6th Annual
Biotechnology and Bioinformatics Symposium

BIOT-2009

Proceedings

Edited by
Etsuko N. Moriyama, Leen-Kiat Soh, and Stephen D. Scott

University of Nebraska-Lincoln

October 9-10, 2009
General Symposium Information

[Meeting locations]
All BIOT-2009 activities take place in Nebraska Union (#17 in the UNL City Campus Map). See the Nebraska Union Map for location of the rooms:
- Registration: Regency
- Refreshments, lunch: Regency, Heritage, Ballroom (see the program)
- Oral presentations: Auditorium
- Banquet: Ballroom
- Poster session: Ballroom

Note: SCID workshop on Oct 11 takes place in Avery Hall (#32 in the UNL City Campus Map).

[Registration]
Registration desk will be open during the symposium at the Auditorium. You need to pick up your nametag and Proceedings book. The nametag will be required for meals at the symposium.

[Student/postdoc scholarship awards]
IMPORTANT! If you received a scholarship, you need to obtain your reimbursement forms at the registration desk. If you fail to pickup your reimbursement form during the symposium, you will not be able to receive your reimbursement.

[Oral presentation]
- All oral presentations are given in the Auditorium.
- Both Macintosh and Windows laptops with PowerPoint will be available.
- The length of a talk is 30 minutes for a full paper (shown with an F-number) and 20 minutes for an extended abstract (shown with an E-number). Please keep 3-5 minutes at the end of your talk for Q & A.
- All presenters, please copy your slide files to the laptop during the break before your presentation.
- Poster space will be also available for all accepted papers and extended (2-page) abstracts. See below for the information for poster presentation.

[Posters]
- Poster session is in the Ballroom. Refreshments will be provided during the poster session.
- See the program when you need to be at your poster.
- **Poster Size:** 3.5' or 42" (width) x 7' or 84" (height). Two posters may share the same poster board. So do not exceed the maximum width. Pushpins will be provided.
- **When to put up posters:** Posters can be put up any time from 12:00 pm to 5:30 pm on Oct 9 (Fri). Find your poster number on the board.
- **When to take down posters:** Please remove your posters by 1:00 pm on Oct 10 (Sat). All posters left after this time will be discarded.
## BIOT-2009 Program

Rooms: Auditorium (A; *in front of the Auditorium), Regency (R), Ballroom (B)  
Talks: Full paper (F-number), Extended abstract (E-number)

<table>
<thead>
<tr>
<th>October 9 (Friday)</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 am - 9:00 am</td>
<td>Registration</td>
</tr>
<tr>
<td>9:00 am - 9:10 am</td>
<td>Opening remarks</td>
</tr>
</tbody>
</table>
| 9:10 am - 10:15 am | **Keynote address 1**  
Chair: Stephen Scott (UNL, Computer Science & Engineering) | A |
| 10:35 am - 11:05 pm| **Session 1A.**  
Chair: Leen-Kiat Soh (UNL, Computer Science & Engineering) | A |
| 10:35 am - 11:05 pm| **F1.** Bagged gene shaving for the robust clustering of high-throughput data  
*Bradley M. Broom, Erik P. Sulman, Kim-Anh Do, Mary E. Edgerton, and Kenneth D. Aldape* | R |
| 11:05 pm - 11:35 pm| **F2.** Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic  
*Pavan K. Attaluri, Ximeng Zheng, Zhengxin Chen, and Guoqing Lu* | R |
| 11:35 pm - 12:05 pm| **F3.** Application of committee kNN classifiers for gene expression profile classification  
*Manik Dhawan, Sudarshan Selvaraja, and Zhong-Hui Duan* | R |
<p>| 12:00 pm - 1:15 pm | Lunch | R |</p>
<table>
<thead>
<tr>
<th>October 9 (Friday)</th>
<th>Room</th>
</tr>
</thead>
</table>
| **Keynote address 2**  
Chair: Leen-Kiat Soh (UNL, Computer Science & Engineering) | A |
| 1:15 pm - 2:20 pm  
*Vasant Honavar* (*Iowa State University, Computer Science*)  
Comparative Analysis of Macromolecular Interaction Networks | |
| 2:20 pm - 2:40 pm  
Coffee break | R |
| **Session 1B.**  
Chair: Etsuko Moriyama (UNL, Biological Sciences & Center for Plant Science Innovation) | A |
| 2:40 am - 3:00 pm  
E1. Association analysis in structured plant populations, an adaptive mixed LASSO approach  
*Dong Wang* | |
| 3:00 pm - 3:20 pm  
E2. A dynamic programming based algorithm to identify coordinately differentially activated transcription factors  
*Haiyan Hu and Xiaoman Li* | |
| 3:20 pm - 3:40 pm  
E3. Molecular evolution of sterol-sensing domain in eukaryotes  
*Pooja K. Strope and Etsuko N. Moriyama* | |
| 3:40 pm - 4:00 pm  
E4. Simulating sequence superfamilies for biological hypothesis testing  
*Cory L. Strope, Stephen D. Scott, and Etsuko N. Moriyama* | |
| 4:00 pm - 4:20 pm  
Coffee break | R |
| **Keynote address 3**  
Chair: Etsuko Moriyama (UNL, Biological Sciences & Center for Plant Science Innovation) | A |
| 4:20 pm - 5:25 pm  
*Andrew Benson* (UNL, Food Science & Technology, Core for Applied Genomics and Ecology)  
Reconstructing Microbiomes and Metagenomes from Massive Amounts of Sequence Data | |
| 5:25 pm - 5:40 pm  
Break | R |
Rooms: Auditorium (A; *in front of the Auditorium), Heritage (H), Ballroom (B)
Talks: Full paper (F-number), Extended abstract (E-number)

<table>
<thead>
<tr>
<th>October 9 (Friday)</th>
<th>Room</th>
</tr>
</thead>
</table>
| 5:40 pm - 10:00 pm | **Banquet & Posters**
| Presenters of posters should be present at the posters: |
| 7:30 pm - 8:15 pm: even numbered posters |
| 8:15 pm - 9:00 pm: odd numbered posters |
| B |

<table>
<thead>
<tr>
<th>October 10 (Saturday)</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 am -</td>
<td>Registration</td>
</tr>
<tr>
<td>8:00 am - 9:15 am</td>
<td>Continental breakfast</td>
</tr>
</tbody>
</table>
| | **Keynote address 4 (Sponsor: Center for Plant Science Innovation)**
| Chair: James Alfano (UNL, Plant Pathology & Center for Plant Science Innovation) | A |
| 9:15 pm - 10:20 pm | **Scott Peck** *(University of Missouri-Columbia, Biochemistry)*
| Using Proteomics to Study Signaling and Secretion in Plant Defense Responses |
| 10:20 am - 10:40 am | Coffee break | H |
| | **Session 2A.**
| Chair: Stephen Scott (UNL, Computer Science & Engineering) | A |
| 10:40 am - 11:10 am | **F4. Differential expression analysis of digital gene expression data: RNA-tag filtering, comparison of t-type tests and their genome-wide co-expression based adjustments**
| Yinglei Lai |
| 11:10 am - 11:40 am | **F5. Scaling alignment of large ontologies**
| Suzette K. Stoutenburg, Jugal Kalita, Kaily Ewing, and Lisa Hines |
| 11:40 am - 12:10 pm | **F6. Multiple genome sequence alignment with longest path algorithms**
| Fangrui Ma and Jitender Deogun |
| 12:10 pm - 1:30 pm | Lunch | B |

*viii*
Rooms: Auditorium (A), Heritage (H), Ballroom (B)
Talks: Full paper (F-number), Extended abstract (E-number)

<table>
<thead>
<tr>
<th>October 10 (Saturday)</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Session 2B.</strong></td>
<td>A</td>
</tr>
<tr>
<td>Chair: Guoqing Lu <em>University of Nebraska at Omaha, Biology</em></td>
<td></td>
</tr>
<tr>
<td>1:30 pm - 1:50 pm</td>
<td></td>
</tr>
<tr>
<td>E5. Oligonucleotide optimization for DNA synthesis</td>
<td></td>
</tr>
<tr>
<td><em>Tobias M. Louw, Scott E. Whitney, Joel TerMaat, and Hendrik Viljoen</em></td>
<td></td>
</tr>
<tr>
<td>1:50 pm - 2:10 pm</td>
<td></td>
</tr>
<tr>
<td>E6. Molecular evolution of G-proteins</td>
<td></td>
</tr>
<tr>
<td><em>Brandon L. Fulk, Stephen O. Opiyo, and Etsuko N. Moriyama</em></td>
<td></td>
</tr>
<tr>
<td>2:10 pm - 2:30 pm</td>
<td></td>
</tr>
<tr>
<td>E7. Gap profiling: scoring indels in multiple sequence alignment</td>
<td></td>
</tr>
<tr>
<td><em>Cory L. Strope, Stephen D. Scott, and Etsuko N. Moriyama</em></td>
<td></td>
</tr>
<tr>
<td>2:30 pm - 2:50 pm</td>
<td>H</td>
</tr>
<tr>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td><strong>Teaching Computing + Biology</strong></td>
<td></td>
</tr>
<tr>
<td>Chair: Leen-Kiat Soh <em>UNL, Computer Science &amp; Engineering</em></td>
<td></td>
</tr>
<tr>
<td>2:50 pm - 3:20 pm</td>
<td>A</td>
</tr>
<tr>
<td>Mark Pauley <em>University of Nebraska at Omaha, Information Science &amp; Technology</em></td>
<td></td>
</tr>
<tr>
<td>Integrating Bioinformatics into the Life Sciences: a CCLI Project</td>
<td></td>
</tr>
<tr>
<td>3:20 pm - 3:50 pm</td>
<td></td>
</tr>
<tr>
<td>Leen-Kiat Soh <em>UNL, Computer Science &amp; Engineering</em></td>
<td></td>
</tr>
<tr>
<td>The Renaissance Computing Initiative at the University of Nebraska</td>
<td></td>
</tr>
<tr>
<td>3:50 pm - 4:20 pm</td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td></td>
</tr>
<tr>
<td>4:20 pm - 4:30 pm</td>
<td>A</td>
</tr>
<tr>
<td>Closing remarks</td>
<td></td>
</tr>
<tr>
<td>4:30 pm - 5:30 pm</td>
<td>H</td>
</tr>
<tr>
<td>Co-chairs/Steering Committee meeting</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>October 11 (Sunday)</th>
<th>Avery Hall 115 and 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 am - 1:00 pm</td>
<td>NSF Semantic Cyberinfrastructure for Investigation and Discovery (SCID) Workshop (separate registration required)</td>
</tr>
</tbody>
</table>
Preface

On behalf of the Program Committee we would like to welcome you to BIOT-2009, the 6th Annual Biotechnology and Bioinformatics Symposium held on October 9 and 10, 2009, at the University of Nebraska-Lincoln, in Lincoln, Nebraska. This volume contains the papers and abstracts accepted for presentation at BIOT-2009.

BIOT-2009 follows what has now become the tradition of the BIOT series. The BIOT-2009 provides an affordable scientific event whose goal is to showcase research and development activities in bioinformatics and computational biology, and biotechnology, and to promote future interdisciplinary activity and research in these areas. It brings together scientists in multiple disciplines: computer science, engineering, mathematics, statistics, bioinformatics, biotechnology, life sciences, and medical science. Because different fields have their own tradition, fostering communication between such disparate fields naturally creates unique challenges. For example, in computer science, it is customary to require the submission of full papers for review several months prior to a conference or symposium. Conferences in life sciences generally require that an abstract or extended abstract be submitted for review. The organizing committee of BIOT understands these differences. For BIOT-2009, we accepted both full-length papers and 2-page extended abstracts for oral presentation. We also emphasized poster presentation by providing such opportunities for all of oral presenters. This is to foster informal interactions between students, postdocs, and faculty. Keeping the conference as a single-track event also helps to keep the audience together and facilitate communication across disciplines.

The BIOT conference series have been a growing success. Founded at the University of Colorado at Colorado Springs, the conference has grown into a national event with an increasing number of international participants. The first two events, as well as BIOT-2007, have been held at the founding institution. The 3rd Symposium, BIOT-2006, took place at Brigham Young University in Provo, Utah, and BIOT-2008 was held at University of Texas at Arlington. This year, for the first time, the Symposium is being held in the Midwest, attracting researchers from 10 states across the US including Nebraska, Iowa, and Kansas. Participants are also coming from five countries other than the US. The conference now has a 9–member Steering Committee, coming from 8 institutions in 6 states around the US.

We have received 10 full papers and 8 extended (2-page) abstracts submitted for oral presentation, and 43 poster submissions. While the posters have not been peer-reviewed, the papers and extended abstracts have been reviewed by at least one or, for many, 2 Program Committee members. From the submitted full-length manuscripts, 6 have been selected as regular papers and 2 as extended abstracts. From the extended (2-page) abstracts submitted, 5 have been accepted for oral presentation. All submissions to oral presentation have been invited also to present their works as posters in order to encourage active discussions in the interdisciplinary environment.

This year we are fortunate enough to receive funding from National Science Foundation for BIOT-2009 (DBI-0938224). With this support, we have established a Student/Postdoc Travel Scholarship program, and more than 45 students and postdocs who attended the Symposium and presented their works were awarded the scholarships.

BIOT-2009 could not be made to happen without many people's help. We are grateful to Program Committee members who spent their time reviewing the contributions. Dr. Nick Stojanovic from the University of Texas at Arlington gave us numerous invaluable suggestions based on his own experience from organizing BIOT-2008. Dr. Mark Clement from Brigham Young University has kindly let us access to the server hosting the website for BIOT-2009 and OpenConf. We are thankful to Center for Plant Science Innovation, Center for Biotechnology, Office of Research, School of Biological Sciences, Department of Computer Science and Engineering, and Institute of Agriculture and Natural Resources at UNL for their continuous support. The organization of BIOT-2009 would not have been possible without the strong support by Barb Gnirk, Teresa Loseke, Amy Knight, and Deb Heckens, who have taken care of many details of the event logistics and the fortunate headaches caused by the NSF-supported travel scholarship. Finally we would like to thank all authors who have chosen to present their research at BIOT-2009, and all participants who joined our discussion. Without all of them this conference would not have happened.
We hope that you will enjoy participating BIOT-2009, interacting with people from different disciplines, and staying in The Prairie Capital City, Lincoln, Nebraska.

On behalf of BIOT-2009 Program Committee,

Etsuko N. Moriyama  
School of Biological Sciences &  
Center for Plant Science Innovation  
University of Nebraska-Lincoln  
Lincoln, 68588-0118 NE  
emoriyama2@unl.edu

Leen-Kiat Soh  
Department of Computer Science &  
Engineering  
University of Nebraska-Lincoln  
Lincoln, 68588-0115 NE  
lksoh@cse.unl.edu

Stephen D. Scott  
Department of Computer Science &  
Engineering  
University of Nebraska-Lincoln  
Lincoln, 68588-0115 NE  
sscot@cse.unl.edu
Table of Contents

BIOT Symposium Steering Committee ........................................................................................................ xvii
BIOT-2009 Program Committee ................................................................................................................... xviii

Keynote Abstracts

Information Theory, Data Compression, and Bioinformatics .......................................................... 3
Khalid Sayood

Comparative Analysis of Macromolecular Interaction Networks ............................................... 4
Vasant G. Honavar.

Reconstructing Microbiomes and Metagenomes from Massive Amounts of Sequence Data .............. 5
Andrew Benson

Using Proteomics to Study Signaling and Secretion in Plant Defense Responses ................................. 6
Scott Peck

Teaching Computing + Biology

Integrating Bioinformatics into the Life Sciences: a CCLI Project ....................................................... 9
Mark Pauley

The Renaissance Computing Initiative at the University of Nebraska ................................................. 10
Leen-Kiat Soh, Stephen Scott, and Etsuko Moriyama

Accepted Full Papers

F1. Bagged gene shaving for the robust clustering of high-throughput data ........................................... 13
Bradley M. Broom, Erik P. Sulman, Kim-Anh Do, Mary E. Edgerton, and Kenneth D. Aldape

F2. Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic.................................................................................................................................................. 21
Pavan K. Attaluri, Ximeng Zheng, Zhengxin Chen, and Guoqing Lu

F3. Application of committee $k$NN classifiers for gene expression profile classification ............... 28
Manik Dhawan, Sudarshan Selvaraja, and Zhong-Hui Duan

F4. Differential expression analysis of digital gene expression data: RNA-tag filtering, comparison of $t$-type tests and their genome-wide co-expression based adjustments ......................................................... 33
Yinglei Lai

F5. Scaling alignment of large ontologies ................................................................................................. 41
Suzette K. Stoutenburg, Jugal Kalita, Kaily Ewing, and Lisa Hines

xii
F6. Multiple genome sequence alignment with longest path algorithms .................................................... 50
Fangrui Ma and Jitender Deogun

Accepted Extended Abstracts

E1. Association analysis in structured plant populations, an adaptive mixed LASSO approach .......... 63
Dong Wang

E2. A dynamic programming based algorithm to identify coordinately differentially activated
transcription factors .................................................................................................................................... 65
Haiyan Hu and Xiaoman Li

E3. Molecular evolution of sterol-sensing domain in eukaryotes ............................................................... 67
Pooja K. Strope and Etsuko N. Moriyama

E4. Simulating sequence superfamilies for biological hypothesis testing .................................................. 69
Cory L. Strope, Stephen D. Scott, and Etsuko N. Moriyama

E5. Oligonucleotide optimization for DNA synthesis ................................................................................ 71
Tobias M. Louw, Scott E. Whitney, Joel TerMaat, and Hendrik Viljoen

E6. Molecular evolution of G-proteins ....................................................................................................... 73
Brandon L. Fulk, Stephen O. Opiyo, and Etsuko N. Moriyama

E7. Gap profiling: scoring indels in multiple sequence alignment ............................................................. 75
Cory L. Strope, Stephen D. Scott, and Etsuko N. Moriyama

Poster Abstracts

P427. MicroRNA target prediction and verification in the green alga Chlamydomonas reinhardtii .......... 79
Adam Voshall, Etsuko N. Moriyama, and Heriberto Cerutti

P672. Screening for anti-angiogenic natural inhibitors for vascular endothelial growth factor receptor - An
in silico approach ........................................................................................................................................ 80
Subhojyoti Chatterjee, Deepak Choubey, Biplab Bhattacharjee, Anantharamanan. R., Jayadeepa R. M.,
Anirudh Chowdhary, Sushil K. Middha, and Usha Talambedu

P686. Molecular basis of homocystinuria: identification of mutation in cystathionine beta synthase (CBS)
gene ............................................................................................................................................................. 81
M. A. Kausar and M. D. Bashyam

P707. Codon use in ribosomal protein ........................................................................................................ 82
William L. Duax, Robert Huether, David Dziak, Sanjay Connare, Evan Ashton, Jane Bednarz, Kevin
Gibas, Fiona Hennig, Dana Hogan, Tyler Kirsch, Luke Klein, Charlotte Lane, Daniel Maiman, Courtney
McEachon, Rasheen Powell, Brian Riefler, Derek Robbins, Emma Rutkowski, Patrick Ryan, Adam Schunk,
Jimmitti Teysir, Joseph Trapp, and Isabel Xu
P1192. Tools for identifying homogenous subgroups in large data ............................................................ 83
Yuanyuan Huang, Heike Hofmann, and Dianne Cook

