reflective practitioners. The model is based on the exchange-based learning and action based learning that encourages students to move away from the traditional learning to problem-based learning and an alternative approach, collaborative inquiry and stimulation-based learning.

REFERENCES

[8] CAIPE. The Centre for the Advancement of Interprofessional Education. Received from http://www.caipe.org.uk
Abstract—*Artemisia* plant is the largest genus of the Asteraceae family. The species belonging to this genus are widespread in the western regions from Iraq. *Artemisia* herba Alba known as “shieh” is used in traditional medicine due to its several therapeutic characteristics. The plant was collected from the area of Al-Baghdadi, in Al Anbar Province/ Iraq; essential oils were obtained by hydrodistillation using a Clevenger apparatus. Whereas alcoholic extract was obtained by ultrasonic bath technique. The yields were 0.98% ‘w/v’ and “22%” respectively. Effectiveness of different concentrations of extract against four different types of bacteria was tested. The bacteria involved in this study were: *E. coli*, *Staphylococcus aurous*, *pseudomonas* and *Bacillus pumilus*. Results showed that both of extracts (essential oils and alcoholic extract) have antibacterial activity and can be applied as antiseptic.

**INTRODUCTION**

*Artemisia* plant is from compositae family and Radiate subgenus. There are different species with different names, most of which have bitter, aromatic leaves and more or less possess the same medicinal properties[1]. The chemical studies on *Artemisia* species indicate that all classes of compounds are present in the genus with particular reference to terpenoids and flavonoids[2]. *Artemisia* Herbal Alba includes 1.2% essence based on the weight of dry basis which includes 39% camphor, limonene and 1, 8-cineol (15%), camphene (6%) and alpha-Pinene 5%. Camphor is one of the main ingredients of *Artemisia* Herba Alba and is antiseptic. After alpha-Pinene, 1, 8-cineol is the most abundant ingredient in the essence which has an extensive use in pharmaceutical preparations[3].

The aim of this study is to prepare new formula of *Artemisia* extract and test the effectiveness of different concentrations of this new formula on four different types of bacteria to use it as an antiseptic and sterilizer similar to Dettol.

FURTHER INFORMATION

*Artemisia* plant collection:

*Artemisia* Herbal Alba was collected from the area of Al-Baghdadi, in Al Anbar Province/ Iraq. They were washed with water and dried at the room temperature after being approved by a botanist expert, then grinded to coarse powder by electrical grinder and kept in clean plastic bottle.

Preparation of Essential Oils:

100 g of fresh leaves of *Artemisia* species were added to 500 ml of distilled water, using Clevenger apparatus for 6 hours. Essential oils were obtained, then were dried using anhydrous sodium sulfate and stored at 4°C in dark bottles.

Preparation of alcoholic extract:

500 ml of the Ethanol (60%) was added to 100 g powder of fresh leaves of *Artemisia* species and was placed on the Ultrasonic device for 24 hours. This solution was dried using a spray drying after passing through the 0.22 μm membrane filter, fine powder was then obtained.

Preparation of *Artemisia* antiseptic formula:

*Artemisia* antiseptic 10% (w/v) Concentrate for Cutaneous solution, contains *Artemisia* extract as the active substance. 63.6 g of castor oil had been Mixed with 200 ml ethanol, then 13.4 g of potassium hydroxide solution was added and mixed well for a period of one hour, the volume was completed with distilled water to 1 liter, then 40 ml of pine oil and color were added with stirring for 24 hours, finally 100 g of powdered *Artemisia* were added to the solution with continued stirring for one hour, packaged solution and well closes.

Preparation of the Test Microorganisms:

Bacterial strain was opened in sterile conditions according to the manufacturer’s instruction[4]. The basic culture medium was prepared in tryptone soy broth and tryptone soy agar medium. A storage culture medium was prepared from the resulting culture and was used in the subsequent stages. A part of the culture was inoculated to the nutrient agar slope medium and was incubated at 37°C for 24 hours at room temperature. The colonies of medium surface were washed with normal saline and

---

*Corresponding Author (e-mail: enas.mn@gmail.com).
Enas Mehjen Numan (PhD).
Ibn Al-Betar Research center, corporation of research and industrial development, Ministry of industry and minerals, Al-Jadriyah, Baghdad, Iraq.
bacterial suspension was diluted with normal saline to make the level of absorption at a wavelength of 53 nm to be equal to the 0.5 McFarland.

Evaluation of Antimicrobial activity (NCCLS, 1990):

Antimicrobial activities of the extracts were measured twice using agar dilution tests and based on recommendations by the National Committee for Clinical Laboratory Standards (NCCLS).

Fifteen milliliters of agar Mueller- Hinton medium was mixed with different concentrations of the essential oil of Artemisia and 104CFU/ml of each sample of bacteria was inoculated to it. Plates were incubated for 24 hours at 37ºC. Minimum inhibitory concentration (MIC) was determined as the lowest concentration of each extract that prevents bacterial growth in the culture medium. During the next step, fluid diluting technique was used to determine the minimum bactericidal concentrations of the extracts of Artemisia. To perform this experiment; equal volumes of each bacterial suspension including 105CFU/ml were inoculated to the Mueller - Hinton broth medium containing various concentrations of the extract of Artemisia.

Then mediums were incubated for 24 hours at 37ºC. One hundred micro liters 100 of each liquid culture were subsequently cultured on Mueller - Hinton agar medium and again were incubated at 37ºC for 24 hours. The minimum bactericidal concentration (MBC) was determined as the lowest concentration of the extract of Artemisia that completely killed the bacteria and no growth was observed.

RESULTS:
The antimicrobial activity of Artemisia essential oil on E. coli, Staphylococcus aureus, pseudomonas and Bacillus pumilus were (17mm, 21mm, 5mm and 24 mm respectively) compared with the positive control (standard Gentamysin) 22mm. It was found that the essential oil of Artemisia inhibits the growth of all bacteria. Also, for the antimicrobial activity of Antiseptic solution of Artemisia on bacterial species table (1). The results indicate the effectiveness of the disinfectant solution to Artemisia against different bacteria is similar or more effectiveness than Dettol.

<table>
<thead>
<tr>
<th>Extract concentration</th>
<th>E. coli</th>
<th>Staphylococcus aureus</th>
<th>pseudomonas</th>
</tr>
</thead>
<tbody>
<tr>
<td>7% g/ml</td>
<td>25mm</td>
<td>25mm</td>
<td>-</td>
</tr>
<tr>
<td>4% g/ml</td>
<td>24mm</td>
<td>21mm</td>
<td>11mm</td>
</tr>
<tr>
<td>2% g/ml</td>
<td>24mm</td>
<td>25mm</td>
<td>12mm</td>
</tr>
<tr>
<td>1% g/ml</td>
<td>22mm</td>
<td>16mm</td>
<td>-</td>
</tr>
<tr>
<td>0.5 % g/ml</td>
<td>20mm</td>
<td>21mm</td>
<td>-</td>
</tr>
<tr>
<td>0.25 g/ml</td>
<td>19mm</td>
<td>20mm</td>
<td>-</td>
</tr>
<tr>
<td>Positive control (mg/ ml)</td>
<td>17mm</td>
<td>18mm</td>
<td>16mm</td>
</tr>
</tbody>
</table>

CONCLUSIONS
This study showed an identical similarity between the biological activity of Antiseptic Artemisia solution and Dettol, this finding may be useful in replacing Artemisia Herbal Alba extract as active substance instead of the active chemical compound (Chloroxylenol) in Dettol, especially it is natural product and Environmentally friend.

REFERENCES
Extraction, Characterization and evaluation of the activity of Adhatoda vasica as Pesticide.

Ibn Al-Betar Research center, corporation of research and industrial development, Ministry of industry and minerals, Al-Jadriyah, Baghdad, Iraq.

Abstract— Adhatoda vasica which belongs to (family: Acanthaceae), was collected (leaves) from the gardens of Baghdad University, Jadreia/ Iraq. The leaves were extracted using solvents of varied polarity. Preliminary phytochemical investigations were carried out of the extracts obtained from the leaf of Adhatoda vasica. The presence of tannins, terpenes, alkaloids, resins, glycosides, carbohydrates was indicated by the tests conducted. The effect of ethanol (95%) extract was tested against two kinds of insects (Aphid and house flies).

INTRODUCTION

Adhatoda vasica belongs to the family of Acanthaceae. It is an erect, terrestrial, perennial shrub. The leaves are dark green above and pale yellow below. The flowers are typical, white, arranged in a pedunculated spike[1]. The variations of chemical constituents are observed in plants belonging to the same species grown in different environmental conditions. These variations may occur due to climatic conditions like altitude, temperature, type of soil etc[2]. The plant extract also finds use as an expectorant, abortificient and antiseptic[3].

The aim of this work is to evaluate the activity of Adhatoda vasica plant as pesticide, through tested the toxicity of best concentration of its extract against two kinds of insects (Aphid and house flies).

FURTHER INFORMATION

Adhatoda vasica plant collection:

Adhatoda vasica was collected from the gardens of Baghdad University, Jadreia/ Iraq. They were washed with water and dried at the room temperature after being approved by a botanist expert, then grinded to coarse powder by electrical grinder and kept in clean plastic bottle.

Preparation of ethanolic extract of leaves:

100 g of Adhatoda vasica leaves (powdered) was weighed and soaked in 250 ml of ethanol (95%) with shaking for 24h., the plant extract was then collected and filtered through Whatman No.1 filter paper. The extract was concentrated at 50°C using a rotary evaporator and then air-dried. The dried powder was stored in an airtight bottle.

Toxicity test:

The toxicity test of ethanol extract for Adhatoda vasica with different concentrations (10%), (20%) and (40%), on three groups of mice (BALB/c). The ethanolic extract was injected to mice by intraperitoneal method under the skin, about (0.1-0.2) ml twice daily for 72h.

The results showed that the ethanolic extract (10%) Conc. Was not toxic, while the (20%) Conc. was toxic and the (40%) Conc. was very toxic for the mice.

Other three concentrations of the Adhatoda vasica alcoholic extract were prepared (4,6,8)%. Mixed with larva nutrient and placed in plastic jar. The ratio of killing were estimated after (3, 5, 7, 10) days of treatment, table I.

<table>
<thead>
<tr>
<th>Alcoholic extract Conc. %</th>
<th>Percentage ratio of Death after 3 days</th>
<th>5 days</th>
<th>7 days</th>
<th>10 days</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6.67</td>
<td>23.00</td>
<td>45.00</td>
<td>73.33</td>
<td>37.00</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>27.00</td>
<td>51.33</td>
<td>78.33</td>
<td>42.91</td>
</tr>
<tr>
<td>8</td>
<td>16.67</td>
<td>31.67</td>
<td>56.00</td>
<td>81.67</td>
<td>46.50</td>
</tr>
</tbody>
</table>

CONCLUSIONS

This study showed the toxicity of the Adhatoda vasica extract at special concentration against two kinds of insects (Aphid and house flies) that may be useful to replace it instead of chemical pesticides, therefore it will be bring down pollution levels in soil and water.
REFERENCES


Feasibility Study of Telemedicine Implementation in West of Iran*
(Imam Khomeini Hospital: a case study)

Ali Roozbahani1, Abdolhossein Ayoubi1, Fatemeh Soleimani Roozbahani2
1Amir Kabir Technical university-Biomedical dept. Tehran- Iran
2Islamic Azad University- Borujerd Branch- Technical and Engineering dept- Borujerd- Iran

Abstract—The healthcare organizations that are confronted with the new technology of telemedicine, provide some bases for co-operating, training and consulting. Telemedicine is the use of information Technology (IT) for clinical care of patients. Due to the deprivation of many regions in west of Iran, being Mountainous Region, dispersion of rural context and being away from the center, the access to medical care is a big problem. The study surveys the possibility of implementing the telemedicine in this zone. The research has been done as a cross sectional study. The study proper population include: the specialists, medical staff and IT sector staff of Imam Khomeini Hospital (a major hospital in west of Iran). The population size and sample size were respectively the same. Presenting personally, the researchers used a standard questionnaire of telemedicine assessment tool, provided by Medicare institute to gather data. The data were analyzed by the SPSS Software. Major part of specialists, expressed that consolation is the best usage of telemedicine, and they believed that lack of technical staff is the biggest obstacles of implementing telemedicine. The other results will be present in the paper. The result can help the health and care officials to achieve the best condition for telemedicine implementation.