P1297. Biosynthetic engineering of HSAF, a tetramic acid-containing antifungal macrolactam isolated from the biocontrol agent Lysobacter enzymogenes C3 ............................................................................. 84
Yunxuan Xie, Yaoyao Li, Lili Lou, and Lingcheng Du

P1333. Extraction of globular proteins with reverse micelles using vegetable oils as the organic medium ........................................................................................................................................ 85
Daliya A. George, Cheuk-Man Chu, and David C. Stuckey

P1427. Matched filter for signature extraction in bioinformatics data set .................................................. 86
Ming Li and William Q. Meeker

P1439. Bioinformatics investigation: examination of two algal proteomes in search of carbon-concentrating mechanism components ....................................................................................................... 87
Andrew Brueggeman, Istvan Ladunga, and Donald Weeks

P1446. Phylogenetic analysis of H1N1 2009 influenza virus ..................................................................... 88
Ameer Shaik, Aparna Sriram, Sri Harsha Muppaneni, Teja Polapragada, and Zhong-Hui Duan

P1447. Graphical framework for processing GCMS data: integration of qt interfaces .............................. 89
Marie Vendettuoli, Michael Lawrence, Dianne Cook, and Heike Hofmann

P1454. Immune-enhancing antitumor RNA interference therapy for cervical cancer ................................ 90
Richa Singhania, Wenyi Gu, and Nigel McMillan

P1456. Layout of phylogenetic networks and multi-labeled trees ................................................................ 91
Poornima Venugopal and Sergey Bereg

P1466. Intraperitoneal delivery of liposomal siRNA for therapy of cervical cancer .................................. 92
Akul Singhania, Sherry Y. Wu, and Nigel A. J. McMillan

P1467. Knock down of CXCR2 enhances sensitivity to chemotherapy ..................................................... 93
Dhananjay M. Nawandar, Michelle L. Varney, and Rakesh K. Singh

P1468. Integrating EST's and high-throughput sequencing to identify orthologous loci for phylogenomic analysis: a case study among basal ray-finned fishes (Actinopterygii) ...................................................... 94
Jason Macrander and Guillermo Orti

P1474. Identifying differentially expressed human lung microRNAs and their molecular functions ...... 95
Sarita Limbu and Zhong-Hui Duan

P1475. Development of technologies to monitor chromatin proteins in small cell numbers for applications in assisted reproductive technology ........................................................................................................ 96
Dipika Singh, Janos Zempleni, and Angela Pannier

P1484. SUMO (small-ubiquitin-like modifier) in subcellular fractions of Chlamydomonas reinhardtii reveals potential functions .................................................................................................................. 97
Amy R. Knobbe, Kempton Horken, Stephanie Matejka, Don Weeks, and Cheryl P. Bailey
P1486. Establishing influenza A viral mutation database through literature and data mining .......................... 98
Ambreen Kedwaii, Pavan Kumar Attaluri, and Guoqing Lu

P1487. Micromechanics based constitutive relationships for collagenous tissue ........................................... 99
Ranganathan Parthasarathy, Anil Misra, Viraj Singh, and Paulette Spencer

P1489. Enhancement of homologous recombination in the model organism
Chlamydomonas reinhardtii .......................................................................................................................... 100
Thomas M. Plucinak and Donald P. Weeks

P1490. An SVM-based method for linear B-cell epitope prediction by introducing a new tri-peptides
Kernel ........................................................................................................................................................ 101
Bo Yao and Chi Zhang

P1492. Pattern detection in the motor protein prestin: a collaborative approach ........................................... 102

P1495. A Web-based tool for rigidity analysis of proteins ............................................................................. 103
Naomi Fox, Filip Jagodzinski, Yang Li, and Ileana Streinu

P1497. Coevolution between PKD1 and PKD2 ......................................................................................... 104
Wenhu Guo and Chun Ye

P1498. A software tool for surveying the rigidity properties of protein families ........................................ 105
Filip Jagodzinski, Diana Jaunzeikare, Naomi Fox, and Ileana Streinu

P1499. A new approach to compare biological sequences based on motif alphabets ............................... 106
Ramez Mina, Dhundy Bastola, and Hesham Ali

P1500. Molecular phylogeny and evolution of pepsinogen subfamily ....................................................... 107
Michael Wachholtz, Jin-Liang Zhao, and Guoqing Lu

P1503. Evolution of gram-negative bacteria: a view from lipid A biosynthesis enzymes ......................... 108
Stephen O. Opiyo, Rosevelt L. Pardy, Hideaki Moriyama, and Etsuko N. Moriyama

P1517. Chromosome specific detection of NOCTH protein binding sites on mouse genome using Chip-
Seq ............................................................................................................................................................. 109
Zhijie Jiang, Prathibha Ranganathan, Anthony J. Capobianco, and Nicholas F. Tsinoremas

P1519. Comparative genome hybridization of maize inbreds and their wild ancestor, teosinte .............. 110
Kai Ying, Nathan M. Springer, Yan Fu, Tieming Ji, Cheng-Ting Yeh, Yi Jia, Wei Wu, Todd Richmond,
Jacob Kitzman, Heidi Rosenbaum, A. Leonardo Iniguez, W. Brad Barbazuk, Jeffrey A. Jeddeloh, Dan
Nettleton, and Patrick S. Schnable

P1567. Phylomarker: mining phylogenetic markers through comparative genomics of model species and
data mining of public repositories ............................................................................................................. 111
Thaine W. Rowley, Chenhong Li, and Guoqing Lu

P1568. Genome-wide comparison of G-protein coupled receptors in model fish species ...................... 112
Kanwal Naeem, Etsuko N. Moriyama, and Guoqing Lu

xv
P1590. Sorting through the sequence: making sense of microbial communities with amplicon pyrosequencing ......................................................................................................................................... 113
Ryan Legge, Andrew Benson, and Etsuko Moriyama

P1596. Proteomics reconstruction of chemotaxis and acrosome reaction signaling pathways in sea star sperm......................................................................................................................................................... 114
Yuyan Shelly Xue and Hideaki Moriyama

P1612. Hypoglycemic effects and chemical constituents of some indigenous medicinal plants from Lawacherra rain forest of Bangladesh ................................................................................................................. 115

P1614. Development and use of the AllergenOnline.org curated protein database for assessing the potential allergenicity of genetically modified organisms and novel food ingredients ......................... 116
John Wise, Steve L. Taylor, and Richard E. Goodman

P1655. Species-specific duplications of trace amine-associated receptors in vertebrate genomes........... 117
Seong-il Eyun and Etsuko N. Moriyama

P1657. Discovering SNP association patterns with spatial constraint using data mining approach........ 118
Fangrui Ma, Liying Jiang, Etsuko Moriyama, and Jitender S. Deogun

P1659. A web application for the analysis of differential histone modification states from ChIP-seq data................................................................................................................................................................................................. 119
Sridhar A Malkaram, Rong Liu, Michael Fromm, and Jean-Jack M Riethoven

P1664. Comparative miRNA analysis in pathogenic fungi ........................................................................ 120
Kanika Jain and B.B. Chattoo

Author index ............................................................................................................................................. 121
BIOT Symposium Steering Committee

Abdullah N. Arslan
Department of Computer Science and Information Systems
Texas A & M University - Commerce
Commerce, Texas, USA

Bradley Broom
Department of Bioinformatics and Computational Biology
MD Anderson Cancer Center
Houston, Texas, USA

Mark Clement
Department of Computer Science
Brigham Young University
Provo, Utah, USA

William Duax
Department of Structural Biology
State University of New York
Buffalo, New York, USA

Rasiah (Raja) Loganantharaj
Department of Computer Science
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

Jugal Kalita
Department of Computer Science
University of Colorado at Colorado Springs
Colorado Springs, Colorado, USA

James Mattoon
Department of Biology
University of Colorado at Colorado Springs
Colorado Springs, Colorado, USA

Etsuko Moriyama
School of Biological Sciences
Center for Plant Science Innovation
University of Nebraska-Lincoln
Lincoln, Nebraska, USA

Nikola Stojanovic
Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, Texas, USA
BIOT-2009 Program Committee

Symposium Co-Chairs
Etsuko Moriyama, School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska, Lincoln, NE
Stephen Scott, Department of Computer Science and Engineering, University of Nebraska, Lincoln, NE
Leen-Kiat Soh, Department of Computer Science and Engineering, University of Nebraska, Lincoln, NE

Program Committee
Vassilis Athitsos, Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas
Robert Barber, Department of Surgery, University of Texas Southwestern Medical Center, Dallas, Texas
Esther Betran, Department of Biology, University of Texas at Arlington, Arlington, Texas
Chengpeng Bi, Children's Mercy Hospital, School of Medicine, University of Missouri - Kansas City, Kansas City, Missouri
Dominika Borek, Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas
Kun-Mao Chao, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
Keith A. Crandall, Department of Biology, Brigham Young University, Provo, Utah
Ugur Dogrusoz, Department of Computer Engineering, Bilkent University, Ankara, Turkey
Suash Deb, Department of Computer Science and Engineering, C. V. Raman College of Engineering, Bhubaneswar, India
Zhong-Hui Duan, Department of Computer Science, University of Akron, Akron, Ohio
Liliana Florea, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland
Jean Gao, Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas
Preetam Ghosh, Department of Computer Science, University of Southern Mississippi, Hattiesburg, Mississippi
Samik Ghosh, The Systems Biology Institute, Tokyo, Japan
Marico Howe, Department of Information Science, University of Arkansas at Little Rock, Little Rock, Arkansas
Hung-Chung (Joe) Huang, Department of Systems Biology and Translational Medicine, Texas A&M Health Science Center, Temple, Texas
Vasant Honavar, Department of Computer Science, Iowa State University, Ames, Iowa
Lynn Little, Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas
Yaling Liu, Department of Mechanical and Aerospace Engineering, University of Texas at Arlington, Arlington, Texas
Chin Lung Lu, Institute of Bioinformatics and Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan
Guoqing Lu, Department of Biology, University of Nebraska at Omaha, Omaha, Nebraska
Ilias Maglogiannis, Department of Information and Communication Systems Engineering, University of the Aegean, Karlovassi, Samos, Greece
Hideaki Moriyama, School of Biological Sciences & Center for Biotechnology, University of Nebraska-Lincoln, Lincoln, Nebraska
Mark Pauley, IS&T Bioinformatics, College of Information Science and Technology, University of Nebraska at Omaha, Omaha, Nebraska
Yu Qian, Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas
Prahlad Ram, Department of Systems Biology, MD Anderson Cancer Center, Houston, Texas
Alberto Riva, Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, Florida
Abanish Singh, Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas
Nikolai Slavine, Department of Radiology, University of Texas Southwestern Medical Center, Dallas, Texas
Keynote Abstracts
Information Theory, Data Compression, and Bioinformatics

Khalid Sayood

Department of Electrical Engineering, University of Nebraska-Lincoln

The development of data compression algorithms requires an understanding of how information is organized in data. Bioinformatics seems the ideal application area for many of the information theoretic concepts, which underlie data compression algorithms. In this talk we explore if we can actually go further and see these information theoretic concepts arising organically from the area of bioinformatics without reference to data compression.

Biosketch:

Dr. Khalid Sayood received his Ph.D. in Electrical Engineering in 1977 from Texas A&M University. He is currently a Henson Professor of Engineering at Department of Electrical Engineering, University of Nebraska-Lincoln.

Research Interests:

Dr. Khalid Sayood's principal interest is in the search of patterns in data. He indulges this interest by looking at problems in data compression, joint source-channel coding, and various aspects of bioinformatics.
Comparative Analysis of Macromolecular Interaction Networks

Vasant G. Honavar

Department of Computer Science, Iowa State University

Network models are playing an increasingly important role in the interpretation of complex interactions among genes, proteins, regulatory RNAs, small ligands and other signaling agents. This talk will (a) review different classes of network models and their applications in computational systems biology; (b) discuss the comparative analysis of network models of biomolecular interactions across different species or tissues and how it has emerged as an important tool for identifying conserved modules, predicting functions of specific genes or proteins and studying the evolution of biological processes, among other applications; and (c) describe a class of graph-kernel based algorithms for aligning large biomolecular networks by decomposing such networks into subnetworks and computing the alignment of the networks based on the alignment of the subnetworks. The resulting algorithms as part of BiNA, an open source, modular, extensible, and customizable biomolecular network alignment toolkit. The results of our experiments with protein-protein interaction networks show that the proposed approach is competitive with the state-of-the-art algorithms for aligning such networks. This work has been carried out in collaboration with Fadi Towfic, a Ph.D. Student in Bioinformatics and Computational Biology.

Biosketch:

Dr. Vasant Honavar received his Ph.D. in Computer Science in 1990 from the University of Wisconsin at Madison, specializing in Artificial Intelligence. Since 1990, he has been on the faculty of Iowa State University (ISU) where he is currently a Professor of Computer Science and of Bioinformatics and Computational Biology and Director of the Center for Computational Intelligence, Learning & Discovery. He served as the chair (during 2003-2005) of the ISU Bioinformatics and Computational Biology Graduate Program which he helped establish in 1999 with support from an Integrative Graduate Education and Research Training (IGERT) award. Dr. Honavar’s research has resulted in foundational contributions in machine learning, bioinformatics and computational biology. He has published over 200 research articles in peer-reviewed journals and conferences (including several that have received best paper awards). Eighteen Ph.D. students and twenty-two M.S. students have graduated under Dr. Honavar's supervision. Dr. Honavar serves on the editorial boards of several scientific journals and program committees of several conferences. Honavar was chosen for the Iowa Board of Regents Award for Faculty Excellence in 2007 and the Iowa State University College of Liberal Arts and Sciences Award for Research Excellence in 2008.

Research Interests:

Artificial Intelligence, Bioinformatics, Computational Biology (especially modeling and analysis of macromolecular interactions), Machine Learning (especially algorithms for learning predictive models from richly structured data), Data Mining (especially scalable algorithms for learning predictive models from large, distributed, and semantically disparate data), Logical, Probabilistic, and Decision-Theoretic Knowledge Representation, Information Integration, Service Composition, Social Informatics, Security Informatics
Order from Chaos: Reconstructing Microbiomes and Metagenomes from Massive Amounts of Sequence Data

Andrew K. Benson

Department of Food Science and Technology, University of Nebraska-Lincoln

Microbiomes (collections of microbial species that are unique to specific ecosystems) numerically dominate soil, water, intestinal and other environments, and have attracted significant effort aimed at defining and manipulating them. Pioneering work in this field focused on sequencing 16S rRNA gene that is conserved among all Bacteria and creating large databases of sequences from known organisms that can be referenced to sort through a diverse microbial population. Advances in Next Generation DNA sequencing technologies (e.g. pyrosequencing) now make it routine for investigators to query composition of complex microorganisms by sequencing thousands to hundreds of thousands of 16S rRNA gene sequences from DNA extracted from environmental, clinical, or even food samples. This approach has already found many applications in medicine, agriculture, and the environment. These analyses are performed by deep pyrosequencing of PCR products amplified from the 16S rRNA gene; subsequently, each complete 16S rRNA sequence read must be assigned a taxonomic status by comparing it to a database of sequences from known and previously classified taxa. Word-based approaches, such as the CLASSIFIER algorithm, have developed to be very efficient at this task but suffer from the fact that 50% of more of the sequences from any given environment are unique and not present in the database, leaving us blind to classifying and quantifying significant portions of the microbiome. To circumvent this problem, alignment and distance estimates between millions of sequence reads can be used to groups similar sequences into Operational Taxonomic Units (OTUs). However, the hundreds of millions of computational operations that are required by this so-called OTU-picking approach make it computationally daunting, and current algorithms simply cannot handle such a volume of data. At the same time, microbiome studies are moving beyond descriptions of different microbiome compositions to quantifying factors that control it demanding high-throughput computational methods and databases to process and warehouse microbiome data. Furthermore, the blossoming of new capacities in microbial ecology has recently produced an explosion of interest in metagenomics, the study of the collective genetic capacity of all organisms in an ecosystem. Metagenomes are assembled from shotgun sequencing of environmental DNA samples, and contain sequence reads from fragments of all species in a given environment. This approach provides a less biased and more comprehensive view of both species content and genetic content than does the microbiome approach. Computationally, metagenome analysis depends on accurate assembly of shotgun sequences from random fragments in a complex mixture of organisms, a task that is exceedingly more complex than assembling a single genome. Moreover, as sequencing platforms progress in their capacity, the computational challenges will also progress from assembling complex and massive jumbles of random sequences into individual contigs to assembling these individual contigs into whole genomes. This challenge will be accompanied by a growing demand for high-throughput gene-finding and functional assignment.

Biosketch:

Dr. Andrew Benson received his Ph.D. in Microbiology in 1992 from the University of Texas Health Science Center at San Antonio. He is currently a Professor at the Department of Food Science and Technology and the Director of the Core for Applied Genomics and Ecology at University of Nebraska-Lincoln.
Using Proteomics to Study Signaling and Secretion in Plant Defense Responses

Scott C. Peck

Department of Biochemistry, University of Missouri-Columbia

Plants perceive potential pathogens by recognizing pathogen-associated molecular patterns (PAMPs) through plasma membrane (PM) receptors. Recognition of flg22, a 22 amino acid PAMP derived from the bacterial flagellum, by the receptor-like kinase FLS2 induces defense signaling responses and contributes to innate immunity by restricting bacterial invasion. A quantitative phosphoproteomic analysis of Arabidopsis PM proteins revealed specific phosphorylation sites changing in response to treatment with flg22 using a [Nühse et al., (2007) Plant J 51:931]. By pursuing reverse genetic investigations of differentially phosphorylated protein candidates, we have discovered that a specific syntaxin, SYP132, appears to be the cognate t-SNARE required for secretion of antimicrobial proteins and/or compounds, implicating protein secretion as a major determinant of resistance. [Kalde M, et al., (2007) PNAS 104:11850]. Subsequent proteomic analyses of proteins secreted during infection by different genotypes of bacterial pathogens revealed a complex interplay between Type III secretion from the bacteria, host resistance responses, and alteration of the host secretome [Kaffarnik et al., (2009) Mol Cell Proteomics 8:145]. Because of our interest in the dynamics of PM proteins during bacterial infections, we have been developing improved strategies for simplified analyses of PM proteomes without the need for two-phase partitioning. Our progress in method development and initial application to the study of plant defense responses will be discussed.

Biosketch:

Dr. Scott Peck received his Ph.D. in Botany and Plant Pathology in 1995 from the Michigan State University. He is currently an Associate Professor in Department of Biochemistry, University of Missouri-Columbia.