Keywords—telemedicine; feasibility study; telehealth; west of Iran; Imam Khomeini Hospital

INTRODUCTION

The new area of healthcare is the field of co-operation of physicians, hospitals, medical center, the financial and insurance experts, in a virtual environment, in which the subject is, goal, profit, and instrument in policy and management of information medical care (1). The application of IT in health industry, especially in hospitals and medical centers, provides a great potential for improving the quality of services provided and efficiency of the effectiveness of personnel (2). The rapid development of technology and health informatics has encouraged the hospitals to gain the necessary infrastructures and skills in order to provide better quality and more developed health services (3). Today the health care organizations are faced with a new technology called telemedicine (4). Which provides some channels for a long distance collaboration, training, and consolation (5). In fact the telemedicine is the use of communicating technology for the clinical care of patients, and includes different mechanisms of offering electronic services. Telemedicine is a tool that enables providers to provide health care services to the patients who are far away.

Telemedicine systems have been transformed during the recent decades and have spread in 1990s, due to the major advances of communicating technology and the reduction of equipments and transportation costs (6).

In 90s two advances were made in technology which lead to increasing the interest in telemedicine. The first one was the increasing spread of broad band width with top speed around the world. The second case was inventing some tools, capable of capturing and transmitting data and images in digital format (7). In telemedicine not only the hardware and software play important roles but also care providers and users are considered (8). Four factors are involved in the success of telemedicine: motivation of therapist, appropriate infrastructure, ongoing investment, and technical support (4).

According Bashshur theory, Grisby believes organizational factors influencing the failure of telemedicine technology in manufacturer institutions, are the mechanisms of this technology, i.e. insufficient planning, poor design, and conflicting interests and expectations of these groups to telemedicine. In fact those who have to accept the telemedicine are not sure about the ways of improving the patient’s life and their reliability (10). In general it can be said that the current generation’s task is to maintain the quality of care through using the open attitudes with the advantages of technology.

In other countries, many studies have been done about the conditions of using new cases and evaluation of systems in use. In Iran paying attention to the theoretical foundations of electronic health has been began since the middle of the last decade and actions have been begun in recent years to operate some systems and software as HIS, EMR, … . In the field of telemedicine also a case of operating telemedicine in Xoramabad, in 2006, second conference of strategies for developing IT, in universities of medical sciences, has been reported. A case of telemonitoring, in Birjand, 2006, has been also reported. In 2008, also a plan named MAPAD is starting in Masih Daneshvari and Firoozkooh hospitals, which intended to use long distance consolation services (13). But its results were not published. The foundation of Martyrs and Veterans also, is going to optimize this technology to
provide services to Veterans with spinal cord injury and those who are unwell.

The west region of Iran with its large geographical area and its various scattered urban and rural areas, is faced with the shortage of specialists in different medical fields. At the other hand the majority of areas of this region are considered as the disadvantaged area, have been created some problems for the patient to visit the specialist and practitioner. According to the importance, benefits, properties, and difficulties of deploying this technology, the present research is to evaluate the feasibility of it in current situation of west of Iran, to offer strategies for using telemedicine in an efficient and effective ways, if necessary.

**METHODOLOGY**

It was a cross- sectional research that was done to determine the feasibility of implementation the telemedicine in west of Iran. Subjects studied include: the possible cases of using telemedicine, identifying obstacles for using telemedicine, determining the condition of medical employee for accepting the telemedicine technology, identifying the condition of executive culture for applying telemedicine to determine the compliance rates of existing infrastructure and technical requirements of telemedicine. The researchers data have been gathered by the use of a standard questionnaire called “telemedicine assessment tool”, which are gathered by Medicare Institute in the U.S. to evaluate readiness of volunteer organizations to employ telemedicine, therefore there is no need for its validity and reliability evaluation. The medicare institute is an institute which determines the health care policy and the cases related to the refund. The questionnaire includes three parts. The first part is completed by the highest executive authority of the hospital, the second part by the head of the IT department, and the third one by the specialist and medical staff who have been working in the hospital. Therefore, the study population consisted of president of hospital, the head of IT department, and 474 specialists who have been working in hospital (total number =478). The population size and the sample size were the same. Through visiting 474 specialists twice, 282 patients completed the questionnaire. By using the questionnaire the collected data, were entered. The SPSS software and analyzed through descriptive as frequency and percentage.

**RESULTS**

Among 282 persons who completed the questionnaire, 200 person (70.9%) were male, and 81 person (29.1%) were female. 224 persons (79.4%) specialist, and 58 (20.6%) professionals who all had the master of science degree. 49 persons (34.8%) were the member of faculty of knowledge. In addition to demographic data, the finding are presented in the following three tables as (a) the distribution of physician by their specialty, (b) prioritization of services used in telemedicine from the experts’ perspective and (c) prioritization of barriers of telemedicine from the physicians’ perspectives.

**Table 1. Distribution of physicians by specialty type**

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>13</td>
<td>9.9</td>
</tr>
<tr>
<td>Heart</td>
<td>10</td>
<td>3.5</td>
</tr>
<tr>
<td>Eye</td>
<td>18</td>
<td>6.4</td>
</tr>
<tr>
<td>ENT</td>
<td>12</td>
<td>4.3</td>
</tr>
<tr>
<td>Kidney and urinary tract</td>
<td>8</td>
<td>2.8</td>
</tr>
<tr>
<td>Anesthesia</td>
<td>18</td>
<td>6.4</td>
</tr>
<tr>
<td>Radiology</td>
<td>4</td>
<td>1.4</td>
</tr>
<tr>
<td>Pathology</td>
<td>6</td>
<td>2.1</td>
</tr>
<tr>
<td>Obstetrics and gynecology</td>
<td>24</td>
<td>8.5</td>
</tr>
<tr>
<td>Pediatric</td>
<td>34</td>
<td>12.1</td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>6</td>
<td>2.1</td>
</tr>
<tr>
<td>General surgery</td>
<td>24</td>
<td>8.5</td>
</tr>
<tr>
<td>Infectious diseases</td>
<td>4</td>
<td>1.4</td>
</tr>
<tr>
<td>Dermatology</td>
<td>6</td>
<td>2.1</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>10</td>
<td>3.5</td>
</tr>
<tr>
<td>Emergency medicine</td>
<td>6</td>
<td>2.1</td>
</tr>
<tr>
<td>Psychology</td>
<td>10</td>
<td>3.5</td>
</tr>
<tr>
<td>Neurology</td>
<td>8</td>
<td>2.8</td>
</tr>
<tr>
<td>Unknown</td>
<td>40</td>
<td>14.2</td>
</tr>
<tr>
<td>Total</td>
<td>282</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 2. prioritization services used in telemedicine from experts’ perspectives**

<table>
<thead>
<tr>
<th>Type of service</th>
<th>Pro</th>
<th>Con</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>consultation Numbe</td>
<td>Number</td>
<td>Percent</td>
<td>Number</td>
</tr>
<tr>
<td>consultation Numbe</td>
<td>242</td>
<td>85.8</td>
<td>18</td>
</tr>
<tr>
<td>Training services</td>
<td>202</td>
<td>71.6</td>
<td>22</td>
</tr>
<tr>
<td>Referral of patients</td>
<td>112</td>
<td>61</td>
<td>48</td>
</tr>
<tr>
<td>Purchase of equipments</td>
<td>154</td>
<td>54.6</td>
<td>32</td>
</tr>
<tr>
<td>Care services</td>
<td>144</td>
<td>51.1</td>
<td>42</td>
</tr>
</tbody>
</table>
The majority of specialists, declared the lack of technical staff, the initial costs, and insurance and refund problems, are the main obstacles to employ telemedicine. The least important obstacles were expressed as competition and confidentially issues, and the result of every country affect on abilities of telemedicine, which will be discussed in infrastructure section.

The results of research of enablers and barriers has been done by group of economic benefits monitoring of the U.S. shows the most attractive aspects of telemedicine include: Rapid response to health care, monitoring patients with chronic diseases and who are living in inaccessible areas, providing the first level of care in inaccessible areas, rapid response to the incidents, home care, specialists shortages compensating and costs reducing.

The important point in this study was the not considering the importance of cost reducing compared with the associated factors with the performance of telemedicine (like as rapid response and patients monitoring). Researchers believe if the cost reducing interpreted to lower payments and lower actual costs, perhaps the results would be changed (15).

Debnath in a research between 1991 and 2003 in the U.K. introduced the factors of education, emergency medicine, general medicine as the most utilities of telemedicine (16).

According to the results of the study and similar studies, it can be resulted that the use of telemedicine is somewhat similar, but the level of information technology infrastructure and the nature of health sector in every country affect on abilities of telemedicine, which will be discussed in infrastructure section.

The results of studies showed the barriers of telemedicine from physicians’ point of view include: Shortage of technical staff, initial costs, insurance and reimbursement issues, running costs and lack of medical and educational staff. The attitude of staff, licensing problems, time constraints, staff resistance, medical privacy, and competition issues are also other expressed barriers with lower priorities. Several barriers to use of telemedicine are mentioned in different studies. The study of enablers and barriers introduced the major barriers of telemedicine as: Strong resistance from local health services providers and their fear of job loss, inadequate insurance coverage, poor quality of diagnosis or care, and purchasing costs. In the study, inability of current mechanisms in acceptance of change has been considered as a major obstacle which is even greater than the fear of lower quality of care (15). Thus providing

<table>
<thead>
<tr>
<th>Utilization obstacle</th>
<th>Major</th>
<th>Partial</th>
<th>No obstacle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>Percentage</td>
<td>Number</td>
<td>Percentage</td>
</tr>
<tr>
<td>Shortage of technical staff</td>
<td>194  68.8%</td>
<td>72   25.5%</td>
<td>16   5.7%</td>
</tr>
<tr>
<td>Initial costs</td>
<td>188  66.7%</td>
<td>66   23.4%</td>
<td>26   9.2%</td>
</tr>
<tr>
<td>Insurance problems and refunds</td>
<td>188  66.7%</td>
<td>54   19.1%</td>
<td>40   14.2%</td>
</tr>
<tr>
<td>Current costs</td>
<td>170  60.3%</td>
<td>76   27%</td>
<td>36   12.8%</td>
</tr>
<tr>
<td>Lack of medical staff</td>
<td>156  55.3%</td>
<td>94   33.3%</td>
<td>32   11.3%</td>
</tr>
<tr>
<td>Education</td>
<td>146  51.8%</td>
<td>82   29.1%</td>
<td>54   19.1%</td>
</tr>
<tr>
<td>Staff attitude</td>
<td>116  41.1%</td>
<td>108  38.3%</td>
<td>58   20.6%</td>
</tr>
<tr>
<td>Licensing problems</td>
<td>102  36.2%</td>
<td>106  37.6%</td>
<td>72   25.5%</td>
</tr>
<tr>
<td>Time limit</td>
<td>88   31.2%</td>
<td>140  49.6%</td>
<td>54   19.1%</td>
</tr>
<tr>
<td>Medical staff resistance</td>
<td>88   31.2%</td>
<td>118  41.8%</td>
<td>76   27%</td>
</tr>
<tr>
<td>Confidentiality issues</td>
<td>80   28.4%</td>
<td>114  40.4%</td>
<td>88   31.2%</td>
</tr>
<tr>
<td>Competition</td>
<td>58   20.6%</td>
<td>102  36.2%</td>
<td>122  43.3%</td>
</tr>
</tbody>
</table>

Table 3. prioritizing obstacles of using telemedicine from experts perspectives.

includes telemedicine, and the executive and the president of hospital will support implementation of telemedicine. There is a telemedicine champion in hospital but no services have been provided there, yet.

**DISCUSSION AND CONCLUSION**

The physicians have recognized the factors of consulting, education, patients referring, equipment purchasing, and care providing as the priorities of using telemedicine.

Among other 268 persons, who did not experienced providing long distance medical education, 164 persons (58.2%) intended to represent such methods, 12 persons (4.3%) declares that they are not ready to provide such education, and 46 persons (32.6%) did not have any comment.

About the administrative culture, the results showed that the hospital has encompasses strategic plan which

\[175\]
telemedicine system requires effective change management (17).

Jennet et al. recognize the factor of resistance of medical staff as the most important obstacle in telemedicine success (18).

Convincing the involved specialists about the usefulness and feasibility of implementing new methods and their active participation in telemedicine services is important for interesting them and inducing them to support the project (19).

The other study has shown that suspicious at the beginning of the telemedicine implementation, after successful implementation, has become desire (20).

In the other side, Debnath’s study introduces the factors of financial constraints, lack of infrastructure, MedicoLegal concerns and resistance to change as the obstacle of telemedicine in Britain (16). This study believes lack of close interaction with the patient’s prevents the active participation of some specialist (16).

Mengistu et al. In a study titled "Telemedicine transmission model for sub-Saharan Africa," mentioned other barriers to the widespread use of telemedicine such as: lack of resources, poor infrastructure, telecommunications (broadband), human and organizational factors, policies and standards, and social - economical issues. This study focuses on the lack of infrastructure in which telemedicine cannot be successful and will continue to work (21).

Investment on telemedicine depends on the fact that telemedicine service compared to other forms of services what has positive characteristics. Discussions about the costs of telemedicine have been more limited to the cases associated with the budget and do not have pervasive economic perspective. Making decisions about the cost of such be widely programs should involve different ideas of patients, health care specialists and organizations which investment in health care (19).

In a study conducted by Kennedy and Yellowlees in two rural health center which used teledermatology services, results revealed from the social point of view (in comparison with the traditional method) costs have been reduced but from the providers point of view costs have increased four-fold. Although the provider point of view is important, but it is only one of the factors involved in the decision making process and by emphasis on the social perspective using teledermatology is justifiable. Selecting the social perspective and avoid focusing solely on cost savings is a good way as other aspects of health care, political and administrative support (22).

Tahir and Wooten’s study of a pilot telemedicine system in Malaysia identified the reasons for its failure as: lack of coordination between telemedicine and traditional system in patients referral, changing the telemedicine coordinators, low bandwidth, the lack of reliability of the network, poor maintenance and change management due to the telemedicine application (23). Issue related to license is one of major problem in the U.S. for teleconsultations (24).

In general it can be concluded that the most important physicians’ priorities about the telemedicine, were offering, providing education and referral of patients. The physicians also expressed the important barriers as: the shortage of technical staff, insurance and reimbursement issues, the initial costs, the current costs, and the shortage of medical staff. Data shows the physicians need to telemedicine is necessary. There is an operational culture for the implementation of telemedicine in a good level. The current band width of hospital can only respond the lowest levels of telemedicine, therefore increasing the band width and providing the necessary IT infrastructure is required for optimal use of telemedicine technology.

REFERENCES

4) Bangert D, Doktor R. The role of organization culture in management of clinical e-health systems, proceeding of 36th Hawaii international Conference on System Sciences. 2003 Jan 6-9: Hawaii, USA.
10) Dargahi H. An investigation about attitude of clinical physicians in the implementation of telemedicine technology in tum's hospitals. The


CNI-1493 administration improves the efficacy of cytotoxic T lymphocytes

Salem Ramadan Sarite\textsuperscript{1*}, Sabine Specht\textsuperscript{2} and Achim Hoerauf\textsuperscript{2}.

\textsuperscript{1}Libyan Academy – Misurata - Libya.
\textsuperscript{2}Institute for Medical Microbiology, Immunology and Parasitology - University of Bonn, Sigmund-Freud-Strasse 25, 53105 Bonn, Germany

*To whom reprint requests should be addressed. e-mail: salemsrs@yahoo.com.

Abstract CNI-1493, is a synthetic guanylhydrazone with anti-inflammatory properties. The most important application of CNI-1493 is the treatment of Crohn’s disease. Adenovirus expressing ovalbumin (AdOVA) induces an increase, as a short peak, of IL-12 shortly after administration, and increased ovalbumin-specific cytotoxic T lymphocyte cells.

Aim of study: To evaluate the safety and immunological effects of the CNI-1493 on the in vivo activity of cytotoxic T lymphocyte cells.

Materials and methods: AdOVA were given i.v. at the same time point to two groups of naïve C57Bl/6 WT mice; Gp1: Received CNI-1493, AdOVA and target cells and Gp2: Received AdOVA and target cells. In addition to the above two groups, a control group (Gp3): Received target cells (Neither CNI-1493 nor AdOVA). Five hours, after the target cells administration, the spleens were taken out of the experimental groups and homogenised in 1xPBS. The ratio of lysed target cells was determined by flow cytometry.

Results: The OVA specific lysis of target cells in immunized C57Bl/6 mice of the experimental groups were 47.2% in G1 and 81.3% in G2.

Conclusion: CNI-1493 is compound which is safe and undamaging to the immunological function of the cytotoxic T lymphocyte cells.

Keywords: CNI-1493, ovalbumin, specific cytotoxic T lymphocyte cells, and immunized C57Bl/6 mice.

INTRODUCTION:

CNI-1493, (N,N’-bis[3,5-bis[1-(aminooimino-methyl) hydrazonoethyl]phenyl] decanedi -amide tetrahydrochloride), formerly known as semapimod, is a synthetic guanylhydrazone with anti-inflammatory properties\textsuperscript{(5)}, including TNF and NO\textsuperscript{(6)}. Initial studies suggested that the mechanism of the anti-inflammatory action of CNI-1493 involves the inhibition of the phosphorylation of p38 MAPK, which plays an integral role in the translation of mRNA of pro-inflammatory cytokines such as TNF\textsuperscript{(5,21,24)}. In addition it involves the inhibition of JNK signaling pathways\textsuperscript{(9)}. Thus, the anti-inflammatory properties of CNI-1493 have been applied in a variety of diseases. Suppression of renal cancer and melanoma was shown in a clinical phase I trial study\textsuperscript{(19)}. CNI-1493 administration reduced myocardial inflammation and myocyte apoptosis. Recently, the compound was also shown to inhibit trans activation of human immunodeficiency virus type 1 (HIV-1) protein Rev\textsuperscript{(14)}, which is essential for virus replication; inhibition was also observed in multi-drug-resistant HIV strains. This inhibitory effect was based on suppression of deoxyhypusine synthase (DHS) and subsequently of eIF-5A biosynthesis, which is an essential cellular cofactor of the HIV-1 Rev regulatory protein\textsuperscript{(24)}. The most important application of CNI-1493\textsuperscript{(17)}, which is now under evaluation in clinical trial phase II, is the treatment of Crohn’s disease, a chronic inflammatory disease of the gut. Moreover, treatment with CNI-1493 resulted in a profound inhibition of lipopolysaccharide (LPS)-induced production of TNF, IL-1, IL-6, and IL-8\textsuperscript{(7)}. In addition to that, our research group had proven that CNI-1493 inhibits \textit{P. falciparum} DHS and it provide in vivo protection against cerebral malaria of mice infected with \textit{P. berghei}\textsuperscript{(15,21)}.

Interleukin-12 (IL-12) is a heterodimeric cytokine that is produced predominantly by macrophages, dendritic cells (DCs), and a variety of other immune cells and serves as a key regulator of cell-mediated immunity\textsuperscript{(29)}. IL-12 has potent biological effects in vitro and in vivo. In addition, it has been shown to (a) induce IFN-γ production, (b) augment cytotoxic function and proliferation of NK and activated T cells, and (c) promote Th-1 type cytokine responses\textsuperscript{(29)}. In addition, Endogenous IL-12 is clearly important for protective immune responses during acute infections with intracellular parasites\textsuperscript{(22)} and bacteria\textsuperscript{(26)}.

In contrast to that, Orange et al, 1995\textsuperscript{(20)} suggested a pathway by which infection with agents inducing IL-12 and/or TNF, during a viral infection, might dramatically limit protective anti-viral immune responses and induce physiologically adverse effects on the host. In agreement with the above findings, when given systemically in a phase I clinical trial, recombinant IL-12 induced multiple serious adverse effects, including renal and systemic toxicity\textsuperscript{(10,16)}. High-dose levels were linked to temporary immune suppression, which would be unfavorable for effective immunotherapy. However, low doses of IL-12 (1-10 ng/d) enhance, whereas high doses (100-1,000...
ng/d) inhibit CD8+ T cell responses. In fact, both CD8+ T cell expansion and presence of virus-specific CTLs are reduced with the higher doses of IL-12[12]. As high dose IL-12 administration induces numerous necrotic lesions and apoptotic cells in lymphoid tissue, the effects on T cell responses appear to be associated with cell death. **Adenovirus expressing ovalbumin (AdOVA)** induce an increase, as a short peak, in IL-12 level[13,18], shortly after administration, and increased ovalbumin (OVA)-specific cytotoxic T lymphocyte cells (CTL)[28].

**Aim of study:** To evaluate, how does the administration of CNI-1493 will affect the in vivo cytotoxicity of CTLs?

**MATERIALS AND METHODS:**

**Animals:** The animal studies were performed according to the guidelines of the German animal rights (license number: 50.203.2-BN). Where C57BL/6-WT mice were obtained from Harlan.

**CNI-1493 and Adenovirus administration:** CNI-1493 was dissolved in DMSO and prepared in PBSx1 at 4mg/kg to be injected intravenously (i.v.).

1x10^8 pfu/mouse of AdOVA were administrated i.v. at the same time to two groups of Naïve C57Bl/6WT mice; Gp1 received CNI-1493 (pre-treated group, treated on -5, -3 and -1 day before AdOVA administration), AdOVA and target cells. Gp2 received AdOVA and target cells (Non-CNI-1493 treated). In addition to the above two groups, a control group (Gp3) received target cells (Neither CNI-1493 nor AdOVA). (Figure 1: Illustrates the experimental time points).

Recombinant Adenovirus type 5 expressing OVA were kindly provided by the Institute for Molecular Medicine and Experimental Immunology (IMMEI), Bonn, Germany.

**Target cells administration:** Each mouse has received 1x10^7 cells from the 58L target cells which given i.v. five days after immunization with AdOVA to the all three experimental groups. Five hours later the spleen was taken out and homogenised in 1xPBS. The ratio of lysed target cells was determined by flow cytometry (By using: FACS CANTO Becton Dickinson GmbH, Germany). The ratio of lysed target cells was determined by the following equation:

![Flow Cytometry Image](image)

**RESULTS:**

In particular, the release of increased levels of pro-inflammatory cytokines (i.e. IL-12, IFN-γ and TNF) is of importance in different stages of systemic, infectious diseases, because it has a beneficial role in host defense. However, the pro-inflammatory cytokines could be harmful to the host if produced excessively. The results described here document the effects of CNI-1493 on the CTLs in vivo cytotoxicity.

**CNI-1493 in Vivo cytotoxicity assay:** 4 mg/kg of CNI-1493 was administrated i.v., for 5 days before Ad-OVA infection (Gp1). The generation of OVA specific cytotoxic T cells (CTLs) in CNI-1493 treated or untreated mice was evaluated by flow cytometry. The OVA specific lysis of target cells in immunized C57Bl/6 mice of the experimental groups were 47.2% in Gp1 and 81.3% in Gp2 as shown in (Fig. 2). On the other hand,
no specific lysis of target cells in Gp3 (specific lysis of target cells was 0%).

Discussion:
Effective cell-mediated immunity, essential for the protection against chronic intracellular infections and cancer, depends on the concordant activity of CD8⁺ T cells (CTL) and CD4⁺ T cells, especially Th1 cells[19]. However, the activation of T cells and NK cells increases the production of the IL-12 receptor (beta1and beta 2), which explains why IL-12 can direct the proliferation and activation of T lymphocytes, NK cells, and NKT cells and can induce both IFN-γ and increased cytotoxic activity[25]. Since, IL-12 enhances the cytolytic activity of cytotoxic T and NK cells, and induces IFN-γ production from T and NK cells[2,29] therefore, the importance of IL-12 will be considered.

This open-label study was designed to evaluate the safety and immunological effects of the CNI-1493 on the in vivo activity of CTLs, by using AdOVA immunized C57Bl/6 mice.