Research Interests:

We study how potential hosts recognize and respond to invading microbes, particularly bacteria. Using proteomic and phosphoproteomic approaches, we have identified numerous proteins involved in cellular signaling and protein secretion. We are currently using a diverse platform of biochemical and genetic approaches to investigate the functions of these proteins.
Teaching Computing + Biology
Integrating bioinformatics into the life sciences: a CCLI project

Mark A. Pauley
College of Information Science & Technology
University of Nebraska at Omaha
Omaha, NE 68182-0116 USA
Email: mpauley@mail.unomaha.edu

Bioinformatics is a young discipline that integrates mathematical and computational techniques with biological knowledge to analyze genetic information. Bioinformatics techniques are playing an increasingly important role in the biosciences. Using these techniques, analysis of the vast amounts of data produced by DNA sequencing and mapping projects such as the Human Genome Project is uncovering hitherto unknown relationships between genes and diseases, having a profound impact on drug development and clinical trials, and is affecting medical diagnostics, pharmacogenomics and agricultural and industrial biotechnology.

It is widely recognized that there is a strong demand for people with bioinformatics skills and the national need for a workforce educated in bioinformatics and biotechnology is well-documented. For example, the worldwide value of bioinformatics is expected to increase three-fold between 2002 ($1.02 B) and 2010 ($3.0 B, estimated) with the fastest growing market expected to be analysis software and services. Furthermore, during this time-period pharmaceutical companies are expected to increase their expenditures in research and development and a major portion of this spending is expected to be in bioinformatics.

Despite its increasing importance, there is a general lack of integration of bioinformatics concepts into the undergraduate curriculum in the life sciences and related disciplines where they could readily be applied. At the University of Nebraska at Omaha (UNO) for example, most biology students graduate with minimal exposure to bioinformatics. Further, although molecular biology is a rich problem domain for computer scientists, most computer science students are not introduced to its potential. Finally, few students, especially at the high school level, are aware of bioinformatics as a scientific discipline or as a possible career path.

The project discussed in this presentation is designed to address the challenges described above. In particular, the goal of our project is to develop, pilot test and disseminate a set of modules in bioinformatics that can be integrated into curricula in the biological sciences, computer science and other disciplines. Each module, consisting of approximately five hours of instruction with a problem-based learning component and an illustrative homework assignment, addresses a fundamental concept in bioinformatics (e.g., algorithms, databases, etc.). In addition, one module has been adapted to introduce bioinformatics at the high school level.

The subjects of the modules cover concepts that are of fundamental importance to bioinformatics but also reflect the strengths of the project team: Introduction to Bioinformatics, Gene Expression and Regulation, Algorithms in Bioinformatics, Science and the Scientific Method, Databases/Data Mining and Bioinformatics Tools. Each of the five investigators on the project is in charge of a least one of the modules. At UNO, the modules are being introduced into the curricula by the authors providing instruction during an initial offering. For those outside UNO, a dedicated website http://ccli.ist.unomaha.edu is available. On this website we provide all instructional materials that have been developed during the course of the project.

To document whether the modules are being developed correctly, whether they are contributing to student understanding of bioinformatics concepts, and whether selected modules have the potential to help recruit high school students into bioinformatics and STEM study, the project is being evaluated outside experts. For each module, a short assessment consisting of ten items that test key ideas have been developed by the project team with review and advice from the evaluators. Data from the assessment taken before and after the modules (pre/posttest) are used to calculate the reliability of the test instrument and the effectiveness of the instruction.

The focus of this presentation will be a discussion of the specific content of the modules and will include our experiences with integrating the content modules into specific courses. Also discussed will be UNO’s unique undergraduate degree program in bioinformatics.
The Renaissance Computing Initiative at the University of Nebraska

Leen-Kiat Soh¹, Stephen Scott¹, and Etsuko Moriyama²

¹Department of Computer Science and Engineering, University of Nebraska-Lincoln
²School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln

In the narrowest sense, "computational thinking" is the mindset that students need to acquire in order to work effectively with computational systems. More broadly, however, it is a way of understanding the world – one that transcends mere methodology and which is likewise transportable across a wide variety of human endeavors. In our view, computer science curricula that do not address this broadening do not address the ways in which computational thinking pervades life in the modern world.

Here at the University of Nebraska, we have embarked on a broad-based curricular and program revision initiative called Renaissance Computing. In our conception, "computational thinking" is neither easily separated from other endeavors nor easily balkanized into a single department. We thus imagine a CS program that is inextricably linked to other domains. We further understand these domains to include not only the subjects ordinarily thought of as cognate with computational thinking (such as bioinformatics), but with such notions as "humanities computing," "arts computing," and "music computing."

Furthermore, we postulate that students of different disciplines will be able to benefit from each other through collaborative activities and cross-pollination of ideas.

Biosketch:

Dr. Leen-Kiat Soh received his Ph.D. in Electrical Engineering with Honors from the University of Kansas. Dr. Leen-Kiat Soh is currently an Associate Professor at the Department of Computer Science and Engineering of the University of Nebraska.

Research Interests:

His research interests are in multiagent systems, computer-aided education, and computer science education. He is the PI of the NSF-funded Renaissance Computing project, which aims at improving CS curriculum for CS majors and non-majors, particularly addressing CS1 courses and revitalized CS minor programs. He has published his work in conferences such as AAAI, IAAI, AAMAS, SIGCSE, and ITiCSE, and in journals such as the Journal of Autonomous Agents and Multiagent Systems, Computer Science Education, International Journal of AI in Education, and IEEE Transactions.
Accepted Full Papers
Bagged Gene Shaving for the Robust Clustering of High-Throughput Data

Bradley M. Broom*, Erik P. Sulman†, Kim-Anh Do‡, Mary E. Edgerton§, and Kenneth D. Aldape¶

*Department of Bioinformatics and Computational Biology, UT MD Anderson Cancer Center, Houston, Texas 77030
†Department of Radiation Oncology, UT MD Anderson Cancer Center, Houston, Texas 77030
‡Department of Biostatistics, UT MD Anderson Cancer Center, Houston, Texas 77030
§Department of Pathology, UT MD Anderson Cancer Center, Houston, Texas 77030
Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic

Pavan K. Attaluri
Department of Computer Science
University of Nebraska at Omaha
Omaha, NE 68182, USA
pattaluri@mail.unomaha.edu

Ximeng Zheng
Department of Computer Science
University of Nebraska at Omaha
Omaha, NE 68182, USA
xzheng@unomaha.edu

Zhengxin Chen
Department of Computer Science
University of Nebraska at Omaha
Omaha, NE 68182, USA
zchen@unomaha.edu

Guoqing Lu
Departments of Biology, University of Nebraska at Omaha, Omaha, NE 68182, USA
glu3@mail.unomaha.edu

Abstract—A phase 6 alert has been declared by the World Health Organization (WHO) in response to the ongoing global spread of the influenza H1N1 virus in humans. Genetic sequence analysis suggests that this pandemic strain evolves from reassortment of swine viruses. The objective of this research is to conduct a series of bioinformatics analyses to characterize currently circulating pandemic influenza viral strains and identify their evolutionary origin. Three groups of sequences (i.e., human, swine, and the latest pandemic human/swine) were used for phylogenetic analysis, decision tree analysis and support vector machine (SVM) analysis. Our results strongly support the finding that the latest pandemic viral strain is of swine origin. To facilitate early detection of human and swine H1N1 viral strains, we have developed a web tool based upon the results obtained in this study and used Hidden Markov Model (HMM) for accurate prediction of influenza H1N1 origin.

Index Terms—H1N1, influenza A virus, machine learning, swine flu

I. INTRODUCTION

Influenza is one of the most important emerging and reemerging infectious diseases, causing high morbidity and mortality in communities (epidemic) and worldwide (pandemic) [1]. The influenza virus is an RNA virus and comprises three types: A, B, and C based upon their protein consumption. Influenza A is the most virulent human pathogen among the three types and is believed to be responsible for the global outbreaks of 1918, 1957, 1968 and 2009 [2]. Influenza A is subdivided into subtypes based on two surface proteins, HA and NA. Mutations on these proteins may result in different influenza subtypes. So far, there are 16 H and 9 N serotypes found in influenza virus. This genetic drift process often results in different strains of H1N1 and H3N2 circulating in humans during annual influenza seasons. Another process called genetic shift undergoes infrequent and sudden changes of genome segments from different viral strains, which is speculated to be the major cause for influenza pandemics [3].

An influenza pandemic occurs when a new influenza virus appears against which the human population has no immunity. In the past, influenza pandemics have resulted in increased death, disease and great social disruption. Influenza A virus caused three major global epidemics during the 20th century: the Spanish Flu in 1918, Asian Flu in 1957 and Hong Kong Flu in 1968. These pandemics were caused by strains of influenza A virus that had undergone major genetic changes and for which the population did not possess significant immunity [4]. The 2009 flu pandemic is a global outbreak of a new influenza A virus H1N1 strain, identified in April 2009 and commonly referred to as Swine Flu [5]. Analysis suggests that the H1N1 strain responsible for the current outbreak first evolve around September 2008 and circulate in the human population for several months before the first cases were identified [6]. Fig. 1 depicts genetic origins of the 2009 swine flu virus.

Fig.1. Genetic structure of influenza A virus (left panel) and the origins of the 2009 swine flu virus (right panel).

In order to effectively deal with the problem of identifying origin of pandemic swine flu viral strains, we have explored approaches using machine learning techniques. Machine learning is a subfield of artificial intelligence and is concerned with the development of algorithms that allow computers to learn, and has found many applications in bioinformatics [7]. Machine-learning approaches are suitable for datasets containing large amounts of data and presence of noisy patterns. The idea behind these techniques is to learn the theory automatically from the data, through a process of inference, model fitting or learning from examples. A number
of machine learning approaches are applied to identify relationships or associations in biological data, to group similar genetic elements, to analyze and predict diseases. Current research domains where machine learning techniques applied are multiple sequence alignment, structure and function prediction, molecular clustering and classification, and expression analysis.

The classical ways of determining the subtype of influenza virus for HA and NA segments are hemagglutination-inhibition (HI) assay and neuraminidase-inhibition (NI) assay which are capable of distinguishing antigenic differences between influenza even of the same subtype. However, as noted in [8], when working with uncharacterized viruses or antibody subtypes, the library of reference reagents required for identifying antigenically distinct influenza viruses and/or antibody specificities from multiple lineages of a single hemagglutinin subtype requires extensive laboratory support for the production and optimization of reagents. Using machine learning methods to predict the lineage of virus is much cheaper and faster, yet usually can still yield high accuracy.

In our current study, we have applied two machine learning techniques (decision trees and support vector machines) to identify the origin of latest pandemic outbreak strains. Furthermore, based on Hidden Markov Model (HMM), we have developed a Web system for the prediction of influenza A virus hosts using the informative positions found through decision tree method. Research on the evolution of latest swine flu outbreak will improve the design of vaccines and diagnostic tools.

II. DATA AND METHODS

A. Data

Three groups of influenza A H1N1 sequences were determined, including Human (found only in human), Swine (found only in swine), and Human_Swine (i.e., the latest pandemic influenza viral strains). Sequences were downloaded from the Global Initiative on Sharing All Influenza Data (GISAID) and the National Center for Biotechnology Information (NCBI) [9]-[10]. Sequence comparison showed that the sequence dataset from GISAID is more complete compared with that of NCBI. We therefore used sequences from GISAID in the analysis. When comparing sequences from different groups, we found that a number of sequences appeared in the human group and in the human swine group as well. These sequences were excluded for further analysis.

Table I shows the count of sequences used for each segment and host. We restricted to a maximum of 150 sequences of each host and of each segment. Part of the data is used for training and the remaining part is used for testing. The sequence data from segments HA, NA, PA, NS, PB1, PB2, M, and NP are considered for SVM and decision tree analysis. We used nucleotide sequences for decision tree analysis, and protein sequences for support vector machine analysis.

<table>
<thead>
<tr>
<th>Segment</th>
<th>Human</th>
<th>Swine</th>
<th>Human_Swine</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA</td>
<td>150</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>M</td>
<td>150</td>
<td>147</td>
<td>150</td>
</tr>
<tr>
<td>NA</td>
<td>150</td>
<td>148</td>
<td>150</td>
</tr>
<tr>
<td>NP</td>
<td>150</td>
<td>150</td>
<td>137</td>
</tr>
<tr>
<td>NS</td>
<td>150</td>
<td>147</td>
<td>111</td>
</tr>
<tr>
<td>PA</td>
<td>150</td>
<td>146</td>
<td>85</td>
</tr>
<tr>
<td>PB1</td>
<td>150</td>
<td>145</td>
<td>93</td>
</tr>
<tr>
<td>PB2</td>
<td>150</td>
<td>146</td>
<td>98</td>
</tr>
</tbody>
</table>

B. Sequence alignment

We used MUSCLE (http://www.drive5.com/muscle/) for multiple sequence alignment. MUSCLE stands for multiple sequence comparison by log-expectation, and is one of the most popular multiple alignment software for protein and nucleotide sequences [11]. MUSCLE can achieve both better average accuracy and better speed compared with several other multiple alignment tools such as CLUSTALW [12]-[13] or T-Coffee [14], by choosing maximum number of iterations and diagonal optimization.

In this study, all H1N1 nucleotide sequences from a specific segment were aligned collectively and then divided into three host groups for training the decision trees. The parameters – maxiters and –maxmb in MUSCLE were set to 2 and 250 MB, respectively. The first two iterations of the algorithm were performed to compromise between speed and accuracy. As for the user submitted sequences, only one iteration is set for prediction.

C. Phylogenetic tree analysis

Phylogenetic analysis is important to understand the evolution of species and of gene and protein families. Phylogenetic trees define the functional subfamilies within protein or gene families with multiple functions. Neighbor-Joining is a method for reconstructing phylogenies from a set of distances between each pair of sequences by successive clustering [15]. Neighbor-Joining can reconstruct trees with additive edge lengths without making the assumption that the divergence of the sequences occurs at the same constant rate at all points in the tree. The trees are generated using CLUTALW2.0 from EBI.

D. Machine learning techniques

The phylogenetic trees classify the sequences and describe the evolutionary relationships. However, inferring phylogenetic trees is difficult and these trees do not always show a clear picture of different clusters. In order to effectively achieve the goal of identifying evolutionary origin of pandemic influenza viral strains, we employed two machine learning approaches for classification analysis of the influenza A virus hosts. Machine learning techniques such as decision tree are able to find the subtle differences in the classifications associated with the real data. The information of critical sites
with sequence data is useful for host prediction of influenza A virus. Specifically, the decision tree approach classifies the sequences of different groups based on the nucleotide information at specific positions, whereas the support vector machine (SVM) approach classifies sequences based on the frequency of amino acids appearing in various sequences. This complementary strategy (one based upon DNA sequences and the other on protein sequences) provides a comprehensive description of influenza A viral classification. As to be described in a later section, results obtained from these two approaches are consistent to each other. Based on the classification results, a Web prediction tool was developed, where Hidden Markov Model (HMM) is used for modeling the informative positions generated from the decision trees. In the rest of this section we provide a brief review of these techniques.

E. Decision tree method

A decision tree is a simple but a powerful machine learning algorithm that has been successfully used for classification problems. The decision tree technique employs a supervised approach for classification, where the leaves on the tree represent classifications and the branches represent conjunctions of features that lead to classification. A series of decisions were made when classifying an instance from root to leaf nodes and the instance was classified to the one associated with the leaf node at the end of the traversal. Each internal node is a decision node and a value of given instance is compared to the decision function to decide which branch to follow. A decision tree is built using a training data set so as to reduce the average depth of each path from root to leaf node. Decision tree classification as a standard machine learning technique has been used for a wide range of applications in bioinformatics [16]-[17]. The software package Weka (Waikato Environment for Knowledge Analysis) (http://www.cs.waikato.ac.nz/ml/weka/), consisting of a number of machine learning algorithms, was used for decision tree analysis of aligned sequences [18]-[19]. The decision trees are generated using the C4.5 algorithm; Weka has its own version of C4.5 known as J48.

We used aligned sequences of each host to train decision tree classifiers in the J48 program of Weka, which allows the most informative nucleotide positions to be found. In each of the iteration steps, one or more critical positions, in which different subtypes can be most likely identified, were determined. These positions were collectively utilized to build HMMs for further host prediction. We applied the cross validation technique for testing. The three groups Human, Swine and Human_Swine strains are trained and then classified.

An example of decision tree is shown in Fig. 2, where the positions can be used to classify hosts, Human_Swine and Human. The position C604 is an input variable. The leaf nodes represent target variables Human_Swine and Human.

F. Support Vector Machine method

Support Vector Machine (SVM) is an alignment-free method which uses vectors to classify objects. One of the advantages of SVM is that it does not depend on multiple alignment, thus it can avoid errors, if any, in multiple alignment files. Another advantage of SVM is that it can classify sequences with low similarity and/or even with very short lengths.

In order to use SVM for classification, first we need to compute the frequency of each amino acid or a certain length of amino acids group. For example, sequence “GPPAV” can be treated in one letter a time (1-mer): “G” with frequency of 0.2 and “P” with frequency of 0.4, etc. Alternatively, it can be treated as two letters at a time (2-mer): “GP” with frequency of 0.25 and “PP” with frequency of 0.25, etc. These amino acid frequencies were treated as vectors, and the distribution patterns of k-mer (up to 3) amino acids were used for classification analysis.

Data from each segment of different origin were divided into two sets with approximately the same number of sequences: one for training and the other for testing. The sequences used for training are selected randomly by our computer program. The software SVM-light was used for classification analysis. Two experiments were conducted: one is trained and tested with Human and Swine sequences, and the other is trained with Human and Swine sequences, but tested with the Human_Swine strains.

G. Modeling using HMM

The Hidden Markov Model (HMM) is used for modeling the informative positions generated from the decision trees. An HMM is a statistical model representing sequences from a gene family. HMMs have a formal probabilistic basis, which is their advantage over other methods [20]. An HMM profile includes more flexible information on a given set of sequences than a single sequence. Therefore, database search methods using profiles is more sensitive to remote similarities than those based on pairwise alignments (e.g., regular BLAST). HMMER, a package that uses Hidden Markov models (HMMs) for sequence database searching, was used to build
HMM models based upon the most informative sites determined by the decision tree method [21]-[22].

To take advantage of most informative sites found by Decision Tree, we built different HMM profiles for the prediction of each influenza A virus host. HMM profiles are statistical representation of a specific group of informative sites. They were built using the program `hmmbuild` in HMMER. These profiles are used in the Web prediction program to determine the host of a viral strain through sequence comparison.

H. Summary of experiment design

Our experiment design is summarized in Fig. 3. In brief, the comparison analysis was conducted with three methods, decision tree, phylogenetic trees and support vector machine. The consensus of the three methods is that the Human_Swine sequences are more closely related to swine origin. Decision tree and HMM were used for web prediction of influenza viral origin.

![Fig.3. Experiment Design](image)

III. RESULTS & DISCUSSION

A. Phylogenetic analysis of HA and NA segments

The tree diagrams (Figures 4 and 5 in the next page) are made by using randomly selected 10 sequences from each group of Human strain, Swine strain and Human_Swine strains. Here, HA and NA segments are chosen as examples.