From the data obtained so far, the CTLs significantly reduced the number of target cells in the presence of CNI-1493 (P < 0.05). On the other hand, there were no statistical differences in the cytolytic activity of CTLs in both groups (present and absent of CNI-1493). These outcomes are in agreement with a large body of data in multiple animal models, which suggests that CNI-1493 can be useful to prevent the macrophage mediated sequelae of disease and trauma, including sepsis, cerebral ischemia, and endotoxemia through inhibition of proinflammatory cytokine responses[6,8,11,15,21,25]. Hence, this study discovered more safety-related properties of CNI-1493 and improvement in the functioning of CTLs. Additionally, this study provides insights that may be of value to protect from IL-12 toxicity and may prove useful as a research tool to elucidate the distinct mechanisms involved in IL-2 host toxicity, which has not been utilized yet. However, additional studies on the interaction between CNI-1493 and IL-12 are needed to provide better understanding and to clarify the relationship between them.

CONCLUSION:
It can be concluded that CNI-1493 is a compound which is safe and undamaging to the immunological function of the cytotoxic T lymphocyte cells.

REFERENCES:
inflammation and endotoxin lethality.

Bianchi, M., Ulrich, P., Bloom, O., Meistrell, M., Zimmerman, G.

Cerami A, Tracey KJ (1996) CNI-1493 inhibits 9-


Cohen, P. S., Schmidtmayerova, H., Dennis, J., Dubrovsky, L.,

Sherry, B., Schmidtmayerova, H., Calandra, T., Zhang, X.,


Bjork, L., K. J. Tracey, P. Ulrich, M. Bianchi, P. S. Cohen, K.


Cohen, P. S., Schmidtmaierova, H., Dennis, J., Dubrovsky, L.,


9-Cohen PS, Nakashiri H, Dennis J, Caragine T, Bianchi M,


Furumoto, K., S. Arii, S. Yamasaki, M. Mizumoto, A. Mori, N.


Hauber I, Bevec J, Heukeshoven J, Kratzer F, Hom F, Choidas A,


Kaiser A, Hammels I, Gottwald A, Nassar M, Zaghloul MS,


Miller, G., S. Lahrs, V. G. Pillaretty, A. B. Shah and R. P.


Orange Jordan S, S. Salazar-Mather Thais P, Opal Steven M,


Specht Sabina, Sarite Salem Ramadan, Hauber Ilona Ilona,

Hauber Joachim, Gorbig Ulf F, Meier Chris, Bevec Dorian,


Tripp CS, Wolf SF, Unanue ER: Interleukin 12 and tumor necrosis factor a are costimulators of interferon g production by natural killer cells in severe combined immunodeficiency mice with listeriosis, and interleukin 10 is a physiological antagonist. Proc Nat Acad Sci USA 90:3725, 1993.


Wingender, G., N. Garbi, B. Schumak, F. Jungerkes, E. Endl, D.


Heit, K. M. Huster, O. Takikawa, S. Akira, D. H. Busch, H.


Molecular Modeling of Interactions of Zwitterionic Histidine Forms with Cu\(^+\) and Cu\(^{2+}\)

Azadeh Manzarzadeh\(^1\), Figen Zihnioğlu\(^1,2\), Cenk Selçuki\(^1,2\)

Abstract—Present study investigates the Cu\(^+\) and Cu\(^{2+}\) complexes of zwitterionic histidine (His) forms by molecular modeling. Calculations have been carried out by Density Functional Theory (DFT) at B3LYP/cc-PVTZ level.

INTRODUCTION

The interactions between histidine and metal ions play essential roles in a wide range of important biological processes including enzyme catalysis and signal transduction. This interest was improved by the discovery of the antibacterial, antifungal and anticancer activity of several metal complexes with histidine ligands which are also found to be highly efficient. Study of the interactions between drugs and transition metals is an important and active research area in bioinorganic chemistry [1-3]. It is well known that metal ions might play a vital role during the biological process of drug utilization in the body.

METHODS

The initial geometries from conformational analysis were re-optimized at B3LYP/cc-PVTZ level. After the geometry optimization step, B3LYP/cc-PVTZ level optimizations were carried out on the previously obtained geometries which were within 5 kcal/mol energy difference relative to the global minimum. Most stable geometries were than selected to interact with Cu\(^+\) and Cu\(^{2+}\) ions at B3LYP/GEN level where GEN basis set was defined as cc-PVTZ for C, O, N, H and LANL2DZ for copper. Initial calculations have been performed in vacuum (gas phase) and the calculations have been repeated in solution using the Polarizable Continuum Model (PCM) mimicking water as the implicit solvent.

RESULTS

<table>
<thead>
<tr>
<th>Conformer</th>
<th>Energy (Hartree)</th>
<th>ZPE (Hartree)</th>
<th>µ (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>006</td>
<td>-745.026</td>
<td>0.163</td>
<td>14.63</td>
</tr>
<tr>
<td>019</td>
<td>-745.015</td>
<td>0.163</td>
<td>12.02</td>
</tr>
<tr>
<td>013</td>
<td>-745.015</td>
<td>0.164</td>
<td>12.02</td>
</tr>
<tr>
<td>002</td>
<td>-745.013</td>
<td>0.164</td>
<td>13.77</td>
</tr>
<tr>
<td>011</td>
<td>-745.013</td>
<td>0.163</td>
<td>8.92</td>
</tr>
</tbody>
</table>

CONCLUSIONS

Metal ion complexes of zwitterionic His forms are more stable than the complexes of neutral His forms.

Both Cu\(^+\) and Cu\(^{2+}\) ion show a preference for specific binding sites.

Our results indicate that Cu\(^+\) can form complexes more easily and they are more stable compared to the Cu\(^{2+}\) complexes when formed.

It is also observed that Cu\(^{2+}\) ions break the bonds in histidine and forms insertion complexes.

ACKNOWLEDGMENT

Some of the calculations are performed on TUBITAK-ULAKBIM (Truba Resources).

REFERENCES

[3] Si-Ming Liao, Qi-Shi Du, Jian-Zong Meng, Zong-Wen Pang and Ri-Bo Huang, Chemistry Central Journal 2013, 7, 44

\(^*\)Corresponding Author (e-mail: cenk.selcuki@ege.edu.tr).

\(^1\)Azadeh Manzarzadeh, Ege University, Graduate School of Natural and Applied Sciences, Biochemistry 35100 Bornova, Izmir, Turkey, (e-mail: azade.manzarzade@gmail.com).

\(^2\)Figen Zihnioğlu and Cenk Selçuki, Ege University, Faculty of Science, Biochemistry Department (e-mails: figen.zihniojl@ege.edu.tr and cenk.selcuki@ege.edu.tr).
Abstract—N-Acetylglucosamine (N-acetyl-D-glucosamine, or GlcNAc, or NAG) is a monosaccharide derivative of glucose. It is an amide between glucosamine and acetic acid. It has a molecular formula of C\textsubscript{8}H\textsubscript{15}NO\textsubscript{6}, a molar mass of 221.21 g/mol, and it is significant in several biological systems. In this study, we carried out a series of conformational analysis and optimizations to investigate the behavior of N-Acetylglucosamine in vacuum and in water, by performing density functional calculations with 6-31G basis set.

INTRODUCTION

It is part of a biopolymer in the bacterial cell wall, built from alternating units of GlcNAc and N-acetylmuramic acid (MurNAc), cross-linked with oligopeptides at the lactic acid residue of MurNAc. This layered structure is called peptidoglycan.

GlcNAc plays several important roles at the cell surface of all organisms. One critical role is in the formation of the cell wall, the innermost layer of which is composed of chitin, a polymer of \(\beta\)-(1,4) linked GlcNAc. GlcNAc is also important for modification of cell-surface proteins. It forms part of the N-linked polysaccharide chain that is added to glycosylated proteins and is a building block in the synthesis of GPI-anchors that maintain certain proteins in the plasma membrane [1].

METHODS

In this study, we aimed to determine the most stable conformer both in vacuum and in water, in order to find out and compare the stability of N-Acetylglucosamine in biological systems. For conformational analysis Spartan08 was used. Density functional calculations at the b3lyp/6-31++g(d,p) level have been performed by using Gaussian 09 program.

RESULTS AND DISCUSSION

According to our preliminary results, the most stable conformer-con003, is stabilized by three H-bonds, between hydroxyl and carbonyl groups of N-Acetylglucosamine.

<table>
<thead>
<tr>
<th></th>
<th>E (Hartree)</th>
<th>(E_{\text{rel}}) (kcal/mol)</th>
<th>(\mu) (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>con003</td>
<td>-744.567</td>
<td>0.00</td>
<td>5.91</td>
</tr>
<tr>
<td>con032</td>
<td>-744.566</td>
<td>0.31</td>
<td>6.03</td>
</tr>
<tr>
<td>con012</td>
<td>-744.565</td>
<td>0.61</td>
<td>6.58</td>
</tr>
<tr>
<td>con011</td>
<td>-744.565</td>
<td>1.03</td>
<td>3.59</td>
</tr>
<tr>
<td>con027</td>
<td>-744.562</td>
<td>2.35</td>
<td>5.43</td>
</tr>
</tbody>
</table>

CONCLUSIONS

After the conformational analysis, we obtained 92 initial structures. Our results indicate that three conformers have exactly the same relative energy and dipole values. First four conformers have slightly different relative energies but after the fifth conformer, the difference between energies became significantly higher.

FUTURE REMARKS

Due to its fundamental importance in cell-cell interaction, we plan to investigate the interactions between the binding site of \textit{Phaseolus vulgaris} PHA-L lectin and N-Acetylglucosamine.

ACKNOWLEDGEMENTS

Some of the calculations are performed on TUBITAK-ULAKBIM (Truba Resources).

REFERENCES

M-Box Riboswitches: Modeling Perspective, Computational Study of Nucleobases and Their Interactions with Mono and Di Mg$^{2+}$ Cations

Ersin Gündeğer, Cenk Selçuki

Abstract — Computational investigation of nucleobases and their interactions with Mg$^{2+}$ have been carried out by using Density Functional Theory. Since it is important to understand the effects of tautomerism in catalytic RNAs, we especially focus on M-box Riboswitches. These calculations are the initial steps of understanding M-box Riboswitch mechanism. We also tried to understand the tautomerism in the presence of Mg$^{2+}$. We observed how the number of Mg$^{2+}$ cations influence canonical-noncanonical equilibria in nucleobases. These results will provide information on how noncanonical nucleobases are stabilized in Riboswitches, and the effect of Mg$^{2+}$ in this stability as well as the catalytical activity.

INTRODUCTION

As a result of developments in bioinformatics tools, scientists have discovered many conversed mRNA that potentially function as riboswitches [1, 2]. Some of riboswitches have unknown ligands that are called "orphan" riboswitches, until Dann et al. (2007), and M-box riboswitches belong to that group [3]. Interactions between tautomers of nucleic acid nucleobases and Mg$^{2+}$ have been investigated to understand the structural and functional properties of M-box riboswitches. Environmental factors including water, chemical modification, metallic cation interaction, irradiation, and excitation play an important role in tautomerization [4].

FURTHER INFORMATION

Presence of metal ions can affect the nucleobase’s electron distribution and consequently tautomeric equilibria [5]. Thus, metal ions play role in formation of rare tautomers. This topic, especially tautomeric equilibria of nucleobases in vacuum and in a microhydrated environment has been studied frequently alongside of nucleobase-metal ion interactions [5]. Nucleobases are highly polar, so metal ions and water can play important role for tautomers of nucleobases. The effects of metal ions on the stability and properties of the nucleobase tautomers have been observed by using molecular modeling tools.

CONCLUSIONS

It has been shown that rare tautomers of DNA and RNA nucleobases have different stabilities in the presence of metal ions. In addition, interactions between two Mg$^{2+}$ ions and nucleobases have been investigated. These calculations can reveal the tautomerism mechanism and the stability of tautomeric nucleotides. Our results at B3LYP/cc-PVTZ level indicate that these rare tautomers form more stable complexes with Mg$^{2+}$ and two Mg$^{2+}$ ions compared to canonical nucleobase-Mg$^{2+}$ complexes. Mono Mg$^{2+}$ calculations indicate that Mg$^{2+}$ prefers to form a bisdendate structure between O and N in all nucleobase-Mg$^{2+}$ interactions except for adenine (Figure 1). Also, results show presence of Mg$^{2+}$ affect tautomerization; for instance, in vacuum rare tautomers-Mg$^{2+}$ interactions are more stable for A1 (dominant tautomer of adenine). A8 is rare tautomer form which is the most seen structure in adenine-cytosine base pair and in effective point mutations [6, 7]. Di Mg$^{2+}$ calculations present that when second Mg$^{2+}$ added, A9 ($E_{rel}=23.66$ kcal/mol) with Mg$^{2+}$ structure ($E_{rel}=11.34$ kcal/mol), A9 is A8’s N9 hydrogen moves to N3 and imine form of adenine. Di Mg$^{2+}$ interaction of A9 has most stable interaction, second Mg$^{2+}$ interact with N9 and it has distance of 2.08 Å (Figure 2).