The sequences in the top right corner of the diagrams (marked by a box) are from Human host. The sequences in the bottom left of the diagram (marked by another box) are from Human_Swine. The rest of the sequences are from Swine host. As presented in the tree diagrams, they indicate that the origin of 2009 H1N1 human influenza A is from swine rather than traditional human H1N1. In order to have a more comprehensive and confident result, alternative approaches using machine learning techniques are needed, as to be described below.

B. SVM analysis

The results of the classification experiment showed above 95% accuracy at prediction and the test results details are summarized in TABLE II. Accuracy is defined as the number of correctly classified sequences divided by the total number of testing sequences. The results shown are the classification results when 3-mer is used and its frequency is computed as a vector input for SVM.

<table>
<thead>
<tr>
<th>Segment</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA</td>
<td>98.00%</td>
<td>97.37%</td>
<td>98.67%</td>
</tr>
<tr>
<td>M</td>
<td>97.97%</td>
<td>98.65%</td>
<td>97.33%</td>
</tr>
<tr>
<td>NA</td>
<td>95.30%</td>
<td>91.46%</td>
<td>100%</td>
</tr>
<tr>
<td>NP</td>
<td>98.67%</td>
<td>98.67%</td>
<td>98.67%</td>
</tr>
<tr>
<td>NS</td>
<td>98.65%</td>
<td>97.40%</td>
<td>100%</td>
</tr>
<tr>
<td>PA</td>
<td>97.97%</td>
<td>96.15%</td>
<td>100%</td>
</tr>
<tr>
<td>PB1</td>
<td>98.64%</td>
<td>100%</td>
<td>97.33%</td>
</tr>
<tr>
<td>PB2</td>
<td>99.32%</td>
<td>98.68%</td>
<td>100%</td>
</tr>
</tbody>
</table>

TABLE II

These results have shown that human and swine groups are well distinguishable, which indicate there are significant differences between them. The outcomes of the experiment that trying to predict the lineage of Human_Swine strain are presented in TABLE III, where 3-mer is used for all segments except PA. Since the results of 2-mer and 3-mer for PA segment differ significantly, the average of 2-mer and 3-mer results are presented instead. The percentages in the Table III indicate the percentages of Human_Swine sequences in the testing data sets that are classified as Swine strain. In other words, all sequences from HA, M, NA, NP, NS, PA, PB1, and PB2 are classified as swine influenza, which means sequences in these segments are more closely related to Swine strain. However, the result for PA segment is inconclusive; it is possibly equally similar to both human strain and swine strain.

<table>
<thead>
<tr>
<th>Segment</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA</td>
<td>100.00%</td>
</tr>
<tr>
<td>M</td>
<td>100.00%</td>
</tr>
<tr>
<td>NA</td>
<td>100.00%</td>
</tr>
<tr>
<td>NP</td>
<td>100.00%</td>
</tr>
<tr>
<td>NS</td>
<td>100.00%</td>
</tr>
<tr>
<td>PA</td>
<td>50.00%*</td>
</tr>
<tr>
<td>PB1</td>
<td>100.00%</td>
</tr>
<tr>
<td>PB2</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

* classified as Human strains by 2-mer SVM whereas classified as Swine strains by 3-mer SVM
Fig. 4. Neighbor-Joining tree of influenza A viral strains based upon HA sequences

Fig. 5. Neighbor-Joining tree of influenza A viral strains based upon NA sequences
C. Decision tree analysis

We have performed 25 iterations for HA, NA, M, NP, PA, NS, PB1, PB2 segments. During each iteration, a decision tree is generated with one or more informative positions which classify the three sets of data (i.e., Human, Swine and Human_Swine). Two thirds of the data is used for training and the remaining one third is used for testing. We applied the cross validation technique which automatically divides the input data into two sets, training and testing. TABLE V summarizes the average accuracy of classification between Human, Swine and Human_Swine strains for each segment of all iterations. A total of 67 nucleotide positions for the HA segment and 64 nucleotide positions for the NA segment have been identified as informative by the decision tree analysis. (We referred the positions determined by decision tree for classifying different groups of data as informative.) The informative positions of HA and NA segments are summarized in TABLE IV. The sequence data from only these positions are collected together to form HMM profiles. All the three groups are classified at a time using decision tree, unlike the case of SVM where every two groups are classified at once.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>HA Positions</th>
<th>NA Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27, 574, 671</td>
<td>47, 681, 728</td>
</tr>
<tr>
<td>2</td>
<td>28, 543, 408</td>
<td>48, 478, 421</td>
</tr>
<tr>
<td>3</td>
<td>29, 604, 625</td>
<td>124, 289, 837</td>
</tr>
<tr>
<td>4</td>
<td>331, 633, 634</td>
<td>281, 439, 793</td>
</tr>
<tr>
<td>5</td>
<td>404, 443, 500</td>
<td>267, 354, 721</td>
</tr>
<tr>
<td>6</td>
<td>597, 841, 1015</td>
<td>58, 293, 842</td>
</tr>
<tr>
<td>7</td>
<td>412, 725, 1036</td>
<td>654, 841, 994</td>
</tr>
<tr>
<td>8</td>
<td>590, 881, 944</td>
<td>176, 351, 1014</td>
</tr>
<tr>
<td>9</td>
<td>423, 623, 631</td>
<td>78, 112</td>
</tr>
<tr>
<td>10</td>
<td>37, 287, 448</td>
<td>731, 927, 1085</td>
</tr>
<tr>
<td>11</td>
<td>717, 857</td>
<td>493, 919</td>
</tr>
<tr>
<td>12</td>
<td>361, 580, 624</td>
<td>208, 643</td>
</tr>
<tr>
<td>13</td>
<td>867, 1111</td>
<td>585, 774, 1183</td>
</tr>
<tr>
<td>14</td>
<td>171, 873, 1126</td>
<td>89, 818, 1076</td>
</tr>
<tr>
<td>15</td>
<td>421, 556, 795</td>
<td>434, 1008</td>
</tr>
<tr>
<td>16</td>
<td>800, 846, 858</td>
<td>717, 982, 1126</td>
</tr>
<tr>
<td>17</td>
<td>32, 918, 1058</td>
<td>990, 1073</td>
</tr>
<tr>
<td>18</td>
<td>321, 446</td>
<td>256, 542, 1045</td>
</tr>
<tr>
<td>19</td>
<td>162, 721</td>
<td>92, 548, 978</td>
</tr>
<tr>
<td>20</td>
<td>221, 230</td>
<td>144, 662</td>
</tr>
<tr>
<td>21</td>
<td>187, 1156</td>
<td>593, 834</td>
</tr>
<tr>
<td>22</td>
<td>254, 553</td>
<td>784, 929</td>
</tr>
<tr>
<td>23</td>
<td>282, 440</td>
<td>789, 822</td>
</tr>
<tr>
<td>24</td>
<td>726, 730, 771</td>
<td>419, 1031</td>
</tr>
<tr>
<td>25</td>
<td>316, 364, 473</td>
<td>680, 936</td>
</tr>
</tbody>
</table>

D. A web prediction tool

In order to assist the task of detecting swine flu origin, we have developed a prediction tool for human influenza A virus hosts, which is now available at http://glee.ist.unomaha.edu/~pattaluri/swine/. The prediction tool has been developed using LAMP technology. We have provided sample data and a simple tutorial on how to use the tool for host prediction. The tool allows users to submit sequences, choose the input segment, and select options to view the result. The given sequence is aligned and the informative positions are extracted to compare with the HMM profiles. The result of prediction will be displayed based on matching scores.

Here we show how to use the web prediction system and what the result page looks like by analyzing a real sequence. We downloaded a Human_Swine (outbreak) sequence from NCBI, copied and pasted the sequence in text area, selected nucleotide type, checked selected options, and clicked Submit button. The prediction result is shown as in Fig. 6, where the sequence is predicted correctly as the Swine type, with a score 116.8 and an E value $4.2 \times 10^{-31}$.

![Fig. 6. Influenza A virus host prediction system](image)

The classification results from our experiments indicate that data from Human and Swine origin are easily distinguishable as they have considerable positions for predicting the host. Also the Human_Swine sequences in the testing data sets are classified as Swine strain. This suggests that the outbreak sequences be more closely related to Swine rather than Human viral strains.
IV. CONCLUSION

Accurate detection of influenza viral origin can significantly improve influenza surveillance and vaccine development. In this research, we applied machine learning techniques to identify the evolutionary origin of the latest human pandemic influenza H1N1 viral strains. Phylogenetic analysis of randomly selected sequences revealed significant differences between human and swine influenza A H1N1 viral strains. Both Support Vector Machine and decision tree methods agreed each other on that the viral strains causing the latest human pandemic are of swine origin. Along with our previous findings [23], this study demonstrated the power of integrating the decision tree and hidden Markov model approaches in classifying influenza A viral subtypes and hosts.

ACKNOWLEDGMENT

This publication was made possible by NIH grant numbers P20 RR016469 from the INBRE Program of the National Center for Research Resources and R01 LM009985-01A1. We are grateful to Dr Ruben Donis of the Centers for Disease Control and Prevention (CDC), US Department of Health and Human Services for constructive discussion. The authors also acknowledge the UCRCA, the University of Nebraska at Omaha, for continuous funding support.

REFERENCES

Application of Committee $k$NN Classifiers for Gene Expression Profile Classification

Manik Dhawan  
Department of Computer Science  
University of Akron  
Akron, OH, USA 44325  
Email: manik.dhawan@gmail.com

Sudarshan Selvaraja  
Department of Computer Science  
University of Akron  
Akron, OH, USA 44325  
Email: sudarshanselvaraja@gmail.com

Zhong-Hui Duan  
Department of Computer Science  
University of Akron  
Akron, OH, USA 44325  
Email: duan@uakron.edu
Differential expression analysis of digital gene expression data: RNA-tag filtering, comparison of $t$-type tests and their genome-wide co-expression based adjustments

Yinglei Lai
Department of Statistics and Biostatistics Center
The George Washington University
2140 Pennsylvania Avenue, N.W., Washington, D.C. 20052
Email: ylai@gwu.edu
Scaling Alignment of Large Ontologies

Suzette K. Stoutenburg
and Jugal Kalita
School of Engineering and Applied Science
University of Colorado at Colorado Springs
Colorado Springs, Colorado 80918
Email: suzette@coloradostoutenburg.com

Kaily Ewing
and Lisa Hines
The College of Letters, Arts and Sciences
University of Colorado at Colorado Springs
Colorado Springs, Colorado 80918
Email: kwing@uccs.edu

Email: kalita@eas.uccs.edu Email: lhines@uccs.edu
Multiple genome sequence alignment with longest path algorithms

Fangrui Ma and Jitender Deogun
Computer Science and Engineering Department
University of Nebraska-Lincoln
Lincoln, NE 68588
Email: fma{deogun}@cse.unl.edu
Accepted Extended Abstracts
Association Analysis in Structured Plant Populations, an Adaptive Mixed LASSO Approach

Dong Wang
Department of Statistics
University of Nebraska
Lincoln, NE 68583
Email: dwang3@unl.edu

I. INTRODUCTION

Recently, there has been heightened interest in performing association analysis in important crop species for its significant potential in dissecting complex traits by utilizing diverse mapping populations. Since population structure is a universal feature for all crop breeding populations, accounting for genetic relatedness between lines in a mapping population is essential for obtaining valid p-values. The development of mixed linear models for plant association mapping has significantly advanced the statistical methodology in this field by providing a general approach for controlling the type I error rate as it takes into account existing population structures. Studies in various plant species have demonstrated that association mapping is a viable approach in detecting marker phenotype associations when proper attention is paid to experimental designs and data analysis. However, in comparison to traditional linkage mapping, important aspects of the association mapping methodology have not been extensively studied. The mixed linear model is currently limited to single marker analysis. But methodologies have not been extensively studied. The mixed linkage mapping, important aspects of the association mapping methodology have not been extensively studied. The mixed linear model is currently limited to single marker analysis. But many complex traits of agronomical importance are regulated by multiple quantitative trait loci (QTLs), their interactions (epistasis), environmental factors, and gene by environment (G×E) interactions [1]. As a result, the lack of knowledge on epistasis and G×E interactions has become one of the major impediments of utilizing genomic information for crop improvement.

In this talk, we report the development of the adaptive mixed LASSO (least absolute shrinkage and selection operator) method that can incorporate a large number of predictors while simultaneously accounting for the population structure. Since its introduction [3], LASSO has attracted huge interest as a method simultaneously performing variable selection and estimation. LASSO can deal with situations where the number of explanatory variables is much larger than the sample size, which is not feasible for traditional regression methods. LASSO and a number of related methods have been successfully applied to generalized linear models, survival analysis, graphical models, latent factor models, among others. LASSO is also related to a class of shrinkage models in the Bayesian framework [2] and to boosting methods in machine learning. A related method, the adaptive LASSO [5], applies different penalty to each coefficient and thus leading to the desirable amount of shrinkage. By extending adaptive LASSO to include random effects for structured populations, we developed the adaptive mixed LASSO method with applications to plant association mapping.

II. ADAPTIVE MIXED LASSO

The adaptive mixed LASSO method is a unified framework for incorporating a large number of predictors (genetic markers, epistatic effects, environmental covariates, and G×E interactions) while accounting for population structures in plant association analysis. For illustration, consider a case where we want to identify significant QTL main effects and epistatic effects in a simple association study using inbred lines derived from accessions of a certain crop species. In the following linear mixed effects model,

\[ y_{ij} = x_{ij}^T \beta + u_i + \epsilon_{ij}, j = 1, \ldots, m, i = 1, \ldots, m, \]

\[ \beta = (\beta_1, \ldots, \beta_p)^T \] is the p dimensional coefficient vector for fixed effects (markers and epistatic effects), \( u_i \) is the random effect of the ith inbred line, and \( \epsilon_{ij} \) is the independent error term. Alternatively, we can write it in the matrix form, \( y = X\beta + Zu + \epsilon \). Here the matrix X is the “design” matrix corresponding to fixed effects, matrix Z represents an incidence matrix that designates each of the n individuals in the sample to one of the m inbred lines. Following usual assumptions of linear mixed models, we assume that \( u \) and \( \epsilon \) are Gaussian vectors with \( u \sim N(0, \sigma_u^2K) \) and \( \epsilon \sim N(0, \sigma_e^2I) \), where K is the relationship matrix for the inbred lines, I is the \( n \times n \) identity matrix, and \( \sigma_u^2 \) and \( \sigma_e^2 \) are variance components associated with inbred lines and residual errors respectively. The difficulty here is that because there could be a large number of genetic markers and epistatic effects, we are likely to have a very large p relative to n, which will make traditional methods ineffective (or totally inapplicable when \( p > n \)). To solve this problem, we apply the weighted LASSO penalty to coefficients associated with genetic markers and epistatic effects to shrink most coefficients to zero. Specifically, the estimator of parameters is obtained by minimizing the following quantity,

\[ \Omega = (y - X^T \beta)^T \Sigma^{-1} (y - X^T \beta) + \lambda_n \sum_{k=1}^p w_k |\beta_k|, \]
where $\Sigma = \sigma_g^2 K + \sigma_e^2 I$ is the covariance matrix of $y$, $\lambda_n$ is the tuning parameter for controlling the amount of shrinkage, $w_k$ is the weight for the $k$th coefficient such that each coefficient will receive different amount of shrinkage. We have proved that the adaptive mixed LASSO estimator for $\beta$ is consistent under mild conditions. Here consistency means that as the sample size increases, the adaptive mixed LASSO method tends to select the correct model (right markers and epistatic effects); at the same time, the magnitude of the estimated values of parameters will converge to the true value. In practice, both $\beta$ and variance components ($\sigma_g^2$ and $\sigma_e^2$) have to be estimated simultaneously and we have developed algorithms to iteratively estimate the regression coefficients and variance components with the number of covariates in the final model determined by the Bayesian information criterion (BIC).

### III. Numerical Results

To demonstrate the property of the adaptive mixed LASSO, we performed preliminary simulation studies using the actual relationship matrix in a maize mapping population [4]. We retrieved data of flowering time and marker (SNP) information on 277 maize inbred lines and carried out a simulation study with three QTL main effects as well as four epistatic effects. Here we use the actual relationship matrix of the maize breeding population described above, as well as 63 SNP markers. Three add-on main QTL effects account for 8%, 10%, and 20% of the phenotypic variation, while the four epistatic effects account for 5%, 8%, 10% and 12% of the phenotypic variation respectively. Two of the epistatic effects involve one marker each with QTL main effect, while the other two epistatic effects are from two markers with no main effect. The adaptive mixed LASSO is then performed with 63 markers and 1953 epistatic effects for a total of 2016 predictors. Among 2000 simulation runs, the selected model by the adaptive mixed LASSO includes all seven actual genetic effects for 99.1% of the runs. In the other 0.9% of simulation runs, the selected model includes six of the seven genetic effects. Of the models including all seven effects, 52% identified the correct model exactly (selecting no other effects), 32% have one or two extra effects in addition to the seven correct effects, the others have three to five extra effects. Figure 1 shows the estimate of genetic effects in one simulation run.

### IV. Conclusion

Our results show that the adaptive mixed LASSO method is very promising in modeling multiple genetic effects (main QTL effects and epistasis) as well as modeling gene by environment interactions when a large number of markers are available and the population structure cannot be ignored. Since no equivalent method has been proposed in the setting of crop association analysis, it is expected to have a significant impact on the study of complex traits in important crop species. Applications to actual data sets from wheat breeding programs has been planned with the potential of influencing plant breeding practices.

### Acknowledgment

The author thanks Drs. K. Eskridge and P.S. Baenziger for valuable discussions.

### References

A Dynamic Programming based Algorithm to Identify Coordinately Differentially Activated Transcription Factors

Haiyan Hu
School of Electrical and Computer Engineering
University of Central Florida
Orlando, Florida 32816-2005
Email: haihu@cs.ucf.edu

Xiaoman Li
Burnett School of Biomedical Science
University of Central Florida
Orlando, Florida 32816-2005
Email: shawnli@mail.ucf.edu

Computational identification of transcription factor (TF) activities associated with a certain experimental condition is important to understanding gene regulation and dynamic cellular mechanism. Because microarray experiments can measure hundreds of thousands of genes’ expression under many experimental conditions, recent studies often attempt to identify activities of a TF through detecting co-expression of its target genes by assuming the equivalence between co-regulation and co-expression. However, even with rapidly accumulated massive amounts of microarray gene expression data and TF binding data, it remains a challenge to identify differentially activated TFs due to at least the following two issues: i) TFs are often regulated at the post-transcription level, thus it is often hard to identify the TF activities by directly measuring the expression change of their corresponding mRNAs; ii) Multiple TFs can be activated coordinately during the same biological process, thus methods to identify TF activities one at a time can generate inaccurate results.

In this paper, we propose a novel algorithm to resolve the above issues and to identify differentially activated TFs from microarray expression data and known TF information. We downloaded 615 TFs and their target gene candidates from mSigDB database [1]. This TF target gene data is collected from curated data in the TRANSFAC database [2] and/or predicted by comparative genomics approaches [3]. Note that the downloaded TF target genes are candidate target genes because a TF may bind with different subsets of its target gene candidates under different conditions.