*Corresponding author (e-mail: cenk.selcuki@ege.edu.tr)
Ersin Gündeğer Ege University, Graduate School of Natural and Applied Sciences, Biotechnology Programme, Bornova, 35100 Izmir, Turkey (egundeger@gmail.com)
Cenk Selçuki Ege University, Faculty of Science, Biochemistry Department, Bornova, 35100 Izmir, Turkey (cenk.selcuki@ege.edu.tr)
Figure 2. A_N1N10-N9_7-4 most stable complex optimized at the B3LYP/cc-PVTZ level.

FUTURE REMARKS

Our preliminary results give useful information for understanding the underlying mechanism and principles for riboswitch properties. This will help manipulation of metabolite binding in different areas [8]. For example, rational design of antibiotics is significant for targeting riboswitches [9]. Besides understanding of regulatory and structural principles, understanding the properties of riboswitches will help designing of synthetic riboswitches that respond to certain ligand [10] and this may have a potential use in many field applications like seeking and destroying an unwanted herbicide in agriculture [11].

ACKNOWLEDGMENT

Some of the calculations have been done using TUBITAK-ULAKBIM (Truba Resources).

REFERENCES


AMPEROMETRIC BIOSENSOR FOR EPINEPHRINE BASED ON LACCASE IMMOBILIZED ON POLYPYRROLE-POLYVINYLSULPHONATE FILM
Fatma Arslan 1, Selma Durmuş 2, Özlem Çolak 3, Halit Arslan 1
1 Department of Chemistry, Faculty of Sciences, Gazi University, Ankara, TÜRKİYE
2 Department of Chemistry, Institute of Sciences, Gazi University, Ankara, TÜRKİYE
3 Türkiye İlaç ve Tibbi Cihaz Kurumu, Ministry of Health, Ankara, TÜRKİYE

Epinephrine, more commonly known as adrenaline [1-(3,4-dihydroxyphenyl)-2-methyloamino-ethanol] is one of the most important neurotransmitters in mammalian central nervous systems and exists in the nervous tissue and body fluid in the form of large organic cations. Epinephrine has also been used in medicine to treat heart attacks and bronchial asthma and in cardiac surgery. Therefore, the analysis of epinephrine in biological fluids is of great importance in medical diagnosis, particularly for patients suffering from Parkinson’s disease.

In this study, an amperometric epinephrine biosensor with immobilization of laccase onto polypyrrole/polyvinylsulphonate composite film was accomplished on the surface of a platinum electrode. The determination of epinephrine was carried out by the reduction of enzymatically generated epinephrinequinone at -0.220 V vs. Ag/AgCl. The optimum working conditions of biosensor with respect to the substrate concentration, pH and temperature were investigated. The storage and operational stability of the biosensor were also studied.

Reaction scheme for epinephrine determination

REFERENCES
Computational Investigation of Interaction of Gluconate Derivatives with Magnetite (Fe$_3$O$_4$) Surface in α-D-glucose coated iron oxide nanoparticles

Armağan KINAL$^a$, Işilay ÖZTÜRK$^a$, Güliz AK$^b$, Habibe YILMAZ$^b$, Şenay ŞANLIER$^b$

$^a$Ege University Science Faculty Department of Chemistry, İzmir, Türkiye
$^b$Ege University Science Faculty Department of Biochemistry, İzmir, Türkiye

Pharmacologically active anticancer drugs reach to tumor tissue with low specificity and they have dose-dependent toxicity in chemotherapy applications. Classical drug administration routes are oral and intravenous. Orally taken tablets or capsules result in irregular pharmacokinetic due to the metabolic pathways and they frequently damage healthy tissues due to low specificity. Nowadays, after the developments in nanotechnology, it is possible reduce these harmful side effects by using nano particular drug delivery systems [1,2]. When α-D-glucose coated iron oxide (magnetite) nanoparticles are loaded with anticancer drug (e.g. doxorubicin), this drug loaded iron oxide nano particles can be directed to tumor tissues via an external magnetic field by mostly eliminating the side effects of classical oral treatment [3-5].

In the present study, we have computationally investigated the thermodynamics and characteristics of the interactions between gluconate derivatives and Fe$_3$O$_4$ surface in α-D-glucose coated iron oxide nanoparticles. Since this system is enormously large for quantum chemical calculations, we used [Fe$_{10}$O$_{12}$]$^{2+}$ surface as the model system to mimic the characteristics of glucose coated iron oxide surface. Approach of several gluconate derivatives to the model [Fe$_{10}$O$_{12}$]$^{2+}$ surface were computationally studied by employing semi empirical PM6 method. The results of PM6 calculations indicated that gluconate and its derivatives spontaneously bind to the surface and they form chemical bonds after exothermic reactions. Gluconate and its derivatives have a tri-dentate bonding character with the iron oxide surface (Fig.1).

REFERENCES

Aydan Yılmaz, Şeyda Çiğdem Özkan, Emine Arslan

*Department of Chemistry, Faculty of Science, Selcuk University, 42031 Konya, Turkey
bDepartment of Biology, Faculty of Science, Selcuk University, 42031 Konya, Turkey
aydan@selcuk.edu.tr

The family of calixarenes plays an important role in host-guest chemistry. The spectacular development of calix[4]arenes as molecular receptors is related to many possible structures (1). They are commonly used in many fields but, very few studies relate to the therapeutical activities of calixarene compounds. Some of these are antimicrobical agents, cytotoxicity, DNA binding, DNA interaction, protection from ultraviolet radiation, anticancer agents, anti-HIV agents, enzyme and protein inhibitors, protein recognition studies. Although many of the metal-complexes of calixarenes were synthesized until now, they have little field of application in DNA studies (2). Generally, metal ions and complexes have great potential in biological studies and applications. Copper is an especially important element for humans with its bioessential activity (3). Cu(II) complexes have great capacity to DNA cleavage or interaction (4). Considering the importance of copper(II) complexes in bacteriology and DNA cleavage studies, we report herein the synthesis and characterization of two new copper(II) complexes of p-tert-butylcalix[4]arenes with amide derivatives. DNA cleavage activities of ligands and copper(II) complexes were also discussed. In addition, the determination of the site of DNA cleavage with BamHI and HindIII restriction enzymes was carried out for all compounds.

REFERENCES

MUTATIONS IN MU PHAGE PROTEINS LEAD LOSS IN LIPASE ACTIVITY

Bilge H. CADIRCI1*, Ihsan YASA2
1Gaziosmanpasa University, Tokat, Turkey
2Ege University, Izmir, Turkey

Bacteriophage Mu is a temperate phage that can infect several species of enteric bacteria. The mor gene product is required for middle transcription, and it is a product of the early transcription. In the present study in order to screen its effect on lipase gene regulation we have mutated the P. fluorescens P21 chromosomal DNA with mini-Tn5 transposon by triparental mating. Lipase negative mutants were selected by Rhodamine B agar method. When characterizing the target and flanking sequences of chromosomally inserted mini-Tn5, an arbitrary PCR was used. The most intense bands after electrophoresis were isolated and sequenced. Sequencing was done on an ABI PRISM 310 automated sequencer.

Open reading frames (ORFs) in the surroundings of the target were easily identified by ORF finder and those corresponding to of which protein was detected by the BLASTX program at the National Center for Biotechnology Information (NCBI). 20 of 3000 mutant colonies were lipase negative. And in some colonies, we got the same mutated Mu phage protein which let us to conclude its high relation to lipase.

*Presenting author.
Address: Gaziosmanpasa Universitesi Muhendislik ve Doga Bilimleri Fakultesi Biyomuhendislik Bolumu, Tokat/Turkey
E-mail: bilgehilal.cadirci@gop.edu.tr
Phone: +90 505 797 56 21
Fax: +90 356 252 15 85

Keywords: Lipase, Mu phage, transposon mutation

This work was supported by a grant from The Scientific and Technological Research Council of Turkey (TUBITAK).
ELECTROCHEMICAL DETERMINATION OF GLUCOSE BASED ON THE İMMOBİLİZATION OF GLUCOSE OXİDASE ONTO POLY(2,3-DI(THIOPHENE-2-YL) NAPHTHALENE -1,4-DIONE) MODIFIED GLASSY CARBON ELECTRODE

M. GULER1, V. TURKOGLU1*, M.R. KIVANC1, A. KIVRAK1*, Z. BASI1, N. TURKOGLU2

zehrabasi@hotmail.com  vedatturkoglu@yyu.edu.tr

1 Yuzuncu Yıl University Faculty of Science Department of Chemistry, Van, Turkey.
2 Yuzuncu Yıl University Faculty of Agriculture, Department of Horticulture, Van, Turkey.

In the study, 2,3-di(thiophene-2-yl) naphthalene -1,4-dione (TNQT) was synthesized. For glucose determination, TNQT was electrochemically polymerized on glassy carbon electrode (GCE) using cyclic voltammetry (CV) technique and glucose oxidase was immobilized on poly(2,3-di(thiophene-2-yl) naphthalene -1,4-dione) modified GCE (poly(TNQT)/GOx/GCE). The biosensing properties toward glucose of the poly(TNQT)/GOx/GCE were carefully studied. Poly(TNQT)/GOx/GCE showed a linear range from 0.6 to 4.58 mM of glucose concentrations and a detection limit of 0.01 mM. The apparent Km value was estimated as 2.13 mM from Lineweaver-Burk graph. The biosensor also displayed good reproducibility and long-term stability.

Key words: Glucose Oxidase, 2,3-di(thiophene-2-yl) naphthalene -1,4-dione, Glassy Carbon Electrode, Cyclic Voltammetry.
PRODUCTION OF POLYVINYL ALCOHOL/GELATIN BASED CYROGEL SCAFFOLDS FOR TISSUE ENGINEERING APPLICATIONS

Seda Ceylan¹,², Nimet Bölgen*²

¹Adana Science and Technology University, Bioengineering Department, Yeşiloba Campus, 46278, Adana, Turkey
²Mersin University, Engineering Faculty, Chemical Engineering Department, 33343, Mersin, Turkey

PVA Polyvinyl Alcohol (PVA) is a hydrophilic, biodegradable and biocompatible synthetic polymer. Recently, PVA hydrogels became very attractive as tissue engineering scaffolds. PVA has been used many different areas for tissue engineering applications such as synthetic heart valves, corneal implants, and cartilage tissue substitutes. Different production techniques can be used to have PVA scaffolds with adequate properties for tissue regeneration. In previous studies related to tissue engineering, PVA has been used in hydrogel form. PVA can be used not only in hydrogel form but also cryogel form can be used in tissue engineering applications. Gelatin is the denatured form of collagen which is the main component of the extracellular matrix of the tissues. In this study, PVA/Gelatin cryogels were crosslinked with freezing–thawing method. SEM images revealed that, with increasing gelatin concentration the average pore diameters increases. At the same time, the degradation rate increases with increasing gelatin concentration. The degradation rate of the produced biomaterials for PVA/Gelatin 100:0 and 50:50 scaffolds, was measured as %8.04 and %51.65, respectively. In addition, Fourier Transform Infrared Spectroscopy (FT-IR) was used to characterize chemical groups in cryogels. Results indicated that, the PVA/Gelatin concentration effects the architecture and characteristic properties of the scaffolds. PVA/Gelatin cryogels with their interconnected porosity, swelling property, biodegradability and mechanically stable structure, have potential to be used scaffolds for tissue engineering applications.

Keywords: Polyvinyl alcohol, Gelatin, Cryogel, Scaffolds, Tissue Engineering

*Correspondance to: Nimet Bölgen, Mersin University, Engineering Faculty, Chemical Engineering Department, Mersin, Turkey. E-mail: nimetbolgen@yahoo.com
Dipolar Ferrocene and Ruthenocene Second-Order Nonlinear Optical Chromophores: A Time-Dependent Density Functional Theory Investigation of Their Absorption Spectra

Seyhan Salman,1,2 Jean-Luc Brédas,2 Seth R. Marder,2 Veaceslav Coropceanu,2 and Stephen Barlow2

1School of Chemistry and Biochemistry and Center for Organic Photonics and Electronics, Georgia Institute of Technology, Atlanta, Georgia 30332-0400, United States
2Genetics and Bioengineering, Istanbul Bilgi University, Eyup 34060, Istanbul, Turkey

The origin of the two prominent solvatochromic near-UV/visible/near-IR absorptions observed for donor–(π-bridge)–acceptor chromophores with ferrocene donors has been investigated using TD-DFT methods. Both chromophores with relatively weak (4-nitrophenyl) and strong acceptors (1,3-diethyl-2-thiobarbituric acid and 3-dicyanomethylene-2,3-dihydrobenzothiophene-1,1-dioxide) were considered, as were ferrocene and octamethylferrocene donors. Computational predictions of optical properties made using the B3PW91 functional were found to be in good agreement with experimental data. The calculations reveal a complex orbital picture that varies from compound to compound, contribution of multiple configurations to some of the important states, and significant contributions from more than one transition to the experimentally observed bands. Natural transition orbitals have been used to gain an understanding of the charge redistribution associated with the transitions. The relatively weak low-energy bands of the ferrocene derivatives were generally found to have both d−d and metal-to-π-bridge/acceptor charge-transfer character. The stronger higher energy bands were found to be associated with charge transfer from cyclopentadienyl rings and the π bridge toward the acceptor group. The experimental spectra of ruthenocene chromophores differ significantly from those of the analogous ferrocene chromophores; however, the calculations reproduce the key differences and indicate a similar origin for the contributing transitions.

Seyda Çiğdem Özkan, Aydan Yılmaz, İsmail Özmen

*aDepartment of Chemistry, Faculty of Science, Selçuk University, 42031 Konya, Turkey
bDepartment of Chemistry, Faculty of Science and Arts, Suleyman Demirel University, Isparta, Turkey
seydaozkan87@gmail.com

Calix[n]arenes can be viewed as examples of [1n]metacyclophanes that possess basket-shaped cavities are composed of phenolic units ortho-linked by methylene bridges (1). Calixarenes can be used in many fields for a variety of purposes and, the biological properties of calixarenes, i.e., antimicrobial agents, antiseptic and anticancer agents, cytotoxicity, DNA binding, DNA interaction recognition, and protection from ultraviolet radiation have recently been investigated. Recently, amido-calixarenes have been used as nucleic acid recognizers (2), enzyme inhibitors (3), potent-DNA binding agents (4). In this study, we prepared six new amide derivatives (mono- and di-) of calix[4]arenes that have some aliphatic and aromatic groups, then investigated the activities of these compounds on the DNA. The interaction between these compounds and pBR322 plasmid DNA has been investigated via agarose gel electrophoresis.

Figure 1. The electrophoresis image of pBR322 DNA after treated with the compounds (P: Untreated plasmid DNA with compounds)

REFERENCES

SNP Prioritization is an important step in Genome Wide Association Studies (GWAS). In prioritization, the SNPs which have some degree of statistical meaning according to their p-values, are selected and subjected to scorings with respect to their associations with genes, pathways, diseases and locations such as intron or exon. METU SNP software was developed in our laboratory in 2011 for SNP prioritization in GWAS analysis. It employs a novel technique for SNP scoring called Analytical Hierarchy Process (AHP) decision making technique for SNP prioritization. It analyzes and calculates score for each SNP according to 56 different features in the dataset. However the weights of these scores are dependent on subjective opinions of molecular biology experts. Moreover the necessities of all of the inspected features in the tree are not clearly known which results in unnecessary calculations in analysis. Since 2011 we have examined GWAS data on several different diseases such as Alzheimer’s disease, prostate cancer, melanoma, rheumatoid arthritis and diabetes. In these studies various data-mining approaches are used on disease specific case-control data to identify SNPs associated with these diseases and to increase our knowledge about SNPs. Recently we are questioning the possibility of adjusting weights and features of AHP objectively by approaching AHP prioritization parameters as a data mining problem. Here we will present our current methodology for AHP optimization, and introduce new databases, such as Regulome DB and GWAS central that will be integrated into the AHP tree for SNP prioritization.
INVESTIGATION OF IN VITRO EFFECT OF Alkanna tinctoria ROOT ORGANIC PHASE EXTRACTS ON HUMAN ERYTROCYTE 6-PHOSPHOGLUONATE DEHYDROGENOSE, GLUCOSE-6-PHOSPHATE DEHYDROGENOSE AND GLUTATHIONE REDUCTASE ENZYMES

M. CAKAR¹, V.TURKOGLU¹, Z. BASI¹, M.GULER¹, M.R. KIVANC¹
¹Yuzuncu Yil University, Faculty of Science, Department of Chemistry, Van, Turkey.

In this study, The effect of Alkanna tinctoria root organic phase extracts on human erytrocyte 6-phosphogluconate dehydrogenase, glucose-6-phosphate dehydrogenase and glutathione reductase enzymes was investigated. Enzyme activity was measured using spectrophotometer method according to Beutler method. Lineweaver-burk graph was drawn for each enzyme. For this purpose, at three different substrates and five different inhibitors (organic phase of plant) concentrations were studied. Also control group graphs were constructed. This process was performed for each enzyme. Inhibition type and \( I_{50} \) values were found from inhibition graphs. Organic phase extract concentrations of plant, 6-phosphogluconate concentrations being the substrate of 6-phosphogluconate dehydrogenase, glucose-6-phosphate concentrations being the substrate of glucose-6-phosphate dehydrogenase, and oxidized glutathione concentrations being the substrate of glutathione reductase were selected among 0.0053-0.021 mg/ml, 0.15-0.6 mM, 0.36-0.72 mM, and 0.2-1 mM respectively. These concentrations were determined as the most suitable values by means of the first procedures. Organic phase extracts of Alkanna tinctoria root inhibited glutathione reductase while they activated 6-phosphoglukonate dehydrogenase and glucose-6-phosphate dehydrogenase. The plant extracts shown non-competitive inhibition effect on glutathione reductase enzyme activity and the \( I_{50} \) value was calculated as 0.0249 mg/ml.

**Key words:** Glutathione Reductase, Glucose-6-Phosphate Dehydrogenase, 6-Phosphoglukonate Dehydrogenase, Alkanna tinctoria, Human Erytrocyte, Enzyme Inhibition, Enzyme Activation.
Functional Proteomics in Lipid Researchs
Funda KARTAL

Ege University, Faculty of Science, Biochemistry Department, 35100 Bornova/İZMİR-TÜRKİYE
funda_kartal@hotmail.com

Lipids serve a great variety of functions. Lipids are vital components of many biological processes and serve as building blocks of biological membranes (e.g., phospholipids, sphingolipids) or of specific proteins (e.g., myristollation, palmitoylation), as substrate for metabolic energy production (long chain fatty acids and fatty acids) and as signaling compounds (long chain fatty acids and fatty acid metabolites). Their digestion, transport, formation, intracellular storage and mobilization are tightly controlled process to ensure overall energy balance.

The regulation of lipolytic activity is crucial to ensure energy balance and to avoid lipotoxic effects as exerted by products of lipolysis such as free fatty acids. Therefore, transcriptional and post-translational regulation are involved in the equilibrium between lipid storage and mobilization. Moreover, substrate accessibility, which may be governed by localization of lipases, is a key regulatory element. Lipid transport as well as lipid synthesis, storage and mobilization is facilitated by complex lipid-protein assemblies, namely lipoproteins and lipid droplets. Proteomic studies of these particles have uncovered protein factors involved in their assembly and transport, but also in lipid synthesis and mobilization.

The increase in the number of genome sequencing projects, there is a concomitant exponential growth in the number of protein sequences whose function is still unknown. Functional proteomics constitutes an emerging research area in the proteomic field whose approaches are addressed towards two major targets: the elucidation of the biological function of unknown proteins and the definition of cellular mechanisms at the molecular level. And based on interactions of proteins with other proteins or small molecules.

Protein–protein interaction techniques are used to gain information on protein localization, protein complex composition and post-transcriptional regulation. For this purpose, a bait protein is either immobilized to specifically bind proteins which interact with the bait, or protein complexes are immuno-precipitated with bait specific antibodies. Apart from tremendously decreasing the sample complexity, information on composition of protein networks can be gained this way. The isolated complexes can then either be studied in depth by MS methods or in a high throughput fashion employing antibody microarray techniques.

References
General Approach to Bioactive Peptide Production

H. Kübra Kani, Ebru K. Kocazorbaz
Ege University, Faculty of Science, Biochemistry Department, Bornova/Izmir, TURKIYE

e-mail: ebru.kocadag.kocazorbaz@ege.edu.tr

Many of the physiological and functional properties of proteins are believed to attribute to biologically active peptides encrypted in the protein molecules. These peptides play an important role human health. These peptides are inactive within the sequence of parent protein and can be released from their precursor proteins by \textit{in vivo} or \textit{in vitro} proteolytic processes with proteases. Proteolytic processing as well as other molecular modifications leads to a wide range of mature products which may vary in different tissues even though they are derived from the same precursor. In the present time, Bioactive peptides (endogenous/exogenous) have been defined as specific protein fragments that have a positive impact on body functions and conditions and more and more peptide-type compounds are introduced and applied in different areas of human health. According to their functional properties, bioactive peptides may be classified as antimicrobial, antioxidative antithrombotic, antihypertensive, opioid, immunomodulatory, mineral binding, anticancer, and etc. The activity of these peptides is dependent on their amino acid composition and sequence. A database named Biopep, summarizing bioactive peptides and their regulatory effects, is available, in which more than 1500 different bioactive peptides are included. Bioactive peptides generally contain 3–20 amino acid units, but in some cases this range may be extended. Variations in peptide yield, composition, and bioactivity exist among different methodologies and should be considered for the selection of one or combined methods for a certain purpose. Peptides can be produced \textit{in vitro} through enzymatic hydrolysis of proteins which is selective and specific, obtaining more predictable with respect to the end products. In addition to \textit{in vivo} and \textit{in vitro} proteolysis of protein precursors, production of bioactive peptides can also be achieved via molecular genetic engineering, in which bioactive peptides with specific amino acid sequences can be synthesized through gene cloning, over expression, and site-directed mutagenesis. The isolation and purification of bioactive peptides are very important for exploration of their physicochemical properties and evaluation of their in vitro and in vivo bioactivities. Bioactive peptides can be separated from a protein hydrolysate mixture by a number of approaches, mainly, different kinds of chromatography and membrane-based separation techniques. Prior to the separation process, a peptide mixture should be extracted in order to remove proteins, enzymes, and other components in the source material following enrichment and separation techniques. Different kinds of chromatography can be used in peptide separation, including high-performance liquid chromatography (HPLC), especially reversed-phase HPLC, ion-exchange chromatography (IEC), capillary electrophoresis (CE), capillary isoelectric focusing (CIEF) separation, gel filtration chromatography (GFC) and hydrophobic interaction chromatography (HIC). One or more of these techniques in combination can be selected based on the properties of the peptides. Recently, affinity chromatography appears as an effective purification method for bioactive peptides. Membrane-based separation processes are techniques in which a component is separated from a mixture as the latter is forced through a porous membrane under applied pressure. Peptide size and/or charge are important features for the retention mechanism of these methods. A wide variety of membranes are available to serve different fractionation purposes. Other than selecting one separation method, different techniques are normally used in combination to achieve best separation of bioactive peptides. Bioactive peptides, once isolated and purified, can be often subjected to mass spectrometry for analyzing the sequence of unknown peptides. Studies on peptides obtained from protein hydrolysates, have shown that these molecules have diverse functional properties offer a promising approach of their production in the food and drug industry. Nevertheless, there is a need for scaled-up production of these compounds.