We determine differentially activated TFs from their differentially expressed target genes. Instead of explicitly define differential expression, we propose to use gene ranking based methods to estimate differential expression [1, 4, 5]. Our problem is then formulated as given a gene list ranked by gene expression difference between two conditions, to identify the smallest subset of TFs that can best explain the expression changes of as many top ranked target genes as possible.

To identify the minimal subset of coordinately differentially activated TFs in the above problem is difficult. First of all, even if target genes of an activated TF are most likely to be differentially expressed, without a way to explicitly define differentially expressed genes, it is hard to determine which genes are the target genes of a TF under the given experimental conditions. Besides, because multiple TFs can be activated coordinately under the given conditions, methods that consider individual TF activities one at a time or dummy methods that simply choose several TFs with most target genes in top ranked genes will only obtain suboptimal solutions.

We design a novel algorithm to identify coordinately differentially activated TFs as following. We first determine for the top ranked $k$ genes, which TFs may have significantly larger number of target genes in the $k$ genes compared with the rest of the genes. In this way, TF with large number of target genes uniformly distributed in the ranked gene list will be filtered out, and only those TFs with their target genes distribution bias toward top ranked genes will be kept for the next step. We then perform a dynamic programming procedure to determine the maximal number of genes in the top $k$ genes that are the target genes of different combination of given TFs. The minimal number of TFs that can explain as many genes among the top $k$ as possible will then be chosen using a Poisson process based statistical method. We repeat the process by permuting $k$ from 1 to $\kappa$ to identify the best subset of TFs.

We tested our algorithm on two microarray datasets: the essential thrombocythemia (ET) data set [6], and the breast cancer data set [7]. For these microarray data sets, we first classified all the samples into two relevant conditions. We next utilized differential t test to test whether a gene is differentially expressed between the two relevant conditions. The differential test is able to assign each gene a p-value. All the genes are then ranked according to this p-value. We downloaded TF target gene sets from mSigDB [1]. Each of these gene sets contain genes that share a TFBS defined in the TRANSFAC, and each of these gene sets is annotated by a TRANSFAC record. There are in total 615 TF target gene sets collected.

We found many literature-supported TFs involved in the corresponding disease conditions.
For ET data, we identified three known TFs corresponding to heat shock transcription factor 1 (HSF1), Hypoxia-inducible factor 1 (HIF-1) and GATA binding protein (GATA). We further extract the identified TFs and their target genes in the protein-protein interaction network downloaded from HPRD. The extracted network shows that CREB binding protein (CREBBP) can interact with all three TFs. Since CREBBP is an important member in JAK-STAT pathway, and the given phenotype is related to JAK2 mutation, this result supports that the association of identified TFs with the given phenotype.

For breast cancer data, we identified three significant TFs corresponding to zinc finger E-box binding homeobox 1 (AREB6), interferon regulatory factor 1 (IRF1) and lymphoid enhancer binding factor 1 (LEF1). We also extract the identified TFs and their target genes in the protein interaction network downloaded from HPRD. In the extracted network, three TFs are connected by a dense protein subnetwork centering on estrogen receptor 1 (ESR1) gene, which is well known to be involved in pathological processes of breast cancer.

We also attempt to compare with other method. Currently, no existing methods can be directly applied to identify a subset of TFs which are coordinately differentially activated under specific conditions. However, with a list of genes ranked by their phenotype relevance such as expression difference between two conditions, for an input gene set, gene set enrichment methods such as GSEA can rank given gene sets in terms of their phenotype-relevance [1] by comparing the number of differentially expressed genes found in a input gene set with the number of genes expected to be found in the input gene set just by chance. By considering all the target gene candidates corresponding to one transcription factor as one gene set, we are able to apply GSEA methods to test the enrichment of a set of target genes corresponding to a TF.

As a result, we found no significantly enriched TF target gene set in general. For example, for the ET dataset, all three target gene sets corresponding to TFs HSF1, HIF1 and GATA are assigned FDR q-value=1. TFs HSF1 obtained a p-value=0.083, HIF1 gained a p-value=0.02268 and GATA’s p-value=0.2633.

In summary, we have developed a novel algorithm, to simultaneously identify a subset of TFs that are activated under a given experimental condition. In comparison with other methods, our approach has two unique advantages. One advantage is that, by considering activities of TFs simultaneously, our approach can systematically identify coordinately activated TFs. It is important to incorporate the interdependency into identification process to provide insight into TF regulatory mechanisms. The other advantage is that, as a side product, our program also identifies TF target genes due to the coupling of identifying activated TFs and identifying their differentially expressed target genes. TF binds to different target gene candidates under different conditions. Our algorithm defines TFs from top ranked target gene candidates and defines top ranked TF target genes from identified activated TFs. Thus, when we output the coordinately differentially activated TF subset, we can simultaneously output the top ranked target genes used during the identification process.

REFERENCES


Molecular Evolution of Sterol-Sensing Domain in Eukaryotes

Pooja K. Strope
School of Biological Sciences
University of Nebraska, Lincoln, Nebraska 68588
Email: poojastrope@gmail.com

Etsuko N. Moriyama
School of Biological Sciences,
Center for Plant Science Innovation
University of Nebraska, Lincoln, Nebraska 68588
Email: emoriyama2@unlnotes.unl.edu

I. INTRODUCTION

The sterol-sensing domain (SSD) is a ∼180 amino acid long region that is conserved in six families of proteins such as hydroxymethylglutaryl-CoA reductase (HMDH), SREBP (sterol regulatory element binding protein) cleavage activating protein (SCAP), Niemann-Pick C-1 type protein (NPC1), Patched, Patched-related and Dispatched [1]. This domain encompasses five transmembrane helices and is involved in sterol-level sensing in the cell. All these classes of proteins thus have functions related to sterols. HMDH is a cholesterol biosynthetic enzyme that is degraded when sterol levels are high. SCAP is responsible for regulating SREBP, a transcription factor of cholesterol biosynthetic genes. NPC1 is responsible for intracellular transport of cholesterol. Patched plays a role in cell differentiation during development and morphogenesis. It is a receptor of hedgehog, a ligand that is bound to cholesterol. Dispatched is involved in releasing the cholesterol-bound hedgehog. The importance of SSD in proper functioning of these proteins has been shown by various mutation experiments. For example, mutations in the SSD region of NPC1 disrupt normal transportation of cholesterol in the cells [2]. Mutation in SSD can be lethal to cells and cause various diseases due to the abruption of cholesterol homeostasis in cells. The role of SSD in sterol homeostasis in cells makes it an important target for bio-medical research in understanding and curing cholesterol-related diseases.

Despite its importance, SSD has not been thoroughly studied in its own entirety. In order to elucidate the molecular evolution of SSD and related protein families, in this study we examined SSDs in various eukaryotic species as well as in different protein families. Existence of HMDH proteins with and without SSDs, for example, provided us with an excellent opportunity to study how their functions and domains have been acquired during the evolution of this protein family.

II. MATERIALS AND METHODS

The annotated SSD-containing proteins were gathered from the UniProt database [5] and the SSD region of these sequences were extracted. These subsequences show clear clustering of different protein families as shown in Fig. 1. Non-redundant set of these subsequences were aligned using ClustalW [6] and a profile hidden Markov model (HMM) was built using the HMMER [7] package. Profile HMMs are probabilistic models of the multiple alignment of a group of related sequences and this method works well in classifying members of a well-conserved protein family or domain. Various other alignment-free methods including support vector machines and partial least squares utilizing amino acid compositions, physico-chemical properties were also tried. However, the results from these methods were not as good or better than HMMs (P.K.S. and E.N.M., unpublished data, 2009).

The eukaryotic species searched included 19 fungal, 6 plant, 3 green algae, 1 amoeba (Dictyostelium discoideum), 1 choanoflagellate (Monosiga brevicollis), and 8 animal species. The predicted protein sets of each species was searched using...
Fig. 2. The maximum-likelihood phylogenetic tree of full sequences of HMDH hits. The red colored sequences have SSD, whereas the black colored sequences lack SSD.

the profile HMM. Results were gathered using an e-value threshold of 0.001. This e-value threshold was determined after searching in annotated proteome of human. The identified SSD regions from these species were aligned including the 59 training subsequences using MAFFT [3]. Phylogenetic trees were reconstructed with the maximum-likelihood method using RAXML [4]. This allowed us to identify which different families (HMDH, NPC1, Patched, Patched-related, SCAP or Dispatched) do the hits belong to.

Not all of HMDH proteins have SSDs. In order to find all HMDH proteins, both SSD-containing and non-SSD types, another profile HMM was built using non-redundant set of 46 HMDH sequences and was used to search in the species mentioned before. To further ascertain that HMDH was really absent in certain cases, we checked for the presence or absence of upstream and downstream enzymes to the HMDH enzyme in the sterol biosynthetic pathway. These enzymes are the hydroxymethylglutaryl-CoA synthase (HMCS) and the mevalonate kinase (MK) [8]. Separate HMMs were built using these families of proteins and searched in the proteomes.

III. RESULTS AND DISCUSSION

As shown in Fig. 2, all plant HMDHs lack SSD. All fungal HMDHs have SSD except in Ustilago maydis. Some fungal species such as Rhizopus oryzae, Aspergillus nidulans, and Aspergillus oryzae have two copies of HMDHs, one with SSD and the other without SSD. Three closely related Saccharomyces species, S. bayanus, S. cerevisiae, and S. para-
doxus have two copies of HMDHs, both with SSD. Human, pipid frog (Xenopus tropicalis), Drosophila, Branchiostoma floridae and Trichoplax adhaerens have HMDHs with SSD. The HMDHs of C. elegans, Dictyostelium discoideum and Monosiga brevicollis lacked SSD. No HMDH was found in Ciona intestinalis and Nematostella vectensis or the 3 green alga.

HMCS enzyme was found in all organisms except Chlamydomonas reinhardtii. MK enzyme was absent from the three green alga, T. adhaerens and C. intestinalis, while being present in the rest of the organisms. The presence of HMDH in organisms where HMDH and MK are absent indicates that these organisms may have HMDHs that are extremely diverged from the currently known HMDH sequences, and therefore they were not identified by our model. It is evident from Fig. 2 that HMDH acquired SSD at a common ancestor of animals and fungi. Some fungal HMDH seems to have lost this domain over time.

Among the other SSD-containing proteins, NPC1 was present in almost all organisms. SCAP was present in only 9 fungal and most animal species while being completely absent from plant and green algal species. Patched and Patched-related proteins were found in 2 green algae, 1 moss, all animals, and a choanoflagellate while being absent from all other plants, all fungi and also from D. discoideum. Dispatched was absent from all fungi, all plants and green algae, and present in only 6 animal species.

The presence of SSD in all the euakaryotic organisms that we examined (except in maize, whose genome sequencing may not be complete) indicates that SSD existed before the euakaryotic divergence. Remote similarity between SSD and some bacterial proteins have been found but further analysis needs to be done to ensure that they are truly evolutionarily related.

REFERENCES

Simulating Sequence Superfamilies for Biological Hypothesis Testing

Cory L. Strope  
Computer Science and Engineering  
University of Nebraska  
Lincoln, Nebraska 68588–0115  
Email: corystrope@gmail.com

Stephen D. Scott  
Computer Science and Engineering  
University of Nebraska  
Lincoln, Nebraska 68588–0115  
Email: sscott@cse.unl.edu

Etsuko N. Moriyama  
School of Biological Sciences, Center for Plant Science Innovation  
University of Nebraska  
Lincoln, Nebraska 68588–0660  
Email: emoriyama2@unlnotes.unl.edu

I. INTRODUCTION

Sequence simulation is an important tool in validating biological hypotheses as well as testing various bioinformatics and molecular evolutionary methods. Hypothesis testing relies on the representational ability of the sequence simulation method. Simple hypotheses are testable through simulation of random, homogeneously evolving sequence sets, which is akin to a closely related group, e.g., sequence families. However, testing complex hypotheses, e.g., highly-diverged sequences with domain or lineage-specific evolution, requires more complex simulation models. As many genome projects gather more biological sequences, such complex hypothesis testing is rapidly becoming the norm rather than the exception in biological analyses; in other words, analyses and methods that were originally used for sequence families now must be extended to sequence superfamilies. Methods are being introduced to assist in understanding and drawing information from such superfamily-level analyses. However, validation of such methods has been either weak or lacking entirely. This is because sequence simulation methods available now are not capable of simulating the evolution of superfamilies of sequences.

Superfamilies of sequences are those that have diverged over a long evolutionary time, where dynamic evolutionary events such as insertion and deletion (indels), in addition to substitutions, have occurred. Though they are distantly related, remote similarities can be inferred through examining their 3-dimensional structure and their related (though often different) function. Simulation methods therefore must preserve structural elements, and allow functional elements to independently change between subtrees (lineages). While many indel-capable sequence simulation methods exist [1], [2], [3], [4], [5], these methods are unable to preserve structural or functional elements. indel-Seq-Gen version 1.0 (iSGv1.0) [6] was created in order to conserve structural elements, by allowing subsequences to be simulated using, e.g., input of different amino acid frequencies for the substitution processes to model different cellular environments. Global conservation of functional regions was also added by disallowing indel occurrence in such regions.

We extend superfamily evolution with indel-Seq-Gen version 2.0 (iSGv2.0) [7], which incorporates heterogeneous protein and DNA evolution, advanced functional site conservation using PROSITE-like regular expressions, lineage-specific evolution, and subsequence length constraints. We also output indel event information that can be used to validate the performance of methods that infer indel events, e.g., multiple sequence alignment methods. Finally, we uncover a flaw in the modeling of indels in state of the art simulation methods that changes the simulation results depending on the number of branching events during the evolutionary path.

II. METHODS

To validate the ability of iSGv2.0 to simulate sequence superfamilies, we chose to model the lipocalin superfamily. The lipocalin superfamily is highly diverged globular protein superfamily sharing a conserved structure of 8 β-strands that form a β-barrel. We used a template multiple alignment of 23 lipocalin sequences from an evolutionary study done by Sánchez et al. [8] to obtain a guide tree and to calculate the amino acid frequencies for β-strand and coil regions. We modeled the sequence structures by setting the amino acid frequencies for each region as calculated. We also parameterized this alignment for SIMPROT, ROSE, and Seq-Gen [9] (the substitution engine of iSGv2.0). We simulated 5 datasets with each method, predicted the structures of the resulting sequences using PSIPRED [10], and plotted the percentage of sites predicted as β-strands.

To examine if the indel simulation is done properly, we tested seven indel-capable sequence simulation methods: EvolveAGene3 [5], ROSE [1], DAWG [2], SIMPROT [4], iSGv1.0 [6], iSGv2.0 [7], and MySSP [3]. We simulated 1000-character long sequence datasets for each simulation method, using 4 guide trees with a varying number of branching nodes (0, 1, 3, and 7) and the total tree length of 8 substitutions per site. Using length 4 indels, with 1 indel occurring every 50 substitutions, we simulated datasets using (i) only insertions and (ii) only deletions. For each of the 4 guide trees, we then examined the resulting true alignments to determine if the methods were affected by the number of branching nodes.
II. RESULTS AND DISCUSSION

Figure 1 shows the results of the protein sequence superfamily simulation. iSGv2.0 outperforms all other protein capable simulation methods, where β-strand sites are correctly predicted at 37% accuracy, 5% better than the next best method, although slightly more coil sites are also predicted as β-strands. The correct predictions in the reference alignment far surpass those from the simulation methods, suggesting that there is room for improvement.

Figure 2 presents the effects of the flawed indel models using a variety of parameters. iSGv2.0, EvolveAGene3, and DAWG correctly produce consistent simulation results regardless of the number of branching points in the guide tree. ROSE, iSGv1.0, SIMPROT, and MySSP, which do not adjust the length of the sequence after each indel, produce differing results depending on the number of branching points in the guide trees.

Its unique abilities to preserve structural and functional elements, create lineage-specific functional divergence, and model indels properly, allow iSGv2.0 to simulate the necessary range of events seen in the highly divergent evolution of sequence superfamilies.
Oligonucleotide Optimization for DNA Synthesis

Abstract
Tobias M. Louw, Scott E. Whitney, Joel TerMaat and Hendrik Viljoen
Department of Chemical and Biomolecular Engineering
University of Nebraska
Lincoln, Nebraska 68588-0643
Email: tobilouw@gmail.com

I. INTRODUCTION
DNA molecules from hundreds to thousands of base-pairs in length can be de novo synthesized using the polymerase chain reaction. The DNA constructs are assembled from smaller oligonucleotides (single stranded DNA molecules, typically with a length of up to 50 base-pairs) which overlap to form the complete gene sequence.

The main difficulty in DNA synthesis is in designing the set of oligonucleotides used so that certain criteria are met. Specifically, the melting temperature amongst oligonucleotide hybridization pairs must not vary greatly and the length of each oligonucleotide must not exceed fifty nucleotides to reduce cost as well as synthesis errors.

To make use of reliable gradient optimization methods, an objective function can be derived from the DNA sequence to be synthesized. This objective function is minimized using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method to find the optimal oligonucleotide set.

II. MELTING TEMPERATURE
Consider (1 – 3), used to calculate the melting temperature of a DNA sequence:

\[ T_m = \frac{\Delta H^0}{R \ln C + \frac{\Delta H^0 - \Delta G^0}{310.15} + \Delta H^0 f_{corr}} \quad \cdots (1) \]

\[ \Delta H^0 = \Delta H^0_I + \sum_{k=1}^{n-1} \Delta H_{NN}^k + \Delta H_I^p \quad \cdots (2) \]

\[ \Delta G^0 = \Delta G_I^0 + \sum_{k=1}^{n-1} \Delta G_{NN}^k + \Delta G_I^p + \Delta G_{sym} \quad \cdots (3) \]

The symbols retain their standard thermodynamic interpretation. The subscript \( NN \) indicates “nearest neighbor” enthalpy and Gibbs free energy, whereas \( I \) indicates “initial”. The integer \( n \) is the sequence length (in nucleotides). \( f_{corr} \) is a correction factor to include the effects of magnesium, monovalent salts and GC-content on the Gibbs free energy. \( C \) is the oligonucleotide concentration (of each oligonucleotide type, assuming they are in equal concentration) in moles/liter.

In order to choose the optimal oligonucleotide set, it is desirable to predict the melting temperature of a specific subset of the DNA sequence. If the global parameters such as the salt, magnesium and oligonucleotide concentrations are known, one can determine the melting point of any subset of the sequence, given the start and end point of each piece \( (x_1 \text{ and } x_2) \). Let us first create a function named \( H_I(x) \) (4):

\[ H_I(x) = \begin{cases} 0.42 \ \text{kJ/mole} & x \rightarrow \text{G/C} \\ 9.63 \ \text{kJ/mole} & x \rightarrow \text{A/T} \end{cases} \quad \cdots (4) \]

The value of \( H_I(x) \) depends solely on whether the sequence has a G/C or A/T base-pair at position \( x \) where \( x \in \{0, 1, 2, \ldots, n\} \) and \( N \) is the total number of base-pairs in the sequence.