References

Preparation of encapsulated therapeutic L-asparaginase nanoparticles

Didem Songurtekin, Duygu Asker, Taylan Kurtuluş Öztürk, Ali Kılınç
Ege University, İzmir, TURKEY

The specificities of enzymes and their having metabolic activities put forward successful results as therapeutic agents in treatment. With the usage of protein and enzymes, which are described as new generation medication, a chance of outstanding strategies for the treatment of cancer, diabetes, immune system and infection diseases has been provided now.

On the other hand designing nano sized immobilization strategies offer promising therapeutic potential for several metabolic enzymes.

Here we report a feasibility study for application potential of encapsulated therapeutic enzyme. By this way, which can be applied to various enzymes, provide critical advantage on carrying soluble enzyme nanoparticles to target area in active form and creating metabolic influence. For this aim recombinant E. coli L-asparagine amidohydrolase (EC 3.5.1.1; asparaginase) is produced and characterized. Active single enzyme nanoparticles¹ (SEN) were developed with the method of nanosized encapsulation in hydrophilic polyacrylamid network.

REFERENCES


*Corresponding author: ali.kilinc@ege.edu.tr
Nanoencapsulation of L-asparaginase in Biosilica Support

Taylan Kurtuluş Öztürk, Duygu Asker, Didem Songurtekin, Ali Kılınç
Ege University, Izmir, TURKEY

Proteins and enzymes, which are described as new generation medication today, provide chances for the improvement of considerable strategies in the treatment of several diseases\(^1\). However, several factors that limit clinical usage of enzymes and the other proteins are. In order to overcome these holdbacks, therapeutic proteins can be improved by protein engineering and immobilization methods.

Nano sized immobilization strategies of therapeutic enzymes with various material types have been developed. Besides the conventional strategies of immobilization, bio templated silica condensation, mimicked from diatoms as an immobilization method\(^2\), has great potential to design nano sized active enzyme nanoparticles.

In this study, transferring R5 oligopeptide to \textit{E. coli} L-asparagine amidohydrolase enzyme (EC 3.5.1.1; asparaginase), therapeutic activity is proven and actively used, enzyme has been produced as recombinant. After that, with trapping this enzyme into a biosilicananocage, active enzyme nanoparticles have been prepared in biosilica structure.

REFERENCES

Adaptive-Network Fuzzy System for the Prediction of Binding Affinity of Peptides

Volkan Uslan, Hanife Catman,

High dimensional, complex and non-linear nature of the post-genome data often adversely affects the performance of predictive models. Fuzzy system is a method that has been widely used to model such non-linear systems. Fuzzy system is good at modelling uncertainty and yielding a set of interpretable if-then rules. The main aim of this paper is to develop an adaptive-network fuzzy system in order to find a feasible solution for the prediction of binding affinity peptides. This problem is one of the difficult and complex modelling problems in bioinformatics due to the diversity of peptides discovered. Fuzzy clustering is used in the premise part of fuzzy system to approximate the membership functions that characterize each fuzzy set. The proposed model successfully applied to this problem and yielded promising results.

REFERENCES

Measurement of Refractive Index using Terahertz Reflectance Mode on Fixed Quartz Sample Holder
Martin Mueller-Holtz, Huseyin Seker and Geoff Smith*

Abstract— Measurements in Terahertz reflection mode can be difficult as a quartz sample holder is typically required to hold the sample in the focus of the incidence beam. In this paper, we present a methodology to account for misalignment errors and displacements between the quartz and a reference mirror by using the Fresnel equations in conjunction with a correction variable. De-ionised water has been chosen as validation data and shows a good agreement with the values presented in the literature.

INTRODUCTION

Measuring the optical properties in the Terahertz regime of biological samples has been of increased interest in recent years [1]. Due to the high absorbance of water in biological tissue and the sample thickness it is almost always necessary to measure the sample in reflection mode.

For the calculation of those properties, the simplified Fresnel equations are often used by neglecting the polarisation and incidence angle [1]. In this paper, we present a possible way of calculating the optical properties using the full Fresnel equations in conjunction with a correction factor. The work carried out is based on the works from Jepsen et.al. [2] but with a different initial situation.

In order to calculate the frequency dependent, complex refractive index from a time domain measurement in the Terahertz regime in reflectance mode, typically a reference mirror, a baseline (air) is needed in addition to the sample data. In practice, it is not always a simple task to measure those values [3] correctly as small differences, like shifts between measurements in the size of some microns, lead to a phase change in a frequency domain, which causes an unstable calculation of the refractive index.

THEORY

A. General Fresnel equations for electromagnetic waves
For every interface of two materials (j, j+1) with different refractive indices, the general Fresnel equations use the parallel and perpendicular parts (superposition) of the reflection (r) and the transmission (t). Considering an equal magnetic permeability (or permeability of 1), the four Fresnel equations are defined as:

\[ r^\parallel_j = \hat{n}_j \cos(\theta_j) - \hat{n}_{j+1} \cos(\theta_{j+1}) \]
\[ r^\perp_j = \hat{n}_j \cos(\theta_j) + \hat{n}_{j+1} \cos(\theta_{j+1}) \]
\[ t^\parallel_{j+1} = \frac{2 \hat{n}_j \cos(\theta_j)}{\hat{n}_j \cos(\theta_j) + \hat{n}_{j+1} \cos(\theta_{j+1})} \]
\[ t^\perp_{j+1} = \frac{2 \hat{n}_j \cos(\theta_j)}{\hat{n}_j \cos(\theta_j) + \hat{n}_{j+1} \cos(\theta_{j+1})} \]

where \( r^\parallel \) and \( r^\perp \) are the reflected wave from an interface \( j, j+1 \) and \( t^\parallel, t^\perp \) are the transmitted wave through the medium \( j+1 \) [4] as schematically shown in Fig.1. Both media have a (complex) refractive index (n) and an incidence angle (\( \theta \)).

This project receives funding from EMDA and the Hope Against Cancer charity (Leicester, UK). The research was approved by the NHS Research Ethics Committee

M. Mueller-Holtz and G. Smith are with The Pharmaceutical Technologies Research Group, the Faculty of Health and Life Sciences, De Montfort University, Leicester, UK (e-mail: mmm@iee.org and gsmith02@dmu.ac.uk)

H. Seker is with The Bio-Health Informatics Research Group, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK (e-mail: hseker@dmu.ac.uk)

* Corresponding author: G. Smith (e-mail: gsmith02@dmu.ac.uk)
Since typically \(n_{j+1}\) and \(\theta_j+1\) are unknown Snell's law and trigonometric functions can be used to eliminate \(\cos(\theta_j+1)\) leaving only \(n_{j+1}\) as the unknown.

\[
\cos(\theta_{j+1}) = \frac{\sqrt{\hat{n}_{j+1}^2 - \hat{n}_j^2 \sin^2(\theta_j)}}{\hat{n}_{j+1}} \tag{5}
\]

With the insertion of Eq.5 in Eqs.1-4, the second angle is eliminated.

The "outcome" of electro-magnetically wave propagating through mediums with different refractive indices (and absorption coefficient) typically requires knowledge of the original emitted wave to create the ratio of the amount reflected or transmitted with the emitted value. For a transmitting wave travelling through a medium the equation becomes

\[
E_{\text{emitted}} = \frac{E_{\text{transmitted}}}{T} \tag{6}
\]

and for the reflected wave

\[
E_{\text{emitted}} = \frac{E_{\text{reflected}}}{R} \tag{7}
\]

where \(T\) and \(R\) are transfer functions constructed from the Fresnel reflection and transmission coefficients.

B. Transmission through sample holder

In cases, where a sample holder is fixed or permanently mounted, the mirror reference is taken in contact with the sample holder. Most frequently, a quartz sample holder is used, as it shows a constant refractive index and a negligible absorbance [5] but changes the angle of 30°. Since the THz module redirects the beam to the sample with an incidence angle of 30°. Hence, the reflectance factor is needed as we cannot consider the beam produced by the machine to be completely unpolarised, consisting of equal parts of the parallel and perpendicular beam.

\[
E_{\text{air, meas}} = E_{\text{air, calc}} \cdot \text{cf} \tag{11}
\]

This is valid as long as the measured baseline is in between the theoretical calculated parallel and perpendicular parts of the reflected energy waveform.

Re-inserting the calculated baseline from Eq.10 into Eq.7, the calculation of an unknown sample is then

\[
E_{\text{sample}} = \frac{t_{12}^1 t_{12}^1 \hat{r}_{31}^1}{t_{12}^1 t_{12}^1 t_{23}^1 t_{23}^1 t_{23}^1} + \frac{t_{12}^1 t_{12}^1 \hat{r}_{31}^1}{t_{12}^1 t_{12}^1 t_{23}^1 t_{23}^1 t_{23}^1} E_{\text{air, calc}} \cdot \text{cf} \tag{12}
\]

Inserting Eq.1 to Eq.4, Eq.11 is solved in regards to the unknown (complex) value \(n_{\text{sample}}\) numerically.

**Methodology and Setup**

A. Terahertz system setup

Terahertz waves were generated using a Teraview TPS spectra 3000™ in combination with the Imaga™ module provided by Teraview, Cambridge, UK. Terahertz waves are generated by firing a femto-second-laser beam on a superconductive dipole switch (HTGaAs). For focusing reasons, directly behind that switch is a highly resistive hyperhemispherical silicon lens [6]. By using a beamsplitter, the beam divides into a reference and a sample beam. The reference beam is used for the detection of the THz wave by using an optical gated switch, in this case a ZnTe diode.

For the purpose of measuring biological tissue, it is almost always necessary to use terahertz in reflectance mode rather than transmission mode [3]. The reflectance module redirects the beam to the sample with an incidence angle of 30°. Since the THz-signal is very sensitive to the exact position of the sample, a SiO² quartz sample holder with a refractive index of 1.95 is used. The SiO² sample holder has a negligible absorbance [5] but changes the incidence beam angle towards the sample to 14.5°.

To avoid artefacts from the beam travelling through air under the sample holder, the underlying chamber has been purged with nitrogen.
B. Measurements

A SiO²-quartz sample holder has been purchased from ISP-optics, with a constant refractive index of 1.95 measured in transmission mode (data not shown in this study) and negligible absorbance. Measurements of the baseline, reference and de-ionized water at room temperature have been carried out on four consecutive days.

RESULTS

A. Baseline measurements and correction factor

By correcting the calculated baseline from a mirror measurement, the resulting energy waveform shows the same maximum and minimum energy values but also a displacement on the optical delay as shown in Figure 2. Table 1 shows the maximum values of the measured and simulated.

It can be assumed that the optical delay displacement is due to an air gap between the mirror and the sample holder introduced by surface roughness of the solid materials.

By transforming the time domain signal into the frequency domain using the Fast Fourier Transform, the same equations are applied to the frequency domain data. Using the Fresnel equations, the calculated and simulated refractive index of air are shown in Fig.3. We expect the refractive index of air to be close to one.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>E_{max} BS meas</th>
<th>E_{max} BS sim</th>
<th>CF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air I</td>
<td>0.1843</td>
<td>0.1917</td>
<td>0.9614</td>
</tr>
<tr>
<td>Air II</td>
<td>0.1872</td>
<td>0.1967</td>
<td>0.9517</td>
</tr>
<tr>
<td>Air III</td>
<td>0.1917</td>
<td>0.1977</td>
<td>0.9700</td>
</tr>
<tr>
<td>Air IV</td>
<td>0.1794</td>
<td>0.1847</td>
<td>0.9714</td>
</tr>
</tbody>
</table>

B. Validating data of de-ionised water

Using the corrected baseline we investigate a more complex material, de-ionised water, which changes the index of refraction over time and has a high absorbance in the Terahertz regime [7]. Distilled water was chosen as validation of our methodology as we are working towards building an understanding of water in interaction with other materials. Fig.4 shows the frequency dependent refractive index of the sample measurements.

De-ionised water has been very well investigated in recent years, which allows us a good verification with the literature values obtained from [8-10].

CONCLUSIONS

This paper demonstrated the use of Fresnel equations to calculate an accumulated baseline. Simulated results of air were compared to the measured ones. It was found that the simulated baseline data improves the quality of the validation data, in this case de-ionised water, which corresponded well to the ones described in the literature.

Having developed a stable method for the measurement of frequency-dependent, complex refractive
indices allow us to investigate the optical properties of unknown samples. Furthermore, having the absorption and refractive index, the complex permittivity can be calculated, which will then expected to help us better understand the interaction of terahertz beams with matter.