To determine the enthalpies contributed by the sets of nearest neighbors, we set up a similar function, named \( H_{NN}(x) \) (5):

\[ H_{NN}(x) = \begin{cases} 0 & x = 1 \\ \sum_{k=2}^{x} \Delta H_{NN}^k & x \geq 2 \end{cases} \quad \cdots (5) \]

Notice that \( H_{NN}(x) \) is the sum of all the nearest neighbor pairs up to \( x \). We can therefore determine the nearest neighbor enthalpies between \( x_1 \) and \( x_2 \) by subtracting \( H_{NN}(x_1) \) from \( H_{NN}(x_2) \). The total enthalpy can now be calculated using (6):
\[ \Delta H^0(x_1, x_2) = H_f(x_1) + H_{NN}(x_2) - H_{NN}(x_1) + H_f(x_2) \quad \ldots (6) \]

The same can be done to determine \( \Delta G^0 \).

The correction factor \( f_{corr} \) is calculated using an empirical formula based on the magnesium and monovalent salt concentration as well as the GC-content of the sequence subset. The fractional GC-content is calculated using (7–8):

\[ GC(x) = \sum_{k=1}^{x} g_c(k) = \begin{cases} 1 & k \to G/C \\ 0 & k \to A/T \end{cases} \quad \ldots (7) \]

\[ f_{GC}(x_1, x_2) = \frac{GC(x_2) - GC(x_1 - 1)}{x_2 - x_1 + 1} \quad \ldots (8) \]

We will define derivatives for each of the functions defined by (5–7) as follows (9):

\[ \frac{\partial \psi(x)}{\partial x} = \frac{\psi(x - 1) + \psi(x + 1)}{2} \quad \ldots (9) \]

Where \( \psi \) is either one of the functions listed above. As each function is differentiable in terms of \( x \), the derivative of the melting temperature of the subset defined by \( x_1 \) and \( x_2 \) exists in terms of \( x_1 \) and \( x_2 \) by the chain rule.

The melting temperature of a subset of a DNA sequence can now be calculated using a continuous, differentiable function of \( x_1 \) and \( x_2 \).

III. OPTIMIZATION OF OLIGONUCLEOTIDE SET

A DNA sequence of total length \( N \) can be divided into \( n + 1 \) overlaps. These overlaps can be defined by a \([\{n-1\} \times 1]\) vector \( \mathbf{x} \), each with a melting temperature \( T_i \). Most oligonucleotide consists of two overlaps, thus the length of each oligonucleotide can be calculated by (10):

\[ L_i = \begin{cases} x_{i} & i = 1, 2 \\ x_{i} - x_{i-2} & 3 \leq i \leq n \\ N - x_{i-2} & i = n, n + 1 \end{cases} \quad \ldots (10) \]

The oligonucleotide set will be chosen by optimizing the following objective function (11):

\[ O(\mathbf{x}) = W_1 (\sum (T_{ave} - T_i)^2) + W_2 \left( \sum e^{L_{max}/2} \right) \quad \ldots (11) \]

Where \( W_1 \) and \( P_1 \) are weights used to fine tune the objective function. The oligonucleotide lengths are constrained by \( L_{max} \). Equation (11) is continuous and differentiable and can be optimized using the BFGS method.

IV. RESULTS

The \( rpoB \) DNA sequence with a length of 3519 base-pairs (64.3% GC-content) was optimized as described above. The overlap melting temperatures are shown in Fig. 1 (\( \mu_T = 61.4^\circ C, \sigma_T = 1.0^\circ C \)). The oligonucleotide lengths ranged from 24 to 42 nucleotides \( (\mu_L = 30.9, \sigma_L = 3.2) \).

Fig. 1. Overlap \( T_m \) for \( rpoB \) sequence \( (n = 3519) \)

REFERENCES


Molecular Evolution of G-Proteins

Brandon L. Fulk
School of Biological Sciences
University of Nebraska - Lincoln
Lincoln, Nebraska 68588-0118
Email: brandon.fulk@gmail.com

Stephen O. Opiyo
School of Biological Sciences
University of Nebraska - Lincoln
Lincoln, Nebraska 68588-0118
Email: sopiyo@unlserve.unl.edu

Etsuko N. Moriyama
School of Biological Sciences
Center for Plant Science Innovation
University of Nebraska- Lincoln
Lincoln, Nebraska 68588-0118
Email: emoriyama2@unlnotes.unl.edu

I. INTRODUCTION

G Protein Coupled Receptors (GPCRs) are seven transmembrane proteins that provide some of the most important signaling pathways for plants and animals via guanine nucleotide-binding proteins (G-proteins). The disparity of repertoire size is noted in the number of GPCRs associated with these G proteins. The human genome has approximately 800 GPCRs whereas there have only been 24 predicted in the Arabidopsis thaliana genome [15].

Figure 1 shows the common model for G-protein signaling cycle. Gα, Gβ, and Gγ make up the cytosolic portion of this model and the GPCR makes up the transmembrane portion. Once a ligand comes in contact with the GPCR, it causes the GDP-bound Alpha protein to exchange its GDP for a GTP. From this point, the Alpha protein disassociates from the Beta-Gamma dimer and the GPCR and effector proteins interact with these free proteins in the cytosol. To complete the cycle, Alpha exchanges its GTP for GDP and it reassociates with the GPCR and a Beta-Gamma dimer.

The evolutionary history of the plant and animal G proteins is also different. While animals show multiple gene duplication events among all G proteins, such is not the case in plants. The genome of Homo sapiens contains 23 Alpha, 5 Beta and 12 Gamma G proteins. This is in contrast to the genome of Arabidopsis thaliana that has 1 Alpha, 1 Beta and 2 Gamma G proteins [8].

The objective of our study is to examine the distribution of G proteins among the plant, metazoan, as well as basal metazoan lineages and to elucidate the evolutionary history of these proteins. Our results thus far have proven to be promising yet these G proteins in algae remain elusive.

II. METHODS

The first set of sequences from Viridiplantae were obtained by using a keyword search in the Entrez protein database provided by The National Center for Biotechnology Information (NCBI). After reducing redundancies in the data set, preliminary phylogenetic trees were reconstructed and used to cluster the Alpha, Beta and Gamma G proteins. Alignments from each of these clusters were produced using CLUSTALW [13], MUSCLE [4] and MAFFT [10]. Phylogenetic analysis was done using the JTT substitution model [9] and the neighbor-joining method [17] with bootstrap analysis [5]. The maximum likelihood method implemented by PhyML [7] was also used with bootstrap analysis.

Profile hidden Markov models (HMMs) were built using the program suites SAM [12] and HMMER [3]. The training set from Alpha contained 22 plants, 23 animals, 5 fungi and 12 protist sequences. The model for Beta contained 10 plants, 11 animals, 8 fungi and 1 protist sequence. Gamma’s model contained 3 plants, 16 animals, 6 fungi and 1 protist sequence. These profile HMMs were used to mine against 29 genomes of plants, algae, protists and diatoms obtained from the Joint Genome Institute (http://genome.jgipsf.org), The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org), The Dicytostelium discoideum Genome Project (http://dictybase.org), The Rice Genome Annotation Project (http://rice.plantpathology.msu.edu) and the Cyanidioschyzon merolae Genome Project (http://merolae.biol.s.u-tokyo.ac.jp). Basic local alignment search tool (BLAST; [11]) was used to perform reciprocal searches for significant hits obtained through the profile HMMs in order to corroborate these proteins are G proteins.

III. RESULTS

The results showed very significant E-values in the dicot and monocot plant genomes. There were also significant hits obtained from diatoms and the choanoflagellate Monosiga brevicollis (currently viewed to be the closest living ancestor of Metazoans [11]). Interestingly, our profile HMM searches did not identify G-protein candidates from the green and red algae with sufficiently low E-values. Table 1 shows this

![Fig. 1 G protein Signaling Cycle](image)

The classical model of the G-protein-signaling complex. The Alpha subunit exchanges its GDP for a GTP and it switches to its active state. Image from [18].
difference in E-values between annotated G protein Beta (Arabidopsis thaliana has an E-value of 6.40E-258) and the green algae Chlamydomonas reinhardtii (with a significantly higher E-value of 3.50E-17). All significant hits underwent reciprocal BLAST searches against the non-redundant protein database. For Beta, E-values >3.50E-17 returned no conclusive evidence of them being G proteins. The rest of the hits were aligned easily with other G proteins. Similar results were also produced using the Alpha and Gamma HMMs.

### Results of G-protein beta mining from genomes using the profile HMM

<table>
<thead>
<tr>
<th>Organism</th>
<th># of Hits</th>
<th>Best E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chaoflagellate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monosiga brevicollis</td>
<td>128</td>
<td>3.50E-17</td>
</tr>
<tr>
<td>Chlorella NC64A</td>
<td>104</td>
<td>6.20E-14</td>
</tr>
<tr>
<td>Chlorella vulgaris C-169</td>
<td>101</td>
<td>2.80E-13</td>
</tr>
<tr>
<td>Micromonas pusilla CCM1545</td>
<td>115</td>
<td>1.60E-15</td>
</tr>
<tr>
<td>Micromonas strain RCC299</td>
<td>132</td>
<td>1.10E-14</td>
</tr>
<tr>
<td>Ostreococcus lucimarinus</td>
<td>96</td>
<td>1.20E-10</td>
</tr>
<tr>
<td>Ostreococcus sp. RCC809</td>
<td>97</td>
<td>1.90E-11</td>
</tr>
<tr>
<td>Ostreococcus tauri</td>
<td>89</td>
<td>7.70E-11</td>
</tr>
<tr>
<td>Volvox carteri</td>
<td>140</td>
<td>4.30E-16</td>
</tr>
<tr>
<td>Heterolobosea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naegleria gruberi</td>
<td>155</td>
<td>8.70E-17</td>
</tr>
<tr>
<td>Monocots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>282</td>
<td>2.00E-264</td>
</tr>
<tr>
<td>Sorgum bicolor</td>
<td>175</td>
<td>2.80E-260</td>
</tr>
<tr>
<td>Mosses and Ferns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physcomitrella patens ssp patens</td>
<td>224</td>
<td>1.40E-239</td>
</tr>
<tr>
<td>Selaginella moellendorffii</td>
<td>310</td>
<td>5.30E-175</td>
</tr>
<tr>
<td>Mycetoza</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dictyostelium discoideum</td>
<td>131</td>
<td>1.20E-258</td>
</tr>
<tr>
<td>Red Algae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyanisochyzon meroiae strain 10D</td>
<td>57</td>
<td>6.20E-11</td>
</tr>
<tr>
<td>Stramenopiles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aureococcus anophagefferens</td>
<td>139</td>
<td>5.30E-163</td>
</tr>
<tr>
<td>Phaeodactylum tricornutum (Unmapped)</td>
<td>2</td>
<td>3.80E-04</td>
</tr>
<tr>
<td>Phaeodactylum tricornutum (Chromosome)</td>
<td>71</td>
<td>2.50E-101</td>
</tr>
<tr>
<td>Phytophthora capsici</td>
<td>120</td>
<td>9.10E-108</td>
</tr>
<tr>
<td>Phytophthora ramorum</td>
<td>138</td>
<td>1.50E-182</td>
</tr>
<tr>
<td>Phytophthora sojae</td>
<td>142</td>
<td>1.70E-182</td>
</tr>
<tr>
<td>Thalassiosira pseudonana (Unmapped)</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>Thalassiosira pseudonana (Chromosome)</td>
<td>78</td>
<td>6.20E-181</td>
</tr>
</tbody>
</table>

**IV. Discussion**

Our profile HMM search did not show any G-protein candidates with confidence from the 10 green and red algal genomes. It is interesting that many extant distant relatives (e.g., Monosiga brevicollis and Dictyostelium discoideum) of Viridiplantae and Metazoa showed very strongly similar proteins as G protein candidates, whereas the algal lineages have failed to produce significant results. It is possible the algal genomes are incomplete but this could not be attributed to an error in the genome sequence quality of algae as the green algae Chlamydomonas reinhardtii has coverage of 12.8x whereas Populus trichocarpa has coverage of 7.5x. Another possibility is that these lineages could have lost their need for G proteins or a different protein has replaced them. This calls for further investigation using alignment-free methods that may be more sensitive to non-canonical and divergent G proteins potentially existing in green and red algae. This will give us a better understanding as to the origin of G proteins and their evolutionary history.

**References**

15. Moriyama, EN, and Opiyo, SO. (in press) Bioinformatics of Seven Transmembrane Receptors in Plant Genomes in “Integrated G Proteins Signaling in Plants” (eds., S. Yalovskly, F. Baluska, and A. Jones), Springer Verlag.
I. INTRODUCTION

Multiple sequence alignment (MSA) is most often the first step of bioinformatics and molecular evolutionary analyses. MSA is done by aligning the nucleotides or amino acids, hereafter referred to as characters, in columns based on their inferred homology by similarity, inserting gaps between positions in order to align the characters. Gaps represent the insertion or deletion of sequence regions, or indels. In addition to characters, homology of the indels can also provide important information to molecular evolutionary analyses. However, gap information is usually ignored in many analyses.

Objective functions used in alignment methods often try to minimize the number of gaps by penalizing steps that introduce gap positions. The effectiveness of such objective functions to correctly place indels has been studied in only two unrealistic test cases (shown in [1]) that were not objectively quantifiable. Recent innovation in MSA methodology is to include homology information of indels into their objective functions [2], [3]. While these methods improved the character matching, gap placement accuracy is unknown. There is an absence of an indel representation that can be used to test indel placement accuracy.

Representing indels requires encoding the indels so that indels can be dealt with an equivalent manner as character data. Such indel coding methods have been used in, e.g., phylogenetic reconstruction [4]. Current methods, however, have drastic trade-offs between the number of indels used as information versus the number of falsely inferred indels. Due to this limitation, “indel coding” as been largely unused.

In this study, we introduce a novel method of representing indels, called gap profiling, and test the accuracy of inferring the placement of indels for the objective functions of six MSA methods. We calculated sensitivity and specificity of indel placement in each method, and use a novel ROC-like measure to rate their accuracies. To assess the validity of our method, we compare the performance of our measure to the standard character-based Sum of Pairs scoring measure.

II. METHODS

Gap profiling is a novel technique for analyzing the accuracy of indel placement by comparing indel positions between the “true”, or benchmark, MSA versus a reconstructed MSA.

We break each MSA M into the set of $N \times (N-1)$ pairwise alignments, where $N$ is the number of sequences in the MSA. For each pairwise alignment $(S_i, S_j)$, we remove columns that contain only gap characters, which are artifacts from the full MSA. We calculate the gap profile for $S_i$ against $S_j$, $GP_{S_i,S_j}$, by incrementing the gap profile position between two characters in $S_i$ when there is one or more gap characters appearing between the two characters in $S_i$, as in Figure 1. This results in a binary gap profile the length of $S_i + 1$ sites long. After calculating all pairwise gap profiles, we obtain a set of $N \times (N-1)$ gap profiles for $M$.

Using our representation, we define gap placement sensitivity as the distance of each gap in the true MSA, $GP_{S_i,S_j}$, to the closest inferred gap in the reconstructed MSA, $GP'_{S_i,S_j}$, and the gap placement specificity as the distance of each inferred gap in the $GP'_{S_i,S_j}$ to the closest gap in $GP_{S_i,S_j}$. We bin these distances as shown in Figure 2, which represent the distances from zero (exact match) to $w$, where $w$ is a given maximum value ($w = 4$ is used in Figure 2).

To test our indel representation, we use the sequence simulation method indel-Seq-Gen version 2.0 [5] to simulate
20 datasets of 16 sequences, each 1000 characters long, varying the substitution and indel rates, number of taxa, and indel length distributions. The “true” MSAs were obtained from the simulation. The reconstructed MSAs were obtained by ClustalW2 [6], MAFFT [7], Muscle [8], ProbCons [9], PRANK [3] (using both PRANKs guide tree, and supplying the true guide tree), and FSA [2] (using standard and maximum sensitivity). We used gap profiling and Sum of Pairs score to compare these reconstructed and “true” MSAs.

We introduce a ROC-like sensitivity vs. specificity plot based on the gap profiles to assess method performance. We use a sliding decision line along the sensitivity bins (see Figure 2), starting at bin 0 to bin w, assuming everything to the left of the line is a true positive inference, while the rest are false positives. For each bin, we calculate the sensitivity as (TP)/(TP + FP). We use the same procedure in reverse (from bin w to bin 0, true negatives are on the right of the line) for specificity bins, where the specificity is (TN)/(TN + FN). We repeat this procedure for w = 5, 10, 25, 50, 100, 250, and 500.

We measure the efficacy of the MSA reconstruction methods by the area under the curve (AUC) of this plot. We compare our measure of indel placement performance for each of the multiple sequence alignment methods against the standard performance measure that is based on character placement, the Sum of Pairs score.

III. RESULTS AND DISCUSSION

Results for the Gap Profile AUC are shown in Figure 3. The Gap Profile AUC generally correlates well with the Sum of Pairs score, although the Sum of Pairs score generally rates programs much higher. Since the substitution rate is low, the Sum of Pairs score is generally unaffected by the rate of indels. The possible exception is PRANK. PRANK is the only method that infers insertions and deletions as different events; inferring a deletion as an insertion will cause one or two character shifts in the alignment, which causes the Sum of Pairs score to be artificially low. The gap profile AUC, however, is robust to small shifts, while also penalizing methods that perform poorly at placing gaps (the so-called “gap magnets”, in which methods concatenate gaps in order to maximize their objective functions, which are commonly tested for accuracy using the Sum of Pairs score), particularly ClustalW2. Both the Sum of Pairs score and the Gap Profile AUC show a declining trend as the proportion of insertions increases, since insertions cause MSAs to be much gapper, and methods create more gap magnet regions. Surprisingly, PRANK, which treats insertions and deletions differently, does not show improvement in the Gap Profile AUC. More investigation is necessary to determine whether this is a problem with PRANK or with Gap Profile AUC. It may be necessary to consider a normalization step to account for the overrepresentation of insertions in MSAs. The gap profile AUC shows promise as a method for evaluating the accuracy of MSA methods in gap placement, leading to new ways to represent indels in objective functions used in MSA methods.

IV. ACKNOWLEDGEMENT

This work was supported by NSF AToL grant 0732863.

REFERENCES

Poster Abstracts
MicroRNA target prediction and verification in the green alga Chlamydomonas reinhardtii

Adam Voshall\(^1\), Etsuko N. Moriyama\(^2\), and Heriberto Cerutti\(^2\)

\(^1\)School of Biological Sciences, University of Nebraska-Lincoln
\(^2\)School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln

MicroRNAs (miRNAs) are a component of the RNA interference (RNAi) machinery, which provides cells with a precise and delicate means of regulating gene expression by a variety of mechanisms. MicroRNA-mediated gene silencing can occur by mRNA cleavage and degradation (found mainly in plants) or by translational repression (found mainly in animals). While it has been established that the green alga Chlamydomonas reinhardtii contains functional miRNAs, little is known about what genes these miRNAs target for silencing. In this study, we developed an in-silico method to predict miRNA targets in the C. reinhardtii genome.

Five potential miRNAs were identified from the sequences of expressed small RNAs isolated from C. reinhardtii strains. They were chosen based on folding patterns characteristic of miRNA precursors. Computational prediction of the targets of these five miRNAs proceeded in multiple steps. First, we identified potential binding-sites using the PITA target prediction program. This program determines the change in free energy (ddg) necessary for the binding to occur [1]. Next, we matched the miRNA sequences against all mRNAs 3’ UTRs predicted by version 6.8 of the Augustus gene modeling program [2]. We also matched them separately against all mRNA transcripts in the Joint Genome Institute (JGI) frozen gene catalog, version 20080828 [3]. Only the most energetically favorable interactions (those with a ddg less than -20 for targets predicted from 3’ UTRs and -15 for those predicted from transcripts) were saved for further analysis. We then used the RNAhybrid program to find the most likely binding site within the 3’ UTR or transcript, requiring a perfect match for the entire seed region (nucleotides 2-8) [4]. For targets predicted from both 3’ UTRs and transcripts, the results were filtered based on gap length and location, whereas those predicted from transcripts were additionally filtered to keep only predictions with one or no mismatches. Finally, we examined all remaining putative targets on the JGI genome browser, filtering out any predictions that did not correspond to a known gene, had no expressed sequence tag data, matched to a transposable element, or for which the target gene was not conserved in the genome of another green alga, Volvox carteri. As a result, we identified six potential targets with matches to their 3’UTRs and another seven with matches to their coding regions.

From the final list, we chose five predicted targets (three from 3’UTRs and two from coding regions) for experimental verification. We are currently examining the transcript levels of these target genes by quantitative real-time PCR in wildtype cells and corresponding RNAi-defective mutants. If the target genes were correctly identified, we expect to see higher mRNA levels in the mutants. We will also test protein-levels by immunoblotting on the same sets of strains. Finally, our prediction strategy will be refined and improved based on the results of this verification.

Screening for anti-angiogenic natural inhibitors for vascular endothelial growth factor receptor-An Insilco approach

Subhojyoti Chatterjee¹, Deepak Choubey¹, Biplab Bhattacharjee²*, Anantharamanan.R², Jayadeepa RM², Anirudh Chowdhary², Sushil K Middha³ and Usha Talambedu³

¹ Amrita Vishwa Vidyapeetham, University, Kerala, India
² Institute Of Computational Biology(IOCB),Bangalore
³ Maharani Lakshmi Ammanni College for Women.Bangalore

*Corresponding Email: biplabbhaiyu@gmail.com

Abstract

Angiogenesis is the formation of new blood vessels. Angiogenesis is a process controlled by certain chemicals produced in the body. Some of these chemicals stimulate cells to repair damaged blood vessels or form new ones. Other chemicals, called angiogenesis inhibitors, signal the process to stop. Angiogenesis inhibitors usually have only mild side effects and are not toxic to most healthy cells. In our study attempt was made to find potent anti-angiogenic inhibitor for vascular endothelial growth factor receptor using natural agents targeting biological processes important in cancer. One group of growth factor receptors critically implicated in angiogenesis is vascular endothelial growth factor receptors. VEGFR2 kinase domain in complex with a benzisoxazole inhibitor is served as a molecular target for our study. The investigational anti-angiogenic inhibitor Pazopanib was considered as a reference drug in this work. Four substances have been approved to control angiogenesis in the therapy of renal cell carcinoma: Sunitinib, Sorafenib, Temsirolimus, as well as a combination of bevacizumab and interferon alpha. Other substances, such as Everolimus, Pazopanib and Axitinib, are currently the subject of clinical trials. Hundreds of natural molecules were selected from various scientific articles. The initial screening of the molecules based the lipinski’s rule of five. Molecules which were satisfying this rule were taken for receptor-ligand interaction study using docking tools HEX and Quantum. Around fifteen molecules were taken as lead molecule and its binding pocket on VEGFR2 was analyzed using SwissPDBviewer and Q-site finder. The molecules which were interacting with VEGFR2 were taken for pharmacokinetics study using ADMET tools. Ames test of the molecules was predicted for probability of mutagenicity on molecular system. Health effects of these molecules in blood, cardiovascular system, gastrointestinal system, kidney, liver and lung were considered for further screening of the molecules. The natural molecules Curcumin, Epigallocatechin gallate (EGCG), Barringtogenol and Finasteride were showing reliable interaction with VEGFR2 and their pharmacokinetics parameters were comparatively good than the pazopanib. The dietary product curcumin and epigallocatechin gallate (EGCG) can be cancer chemopreventive agents and the natural molecules and Finasteride can be effective inhibitors for vascular endothelial growth factor receptor.
Molecular Basis of Homocystinuria: Identification of Mutation in Cystathionine Beta Synthase (CBS) Gene
MA Kausar¹ and MD Bashyam²
¹Department of Respiratory Allergy & Applied Immunology of VP Chest Institute, University of Delhi, Delhi, India
²Molecular Oncology, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad (AP). India

Introduction:
Homocystinuria is a metabolic disorder due to cystathionine beta-synthase (CBS) deficiency, producing increased urinary homocystine and methionine. Mutation in cystathionine beta synthase gene has been identified as a molecular basis for the disease and is also responsible for its clinical heterogeneity. Hence, the present study was performed to identify mutation(s) in CBS gene that are responsible for homocystinuria in any individual.

Methods:
Mutation analysis was carried out in CBS gene in 3 patients of having symptoms of homocystinuria. DNA was isolated from the blood sample of these patients. The exons 3,4,5,6,8,10,11 and 12 of CBS gene were amplified by polymerase chain reaction (PCR) from the patient samples. DNA was eluted out from the after agarose gel electrophoresis. Sequencing was carried out to and compared with the sequence of normal subject to determine the mutation(s).

Results and Discussion:
The sequencing results indicated heterozygous mutation found in 10th exon of one patient at position 1080 of CBS gene. No mutation was observed in any other exon of this patient. Further, there was no mutation found in any exon(s) of other two patients. The heterozygous mutation converted triplet code for alanine, GCC (normal) to GCT (patient). As the mutation was observed in wobble base, it coded for the same amino acid, even after mutation. So, there was no effect on the final protein (i.e. Cystathionine Beta Synthase).

Conclusion:
Symptoms of homcystinuria and elevated level of homocystine were not due to mutation in CBS gene in any of the three patients. They may be due to several other disorders including methylcobalamin deficiency, vitamin B12-responsive homocystinuria etc.
Our data indicate that all ribosomal proteins were originally encoded by the 32 codons that end in the nucleotides G or C and were heavily populated with the four amino acids Glycine (G), Alanine (A), Arginine (R) and Proline (P). Search vectors composed primarily of these amino acids retrieve all members of each of ribosome proteins in bacterial species with no false “hits”. Different combinations of G, A, R and P and appropriate insertions or deletions (INDELS) differentiate each ribosomal protein from all others. We are able to identify combinations of site mutations and INDELS that can be used to subdivide each ribosomal protein ensemble into the 17 classes of bacteria for which the total genomes have been reported. Within each class, the amino acid sequences have remained highly conserved over millions of years. Our data support the hypothesis that the genetic code arose in GC-rich bacteria. The residues G, A, R and P dominate the conserved residues in the fingerprints of ribosomal proteins because of the nature of these amino acids, protein folding, the ribosome, DNA and the genetic code. The eight codons in the universal code that are composed of G and C bases only are two each of G, A, R, and P. Glycine residues are achiral and can readily take up conformations that are rarely sustainable by any other amino acids. Every ribosomal protein has several glycines in conformations that are vital to the integrity of the ribosome. Proline residues are limited in conformational flexibility, introduce turns of a restricted nature into proteins, played a critical role in the folding of the earliest ribosomal proteins and have been retained. The first ribosomal proteins were rich in Arginine that provided a charge balance to the acidic nature of the RNA. Alanines encoding by the codons CGC and CGG may be associated with stability of proteins composed of α-helical bundles stabilized by high Alanine content. Stability of DNA is directly correlated with GC content and codon integrity would be enhanced by this stability. In some ribosomal protein families amino acid variation in a single sequence position correlates with separation of bacterial classes. The most hypervariable amino acid position in the 1300 fully aligned sequences of the L1 ribosomal protein is occupied by 18 of the 20 amino acids. The 18 residues in this position subdivide the L1 ribosomal protein into 21 phylogenetic classes of bacteria. For example, a tryptophan is present in the L1 ribosomal proteins in all firmicutes in Swiss-Prot/TrEMBL and an aspartic or glutamic acid residue isolates 92% of all γ-proteobacteria. The only Eukaryotic proteins that share the universal fingerprint of the ribosomal proteins are the proteins of the small subunit of the ribosome. The universal fingerprint for ribosomal protein S4 captures 911 proteins in Swiss-Prot and over 1000 in TrEMBL. This total includes all of the members of the 21 phylogenetic classes of bacteria and over 740 ribosomal proteins in the chloroplasts of almost all the plants in Swiss Prot/TrEMBL. The amino acids in one sequence position, in the 200 amino acid long S4 protein, separates all eukaryotic proteins, all cyanobacteria, the only plactomycetes, most spirochaetes, and half of the actinobacteria from the other 13 classes of bacteria. It is noteworthy that this partitioning separates the gram-positive and gram-negative bacteria, which have one and two membranes respectively. It has been hypothesized that bacteria with a single membrane are the most ancient. A histidine residue in another position in S4 further isolates all 740 ribosomal proteins in plants in Swiss Prot TrEMBL from all bacterial S4 proteins.

Tools for Identifying Homogenous Subgroups in Large Data
Yuanyuan Huang\textsuperscript{1,2}, Heike Hofmann\textsuperscript{1,2}, and Dianne Cook\textsuperscript{1,2}
\textsuperscript{1}Department of Statistics, Iowa State University
\textsuperscript{2}Program of Bioinformatics and Computational Biology, Iowa State University

Keywords: data mining, biostatistics, subgroup

With an increase in complexity of data sets the task of finding homogeneous subgroups becomes more challenging and yet more important. The presence of homogeneous subgroups allows a more targeted treatment of individuals in all aspects of life. At the example of the National Health and Nutrition Examination Survey (NHANES), 1999–2002, we will discuss visual tools of identifying subgroups of the population with similar dietary habits. In homogeneous sub-populations with “good” quality diet, the health status of individuals is related to various explanatory factors, such as sociodemographic characteristics (sex, age, race and ethnicity, education, etc) and body measurements (tooth retention, body mass index, serum of Phenylalanine, etc). This project studies the NHANES III data using machine learning approaches to data mining and knowledge discovery, to identify the sub-population that benefit most from a change to healthier dietary habits. We compare the results from different multivariate models and discuss possible future applications in medicine that physicians could assess the appropriateness and drug effects of assigning some treatment to a particular patient.
Biosynthetic Engineering of HSAF, a Tetramic Acid-containing Antifungal Macrolactam Isolated from the Biocontrol Agent *Lysobacter enzymogenes* C3

Yunxuan Xie, Yaoyao Li, Lili Lou, Liangcheng Du

Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588
Extraction of globular proteins with reverse micelles using vegetable oils as the organic medium.

Daliya A. George, Cheuk-Man Chu, David C. Stuckey

Department of Chemical Engineering and Chemical Technology, Imperial College London, Exhibition road, London SW7 2AZ, United Kingdom.

Reverse micelles are nanosized water in oil microemulsions, however, most oil (solvent) phases such as isoctane are difficult to work with in industry due to their volatility and cost. An attempt has been made in this work to extract globular proteins such as Lysozyme (MW.14.3 KDa, pH 11) and Bovine Serum Albumin (BSA, MW. 66 KDa, pH 4.9) with reverse micelles of Sodium bis (2-ethylhexyl) Sulfo succinate (AOT) in more industrially friendly solvents such as vegetable oils (Rapeseed Oil, Corn Oil, Sunflower Oil). Although the dissolution of AOT in these oils required more time compared to organic solvents, it was enhanced with suitable co-surfactants such as 1-hexanol. The extraction efficiency of both Lysozyme and BSA were studied by varying parameters such as AOT concentration, pH of the aqueous phase, cation concentration and type in the aqueous phase, and type of oil. With Lysozyme optimized conditions resulted in 90-95% forward extraction and 80-85% back extraction. Very slight precipitation was observed in forward extraction and the maximum extraction was obtained with only 5mM of AOT. With BSA the best extraction under optimized conditions was 80-85% forward extraction and 75-80% back extraction. All three oils studied did not show any significant difference in terms of extraction efficiency although their viscosity varied substantially. The bioactivity of the proteins recovered after back extraction appeared to be reduced and thus needs further investigation. Nevertheless, the use of vegetable oils in reverse micellar extraction appears to be a good substitute for organic solvents, and we are now looking at monoclonal antibody and fragment extraction.
Matched filter for signature extraction in bioinformatics data set
Ming Li and William Q. Meeker
Department of Statistics, Iowa State University

In many bioinformatics research and applications, we need to extract interesting signals or to compare between signals in large multivariate data sets. In most cases those signals are buried in “noise”. Finding methods to increase signal-to-noise ratio is an important statistical problem studied for many years and there exist a set of well developed methods for noise reduction and signal boosting. In large scale multivariate bioinformatics data sets, when the signature of the signal is known, at least approximately, the concept of a matched filter can be applied to increase the signal-to-noise ratio dramatically. This, in turn, makes the automatic signal extraction and detection possible and reliable. Further more, matched filter is the optimal linear filter with the present of white noise.

In this poster, we first conceptually illustrated the idea of matched filter and the mechanism of increasing signal-to-noise ratio in a simple one-dimensional (1D) time sequence data set as shown in Fig. (A) and Fig. (B): the blue lines with triangle symbols are contaminated signal data with noise and the red lines with dot symbols are the pure signal without noise while the green lines with square symbols are the pure white noise. It is clear that the signal-to-noise ratio is increased at least 10 times when compared the original scale results and after matched filter results. Then the matched filter technique is applied to a 3D data set originally from nondestructive evaluation research, and a particular 2D slice of the 3D data set is compared before and after matched filter at Fig. (C) and Fig. (D). Finally we explore the possibility of using a matched filter to find particular segment in a DNA sequence, to compare the difference between two DNA sequences, and to improve complex image analysis results.
Aquatic environments can vary dramatically in the concentration of CO$_2$, due mainly to slow diffusion and interconversion of CO$_2$ ↔ HCO$_3^-$- . Some species of algae, such as *Chlamydomonas reinhardtii*, have developed a mechanism to deal with this variance in CO$_2$ levels, where CO$_2$ is actively transported into the algal cell, thereby maintaining a constant, elevated intracellular level of CO$_2$. Recently, two strains of algae, *Coccomyxa* C-169 and *Chlorella* NC64A, have been genome-sequenced. The presence of a 'carbon-concentrating mechanism' (CCM) in these strains is unknown. To investigate, a list of *C. reinhardtii* proteins involved in the CCM was compiled and sequences gathered. The proteins sequences were blasted against the C-169 and NC64A genomes in order to locate candidate homologs. Candidate homologs were then back-blasted against the *C. reinhardtii* genome in order to verify the specific homology. Further investigation of candidate homologs was performed using multiple alignments and hidden Markov models. Findings from this work showed that both C-169 and NC64A possessed homologs of the majority of *C. reinhardtii* CCM proteins. In particular, both C-169 and NC64A contained homologs of key CCM proteins, such as CIA5, CAH1, and LCIB, which are critical for CCM function in *C. reinhardtii*. Of particular interest were the CIA5 and LCIB homologs found in C-169 and NC64A. Multiple alignment and hidden Markov modeling of *C. reinhardtii* CIA5 and the CIA5 homologs in C-169 and NC64A revealed high conservation in only the N-terminal portion of the proteins, which includes two zinc-finger motifs. This raises questions as to the purpose of the central and C-terminal domains of CIA5. LCIB, which is crucial for CCM function, has three paralogs (LCIC, LCID, LCIE) in *C. reinhardtii*, further emphasizing its importance. In contrast, NC64A has two LCIB homologs and C-169 has a single LCIB homolog. This observation questions the functional necessity of LCIC, LCID, and LCIE in *C. reinhardtii*. 
Since April of 2009, the novel swine-origin influenza A (H1N1) virus (S-OIV) has rapidly spread across the globe. The World Health Organization declared the outbreak a global pandemic in June. The new S-OIV has a unique genome composition and contains genes derived from avian, human, and swine influenza viruses. It has been reported that the new S-OIV may have emerged from the reassortment of previous reported triple-reassortant swine influenza viruses and Eurasian avian-like swine lineages. Human infections with swine influenza A viruses had been reported in the past. Although the infections occurred sporadically, some cases were fatal.

In this study, phylogenetic analysis of human H1N1 viruses was carried out to characterize the similarity and differences between the new S-OIV and previously reported human H1N1 viruses. Human H1N1 virus protein sequences were downloaded from NCBI influenza virus sequence database. The data consists of all samples of H1N1 viruses in human hosts deposited in Genbank before Sept. 2009. Phylogenetic trees for the viruses were inferred using the neighbor-joining method, with genetic distances calculated based on mPAM.

The results of this study reveal relationships among the new S-OIV and the previously reported human H1N1 viruses. The phylogenetic trees based the sequences of four proteins (PB2, PA, NP, NA) present three clearly visible clusters of H1N1 viruses. In the tree based on NA (Fig. 1), 59 samples of S-OIV were clustered together in the second cluster along with five cases in 1976, a 2005 Iowa case, a 1998 Wisconsin case and a few isolated cases in 1988 and 1991. The results from other six proteins (PB2, PA, HA, NA, NS1, NS2) also indicate the new S-OIV is closely related to the Fort Dix swine flu outbreak happened in 1976 also. As reported in the journal Nature (June 25, 2009), the common ancestor of S-OIV and the closest related swine viruses existed between about 10 and 20 years ago. However, the Fort Dix swine flu virus existed over 30 years ago. The results suggest that although some zoonotic influenza virus infections occur occasionally, they can potentially evolve into strains that causes pandemic.
Graphical Framework for processing GCMS Data: Integration of qt interfaces

Marie Vendettuoli\textsuperscript{1,2}, Michael Lawrence\textsuperscript{3}, Dianne Cook\textsuperscript{1,2}, and Heike Hofmann\textsuperscript{1,2}

\textsuperscript{1}Bioinformatics and Computational Biology, Iowa State University, Ames IA
\textsuperscript{2}Dept of Statistics, Iowa State University, Ames IA
\textsuperscript{3}Fred Hutchinson Cancer Research Center, Seattle WA

Analysis of gas chromatography mass spectrometry (GCMS) for the purpose of quantifying and identifying metabolite requires several processing steps: noise elimination, peak detection, component detection, component alignment, retention time correction, component quantification, and normalization. Current data analysis tools (AMDIS, metAlign, MET-IDEA, XCMS, etc) offer an approach that limits users’ control and visualization of intermediary steps, creating ‘black box’ functionality. Additionally, knowledge of a programming language may be required or processing may be fragmented across multiple software programs. This diverts efforts from result analysis and creates danger of inconsistent data analysis between experiments. Furthermore, there is limited access to descriptive statistics and other diagnostic tools to validate the accuracy of generated results.