REFERENCES

Most phenotypes show complex inheritance and continuous variation, which play a central role in evolution and susceptibility to disease. Although the technique of quantitative trait loci (QTL) has helped understanding the regulation of those traits, it has proven difficult to comprehensively determine their complex genetic architecture, partially because phenotype variation often results from the variation of multiple members of a pathway at diverse molecular levels. Using a combination of new proteomics methods and novel computational algorithms we investigated the impact of natural genetic variation on protein concentrations. To accomplish this task we generated an almost complete reference map of the budding yeast proteome for shotgun and targeted proteomics. We used this map in a series of shotgun- and targeted proteomics experiments in a panel of 78 budding yeast strains in order to identify protein-QTL, i.e. genomic regions associated with protein abundance changes. These experiments were informed by computational network analysis. Using a powerful new machine-learning approach we could identify a surprisingly large fraction of protein-QTL being in epistasis with each other.

The network-based analysis facilitated the identification of protein modules, whose members are affected by several independent genetic variants in a coordinated way. This suggests that selective pressure favors the acquisition of sets of polymorphisms that adapt protein abundances at the pathway level.

In a related project, we have generated the first recombinant strain library for fission yeast and conducted a RNA-seq-based expression QTL study, i.e. aiming to identify genomic regions affecting RNA levels of coding and non-coding genes. This data is characterized by a particularly high statistical power. We have developed methods that enable the genotyping of the segregants using only RNA-seq data, and an individualized quantification of gene expression preventing the detection of false cis-eQTL. Using these techniques we showed that non-coding genes are at least as much affected by eQTL as protein-coding genes. We have identified a genetic variant of swc5 that is affecting the deposition of the histone HA2.Z and modifying the levels of more than 1,000 transcripts. Additional analyses, including an antisense expression QTL study, indicated that this variant impacts antisense levels via read-through transcription. The strains, methods, and datasets generated in this project provide a rich resource for future studies.
International Conference on Applied Informatics for Health and Life Sciences
Turkish-German Workshop on Bioinformatics: Recent Developments from Health to Nanotechnology
Kuşadası-TÜRKİYE 19-22 October 2014

Microarray Gene Expression Data-Based Bioinformatics Method for the Diagnosis of Male Hypertension

Nazlı Ucunoglu¹, Arinze Akutekwe² and Turgay Isbir¹

¹ Department of Molecular Medicine, Faculty of Medicine, Yeditepe University, Istanbul, Turkey
(e-mails: naz_zli@hotmail.com; turgay.isbir@yeditepe.edu.tr)

² Bio-Health Informatics Research Group, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK
(email: aakutekwe@dmu.ac.uk)

Hypertension is a chronic medical condition that the blood pressure in the arteries is elevated. Hypertension can lead to damaged organs, as well as several illnesses, such as renal failure (kidney failure), aneurysm, heart failure, stroke, or heart attack [1, 2]. Therefore, it is important to understand the mechanisms of the disease and identify relevant molecular markers of the disease. Therefore, in this study, we used a hypertension data set that was collected from male subjects in Taiwan [3] to identify potential biomarkers. The data set includes 77 patients with non-medicated young-onset male hypertensive cases and 82 male controls. Expression level of 22184 genes was also measured for each subject. For the statistical assessment, ten different subsets of the samples, each of which includes more or less the same number of samples, were constructed for the male hypertension patient and control group. In order to analyse the data set and identify potential genomic biomarkers, two statistical methods were applied, namely, t-test and entropy feature selection methods using 2-fold and 5-fold cross validation methods. In order to validate of the selected genes, K-Nearest neighbour classifier method was utilised. Among these groups, 3 number of biomarkers set were chosen (1,3,9) for 4 tables (t-test; 2-fold and 5-fold; entropy; 2-fold and 5-fold). From these biomarker sets which has the highest accuracy which is the measurement used for the classifier assessment was analysed and taken to the best models for each sub-set table. Each sub-set tables were analysed with each other and we tried to find the most appropriate biomarker. The defined biomarker was searched within database in order to find relationship with the disease. Consequently, highly recurrent and highly accurate candidate genes can be further analysed. Further analysis (both database and wet study ) can be suggested for the highly recurrent genes like Hs. 683236 (null), Hs. 475902, 420541, 656129, 647705 and 657792 as they may be potential biomakers that could further help understand the mechanism of the disease and be used for early diagnosis of the disease.

REFERENCES


<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdulah, S S</td>
<td>171, 135</td>
</tr>
<tr>
<td>Abhussin, A</td>
<td>23</td>
</tr>
<tr>
<td>Adnan, D</td>
<td>111</td>
</tr>
<tr>
<td>Agarwal, D</td>
<td>23</td>
</tr>
<tr>
<td>Agirbasli, Z</td>
<td>107</td>
</tr>
<tr>
<td>Ak, G</td>
<td>187</td>
</tr>
<tr>
<td>Akalin, H</td>
<td>160</td>
</tr>
<tr>
<td>Akbarova, Y</td>
<td>160</td>
</tr>
<tr>
<td>Akdağ, M</td>
<td>132</td>
</tr>
<tr>
<td>Akkmet, M</td>
<td>119</td>
</tr>
<tr>
<td>Aknunc, O</td>
<td>65</td>
</tr>
<tr>
<td>Akütekwe, A</td>
<td>91, 206</td>
</tr>
<tr>
<td>Albayrak, Y</td>
<td>132</td>
</tr>
<tr>
<td>Alkan, C</td>
<td>13</td>
</tr>
<tr>
<td>Ali, A S</td>
<td>171</td>
</tr>
<tr>
<td>Alkaya, Ü</td>
<td>49</td>
</tr>
<tr>
<td>Altundağ, M</td>
<td>183</td>
</tr>
<tr>
<td>Alýunuk, H</td>
<td>53</td>
</tr>
<tr>
<td>Al-Zubaidy, M A</td>
<td>111</td>
</tr>
<tr>
<td>Arslan, E</td>
<td>188</td>
</tr>
<tr>
<td>Arslan, F</td>
<td>186</td>
</tr>
<tr>
<td>Arslan, H</td>
<td>186</td>
</tr>
<tr>
<td>Ayoubi, A</td>
<td>173</td>
</tr>
<tr>
<td>Balasubramaniyan, R</td>
<td>60</td>
</tr>
<tr>
<td>Ball, G</td>
<td>23</td>
</tr>
<tr>
<td>Bader, L</td>
<td>79, 83</td>
</tr>
<tr>
<td>Barlow, S</td>
<td>192</td>
</tr>
<tr>
<td>Başaran, S C</td>
<td>132</td>
</tr>
<tr>
<td>Basi, Z</td>
<td>190, 195</td>
</tr>
<tr>
<td>Batistiou, E</td>
<td>140</td>
</tr>
<tr>
<td>Bekpen, C</td>
<td>13</td>
</tr>
<tr>
<td>Beyer, A</td>
<td>205</td>
</tr>
<tr>
<td>Brídas, J L</td>
<td>192</td>
</tr>
<tr>
<td>Boz, M</td>
<td>160</td>
</tr>
<tr>
<td>Bölgen, N</td>
<td>191</td>
</tr>
<tr>
<td>Bugra-Bilge, K</td>
<td>13</td>
</tr>
<tr>
<td>Burakoğlu, D</td>
<td>149</td>
</tr>
<tr>
<td>Busch, M</td>
<td>19</td>
</tr>
<tr>
<td>Cakar, M</td>
<td>195</td>
</tr>
<tr>
<td>Cavas, L</td>
<td>53, 107</td>
</tr>
<tr>
<td>Ceylan, S</td>
<td>191</td>
</tr>
<tr>
<td>Ceylan, Z</td>
<td>162</td>
</tr>
<tr>
<td>Cinar, A</td>
<td>114</td>
</tr>
<tr>
<td>Coropceanu, V</td>
<td>192</td>
</tr>
<tr>
<td>Çadırcı, B H</td>
<td>189</td>
</tr>
</tbody>
</table>

207
<table>
<thead>
<tr>
<th>Name</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Özören, N</td>
<td>13, 190, 195</td>
</tr>
<tr>
<td>Özkan, Ş C</td>
<td>188, 193</td>
</tr>
<tr>
<td>Özkaya, Ö K</td>
<td>14, 206</td>
</tr>
<tr>
<td>Özkul, Y</td>
<td>160, 13</td>
</tr>
<tr>
<td>Özmen, İ</td>
<td>193, 160</td>
</tr>
<tr>
<td>Öztürk, İ</td>
<td>187, 200</td>
</tr>
<tr>
<td>Öztürk, T K</td>
<td>198, 199, 75</td>
</tr>
<tr>
<td>Özyurt, C</td>
<td>130, 189</td>
</tr>
<tr>
<td>Plach, M</td>
<td>19, 16</td>
</tr>
<tr>
<td>Polyzos, D</td>
<td>6, 132</td>
</tr>
<tr>
<td>Potsika, V T</td>
<td>6, 189</td>
</tr>
<tr>
<td>Protopappas, V C</td>
<td>6, 99</td>
</tr>
<tr>
<td>Riedel, G</td>
<td>1, 103</td>
</tr>
<tr>
<td>Roessler, O</td>
<td>79, 83, 68, 71, 166</td>
</tr>
<tr>
<td>Roozbahani, A</td>
<td>173, 127</td>
</tr>
<tr>
<td>Roozbahani, F S</td>
<td>173, 194</td>
</tr>
<tr>
<td>Roy, J</td>
<td>116, 188, 193</td>
</tr>
<tr>
<td>Sahinalp, S C</td>
<td>13, 187</td>
</tr>
<tr>
<td>Sakar, C O</td>
<td>114, 149</td>
</tr>
<tr>
<td>Saleh, B O</td>
<td>153, 36</td>
</tr>
<tr>
<td>Salman, S</td>
<td>192, 36</td>
</tr>
<tr>
<td>Samol, C</td>
<td>36, 182</td>
</tr>
<tr>
<td>Sarac, F</td>
<td>65</td>
</tr>
<tr>
<td>Sarite, S R</td>
<td>178</td>
</tr>
<tr>
<td>Sarı, Ç</td>
<td>71</td>
</tr>
<tr>
<td>Schroeder, M</td>
<td>15, 116</td>
</tr>
<tr>
<td>Schwab, K</td>
<td>1, 10</td>
</tr>
<tr>
<td>Seker, H</td>
<td>75, 91, 201</td>
</tr>
<tr>
<td>Selcuki, C</td>
<td>182, 183, 184</td>
</tr>
<tr>
<td>Serhat, A</td>
<td>75</td>
</tr>
<tr>
<td>Sezerman, O U</td>
<td>40, 95, 103, 145</td>
</tr>
<tr>
<td>Smith, G</td>
<td>201</td>
</tr>
<tr>
<td>Somel, M</td>
<td>13</td>
</tr>
<tr>
<td>Son, Y A</td>
<td>194</td>
</tr>
<tr>
<td>Songurtok, D</td>
<td>198, 199</td>
</tr>
<tr>
<td>Sözer, N B</td>
<td>103</td>
</tr>
<tr>
<td>Spang, R</td>
<td>36</td>
</tr>
<tr>
<td>Specht, S</td>
<td>178</td>
</tr>
<tr>
<td>Spiridon, I F</td>
<td>6</td>
</tr>
<tr>
<td>Storey, J M D</td>
<td>1</td>
</tr>
<tr>
<td>Sugimoto, M</td>
<td>3, 31</td>
</tr>
<tr>
<td>Şanlier, Ş</td>
<td>187</td>
</tr>
<tr>
<td>Taner, D</td>
<td>158</td>
</tr>
<tr>
<td>Tasan, E</td>
<td>114</td>
</tr>
<tr>
<td>Telefoncu, A</td>
<td>130</td>
</tr>
<tr>
<td>Theuring, F</td>
<td>1, 10</td>
</tr>
<tr>
<td>Türkoglu, N</td>
<td>190</td>
</tr>
</tbody>
</table>
SPONSORS

SIGMA-ALDRICH™

INTERLAB

LABORATUAR ÜRÜNLERİ SAN. ve TİC. A.Ş

Türkiye tek yetkili distribütörü