The \textit{chromatoplots} package, written for the R environment, is a statistical approach to to a unified framework of the pipeline described above. Methods build upon strengths of existing tools, while emphasizing transparency throughout processing. Integration with R graphics and GGobi offers users the ability to view both descriptive statistics and processing diagnostics. Command-line interface affords users opportunity to expand functions beyond current state. A new graphical user interface (GUI) guides analytical statistics and processing diagnostics, eliminating the need to acquire additional programming skills while offering users the opportunity to validate and compare multiple approaches of data analysis. We present the incorporation of \textit{qt interfaces}, a collection of R packages that extend functionality of the GUI to include dynamically interactive graphics. This increases beneficial impact of embedded statistical analysis by allowing researchers to rapidly compare multiple processing parameters or elucidate areas of high interest in early analysis phases.
Immune-enhancing antitumor RNA Interference therapy for cervical cancer
Richa Singhania¹, Wenyi Gu¹, Nigel McMillan¹
¹Diamantina Institute for Cancer, Immunology and Metabolic Medicine, University of Queensland, Princess Alexandra Hospital, Woolloongabba QLD-4102

Cervical cancer, caused by overexpression of human papillomavirus E6 and E7 oncogenes¹, contributes to almost 12% of the global cancer burden in women making it the second most frequent cancer in women worldwide². Even though prophylactic vaccines, Gardasil® and Cervarix®, have been developed, they are of limited use to women who are already infected. Also, owing to prolonged course of cervical carcinogenesis, the cancer burden will not be reduced significantly for several years even after the availability of these preventive vaccines³. So, an effective therapeutic vaccination strategy is required. In recent past, RNA Interference (RNAi) technology has emerged as an effective way to silence cancer-causing genes⁴. The recent findings show that when short hairpin RNA (shRNA) is targeted downstream of the cytotoxic T cell epitope in E7 gene, 5’ mRNA intermediates are generated, which get translated into incomplete proteins and result in increased presentation of target protein and generation of a tumor-protective immune response in mice⁵. It is known that certain viral proteins are rapidly presented to the immune system than normal proteins due to the production of defective ribosomal products (DRiPs) by the virus⁶. Our work aims to bring together the concept of RNAi and DRiP and we hypothesize that RNAi-enhanced adaptive immune response against tumor antigens occurs via increased presentation of defective proteins. Accordingly, we have investigated the fate of RNAi-cleaved mRNA and determined that shRNA treatment decreases the half life of both full length and the 3’ end fragment of the E6/7 mRNA, but the 5’ end fragment appears to be more stable. This would favour its translation into truncated protein, which we will show from polysome analysis and identify the location of 5’ end fragment in cell using immunoprecipitation and confocal microscopy. Also, we will investigate the levels of peptides presented on the cell surface after RNAi treatment by mass spectrometry to address the hypothesis that immune enhancement occurs via increased presentation of defective proteins to the immune system. In summary, the results to date suggest that RNAi therapy need not kill or be delivered to every cancer cell to be effective, thus demonstrating its potential to overcome current efficacy and delivery constraints. By integrating this novel treatment into current chemotherapy, we envisage better outcomes in pre clinical and clinical studies.

Cited References

Layout of Phylogenetic Networks and Multi-Labeled Trees

Poornima Venugopal, Sergey Bereg
1Department of Computer Science, University of Texas-Dallas

Recently, Multi-Labeled (MUL) trees have been used in polyploid studies to construct Phylogenetic networks [1]. We present a novel method to improve the layout of the Phylogenetic networks constructed from MUL trees. The idea is to rearrange the MUL tree to bring identical subtrees closer. We introduce a new metric $MIS$-Distance to quantify closeness of identical subtrees of MUL trees.

Algorithm: MUL-Arrange
1. Identify the maximal inextendible subtrees[1] of an MUL tree $T$
2. Swap the left and right subtrees of the root and compute the $MIS$-Distance. Retain the tree with minimal $MIS$-Distance.
3. Recursively apply the swap operation to the root nodes of the left and right subtrees retaining the MUL tree with the least MIS-Distance at each stage.
4. Repeat Steps 1 through 3 until the MIS-Distance output by consecutive iterations is the same.

Illustration: MUL Tree and Phylogenetic Network of part of tree appearing in [2] on plant taxa Silene ajanensis (S.a), Silene sorensenis (S.s), Silene ostenfeldii (S.o), Silene involucrata (S.i), and Silene uralensis (S.u).

Figure 1: (a) Original MUL Tree  (b) Phylogenetic Network

Figure 2: (a) Rearranged MUL Tree  (b) Phylogenetic Network

References:
Intraperitoneal delivery of liposomal siRNA for therapy of cervical cancer

Akul Singhania1, 2, Sherry Y. Wu1, and Nigel A.J. McMillan1
1Diamantina Institute for Cancer, Immunology and Metabolic Medicine, Queensland, Australia
2Department of Bioinformatics and Computational Biology, Iowa State University, Ames, IA

Since its discovery, RNA Interference (RNAi) technology has been effectively used as a gene therapy to treat a wide range of human diseases, including cancer[3]. Briefly, RNAi can be described as a means to specifically silence the expression of endogenous genes. This can be achieved by different mechanisms involving different classes of small RNAs – siRNA (short interfering RNA), shRNA (short hairpin RNA), miRNA (micro RNA) and piRNA (piwi-interacting RNA)[2]. Cervical cancer, caused by the continual expression of E6 and E7 oncogenes of human papillomavirus, is the second most common cancer affecting women worldwide[3]. Our research employs RNAi technology as a therapy for cervical cancer by utilizing E6/7 specific siRNAs. siRNAs are short dsRNA moieties (~21-23 nucleotides in length) that are loaded into the RISC (RNA-Induced Silencing Complex), which then targets homologous mRNA transcripts by base-pairing, resulting in gene silencing[4]. Although the results in vitro seem promising, delivery of siRNA into an in vivo system still remains a major issue. Naked siRNAs get degraded quickly by RNase activity in serum before reaching the target tissues and even chemically modified siRNA have a short half-life[5]. For this reason, siRNAs are delivered through viral, bacterial, peptide and lipid-based delivery systems[2]. Liposomes, tiny bubble like vesicles made of lipids, have shown promising outcomes in this field due to their stable lipid bilayer structure, which prevents degradation of siRNA packaged in the core of liposomes. Synthetic siRNAs are effective only for 2-3 days when introduced in cells, making repeated dosing necessary[2]. This is not possible by the traditional way of intravenous (I.V.) injection administered via the tail of mice, since it is time consuming, requires skill, and most importantly the tail tissue gets damaged after 3-4 injections. Our project aims to test the efficacy of liposomal siRNA administered by intraperitoneal (I.P.) injection (through the peritoneal cavity) in cervical cancer mouse models, due to ease in injection and convenience of multiple dosing. So far, we have shown the uptake of E6/7 specific liposomal siRNAs in HeLa cells through FACS and knockdown of E6/7 oncogenes in vitro by Western Blot analysis. In in vivo studies, we have successfully delivered liposomal siRNAs to subcutaneous tumors and found that the liposome size and siRNA dose are crucial for efficient delivery. We have also compared the delivery efficiency of cationic and non-charged liposomes and found cationic liposomes to be better. In future we will investigate the mechanism by which liposomes get delivered to different organs, retention and biodistribution of liposomes and also carry out tumor reduction studies in mice. In conclusion, the results to date show that RNAi is an effective tool in cervical cancer therapy, and I.P. delivery of liposomal siRNA has therapeutic efficacy comparable to that of I.V. delivery.

References:

KNOCK DOWN OF CXCR2 ENHANCES SENSITIVITY TO CHEMOTHERAPY
Dhananjay M. Nawandar1, Michelle L. Varney2, and Rakesh K. Singh3
1Department of Biology, University of Nebraska- Omaha
2,3Department of Microbiology and Pathology, University of Nebraska Medical Center

Cancer has always had and still has a huge impact on human society. According to the American Cancer Society, cancer accounts for nearly one quarter of deaths in United States, exceeded only by heart disease. In 2006, there were 559,888 cancer deaths in the US. According to the World Health Organization, deaths from cancer worldwide are projected to continue rising, with an estimated 12 million deaths in 2030. Although prostate cancer is the most prevalent cancer among males, breast cancer is the most prevalent cancer among females.

The current cancer therapy, which is also known as the conventional cancer therapy, mainly consists of either chemotherapy, surgery or radiation therapy. A lot of times these therapies are successful in treating the disease but many times they fail to do so, which leads to the recurrence and progression of the disease. Chemotherapy and radiation therapy have side effects on the human body, which leads to toxicity. As opposed to early stage disease, many challenges exist in the current management of advanced stage breast cancer as there are fewer recognized therapeutic strategies, often due to therapy resistance. The major challenges right now in this field of research are to improve the efficacy of the current therapeutic regimens by limiting its toxicity and the reversal of therapy resistance i.e. to make tumor cells more sensitive to therapy. Recent reports suggest that malignant cells that survive initial chemo- and radiation therapy often express inflammatory cytokines such as CXCR2 ligands, which provides survival benefit making tumors resistant. The specific objective of this study is to develop strategy to manipulate chemokine-chemokine receptor network to develop effective therapy with limited toxicity for drug-resistant breast cancer.

CXCR2 and its ligands like CXCL1, 3, 5 and 8 have been shown to play an important role in inflammation and tumor progression. They have been shown to be associated with the aggressiveness of the tumor cell lines. CXCR2 ligands enhance malignant cell proliferation and survival. There is an increase in the levels of CXCR2 ligands in the patient’s body in response to chemotherapy. When the tumor cells are exposed to chemotherapy drugs, there is an increase in the production of CXCR2 ligands, which bind to CXCR2 leading to therapy resistance. Based on these previous studies, we hypothesize that inhibition of CXCR2 and its ligands signaling can enhance the efficacy of chemotherapeutic agents in malignant breast cancer. The specific objective of this study is that knocking down of CXCR2 expression in malignant cells can enhance chemotherapy response and lower the effective dose of chemotherapy drugs.

We used mice breast cancer lines which expressed different levels of CXCR2 (Cl66-control and Cl66 sh-CXCR2) and examined their response to doxorubicin and paclitaxel. There was the normal activity of CXCR2 in Cl66-control cell lines, while the activity of CXCR2 was knocked down in Cl66 sh-CXCR2 cell lines using the small interfering RNA technique. To determine the therapeutic response of these cell lines in response to treatment with the drugs, we used the cytotoxicity assay. To determine the ability of the drugs to induce apoptosis in these cancer cell lines, we performed the apoptosis assay and to determine how CXCL1 expression varies in these cell lines according to the different concentration of the chemotherapy drugs, we performed the enzyme linked immunosorbent assay (ELISA).

We found a significant difference in CXCL1 expression between the two cell lines at different concentration. We also found that at all the concentrations of chemotherapy drugs, Cl66-control cell lines proliferated more as compared to Cl66 sh-CXCR2 cell lines. Especially at lower concentration of chemotherapy drugs, even though the Cl66-control cell lines were proliferating, the Cl66 sh-CXCR2 cell lines were being inhibited. This suggests that knock down of CXCR2 enhances the sensitivity of therapeutic agents at lower concentrations. Therefore, inhibition of CXCR2 may possibly prove to be an effective adjuvant in treatment of cancer in future.
Integrating EST’s and High-Throughput Sequencing to Identify Orthologous Loci for Phylogenomic Analysis: A Case Study among Basal Ray-Finned Fishes (Actinopterygii)

Jason Macrander¹, Guillermo Orti²
¹ School of Biological Sciences, University of Nebraska-Lincoln
² Department of Biological Sciences, George Washington University

Next-generation, high-throughput sequencing is revolutionizing molecular data collection. Consequently, disciplines that incorporate molecular data are refocusing their attentions in order to take advantage of this technological breakthrough. Following suite, the field of phylogenetics is shifting towards analyses which incorporate partial/full genomes or transcriptomes of model organisms into analyses, adopting the name “phylogenomics”. The strength behind phylogenomic studies comes from the ability to take advantage of large datasets to be used for inference of the relationships among species. The utilization of expressed sequence tags (ESTs) has been shown to be an effective approach to address evolutionary questions when multiple loci are desired. However, until now this has heavily depended upon large data sets available from model organisms. With the advent of next-generation, high throughput sequencing, large data sets for phylogenomic analysis is no longer restricted to model organisms. We have taken the initial steps to do such an analysis by constructing EST libraries from 8 non-model organisms. We used the 454 Genome Sequencer FLX next-generation sequencing platform to construct partial transcriptomes for 7 ray-finned fishes (Polypterus endlicheri congicus, Scaphirhynchus platorynchus, Lepisosteus platostomus, Amia calva, Echidna nebulosa, Osteoglossum bicirrhosum, and Dorosoma cepedianum) and 1 lobe finned fish (Potopterus annecteus). Preliminary data analyses have found ~3000 unique ESTs per species with contigs from 250 – 1600 bp. The software pipeline OrthoSelect (Schreiber et al. 2009) is used to prepare our data in addition to data freely available online. OrthoSelect automatically assigns ESTs to predefined Orthologous Groups (OGs), translates the sequences, eliminates redundant sequences, and creates multiple sequence alignments as suitable input data for phylogenetic reconstruction. We estimate that this analysis will yield ~400+ orthologous gene sets to be used in phylogenomic analysis, the largest data set constructed for analysis of phylogenetic relationships among ray finned fishes.
Identifying Differentially Expressed Human Lung MicroRNAs and Their Molecular Functions

Sarita Limbu and Zhong-Hui Duan
Department of Computer Science, University of Akron, Akron, OH

MicroRNAs are non-coding, short RNAs which regulate the gene expression in cells either by cleaving the target mRNAs or by suppressing the target protein synthesis. Differentially expressed microRNAs affect the normal gene expression and cause diseases such as cancer. MicroRNA expression profiling is possible due to microarray technology. There is a need for development of a computational technique to determine the biological pathways altered by the differentially expressed microRNAs so that the effective study of any type of cancer can be conducted easily whenever required. In this research, a microRNA dataset, consisting of normal and lung cancer data, is analyzed. The microRNA data are normalized and the microRNAs that are differentially expressed in the lung cancer microarray than in the normal lung microarray are obtained using statistical method called T-test. Only those microRNAs with p-values less than 0.01 are considered here for more preciseness. The samples of differentially expressed microRNAs are classified using KNN classifier for five categories of microRNAs: top 5, 10, 20, 40 and 50 microRNAs with 50 randomly generated test cases for each category. The classification was 81% accurate on average for five categories of top microRNAs. The results also indicated that the lung cancer samples can be classified using a small panel of 5 microRNAs with over 85% classification accuracy which means the lung cancer sample and normal lung sample are significantly distinct. The microRNA names made available in our dataset file are very specific names whereas predicted target list are available for more generic microRNA names. 15 unique microRNA names: hsa-mir-21, hsa-mir-210, hsa-mir-30d, hsa-mir-205, hsa-mir-30a, hsa-mir-191, hsa-mir-203, hsa-mir-24, hsa-mir-20, hsa-mir-214, hsa-mir-216, hsa-mir-29b, hsa-mir-30b, hsa-mir-215 and hsa-mir-218, were obtained by processing the top 50 differentially expressed microRNAs. 3103 predicted mRNA targets of those differentially expressed microRNAs are identified from online resource. The up-regulated and down-regulated microRNAs are identified and their targets are stored separately. The molecular functions associated with the predicted targets of both types of microRNAs are retrieved from Gene Ontology Consortium and are represented using GO IDs in directed acyclic graphs. Also, the GO IDs for the human’s entire GeneIDs are retrieved by processing files made available by European Bioinformatics Institute (EBI) through its Gene Ontology Annotation (GOA) project. They are also represented in a direct acyclic graph. The entire GeneID-GOID set of human as well as the GeneID-GOID sets of the predicted targets of both the up-regulated and down-regulated microRNAs are further processed. Fisher’s Exact Test was employed for conducting significance test to find those functional groups which are prominently affected by the differentially expressed microRNAs. The results show that the differentially expressed microRNAs could potentially change the expression of mRNAs and therefore modify the related cellular processes. Significance testing identified 213 molecular function groups which are exhibiting strong association with the differentially expressed microRNAs and hence with the lung cancer disease, and are potentially on the biological pathways altered by these differentially expressed microRNAs.
Epigenetic marks, including DNA and histone modifications, have been shown to be influenced by a variety of environmental factors, such as basic nutrients and hormones, leading to changes in chromatin and cell phenotype. Changes in cell phenotype may lead to genetic disorders and birth defects which are more common in children conceived with the help of assisted reproductive technology (ART) compared to normal pregnancies. Some researchers have postulated this may be due to inappropriate culture media composition for oocytes, prior to fertilization, which could lead to aberrant epigenetic marks. Currently, epigenetic marks are near-impossible to study in single oocytes with currently available epigenetic technologies due to the technical difficulties with collecting a number of oocytes large enough for subsequent analyses. Chromatin immunoprecipitation (ChIP) assays for analysis of epigenetic marks require at least 10,000 cells and heavily depend on availability of antibodies. The aim of this project is to develop an antibody-independent technology to map chromatin proteins and epigenetic profiles in samples consisting of very few cells. For the initial development of this technology, human breast adenocarcinoma (MCF-7) and human breast (MCF-10A) cell lines were used mainly due to their availability and also to study differences in gene expression between cancer and noncancerous cells. DNA adenine methyltransferase identification technique (DamID), developed by the Van Steensel lab in the Netherlands, was adopted for this project. DamID is based on the creation of a fusion protein consisting of E. coli DNA adenine methyltransferase (Dam) and a chromatin protein of interest. Dam is proven to methylate adenine in the sequence GATC. Endogenous methylation of adenines is absent in most eukaryotes and therefore DamID technology works as a unique tagging system to mark the genomic binding sites of the chromatin protein of interest. Holocarboxylase synthetase (HCS), a known chromatin protein, was used as a model compound in mapping experiments. Upon expression of the HCS-Dam fusion protein in cultured cells, Dam targets the native HCS binding sites and mediates local methylation of adenines. Hence, the sequences near a binding site of HCS will be marked with a unique methylation tag, which can be detected using PCR and microarray based assays. HCS-V5-Dam fusion plasmid, Cbx1-V5-Dam plasmid (positive control), and HCS-only plasmid (negative control) have been successfully cloned. Positive control plasmid has been proven to methylate cells upon expression while the negative control does not contain Dam and hence will not result in methylation. Transfection using eGFP-Luc (encoding green fluorescent and firefly luciferase proteins) plasmid has been optimized to provide transfection efficiencies ranging from 0.5% to 42% by varying transfection conditions (DNA to Lipofectamine 2000 ratio; bolus versus substrate-mediated gene delivery) in the MCF-7 cell line. Transfection using eGFP-Luc is currently being optimized in MCF-10A cells. Once plasmids have been sequence verified and transfection efficiencies optimized by reporter genes, each cloned plasmid will be transfected into either cell line. Sequences methylated by Dam will then be amplified and analyzed by cDNA arrays for gene profiling of MCF-7 and MCF-10A cell lines. Once the technology is established in MCF-7 and MCF-10A cell lines, the study will further continue to research with mouse oocytes. (A contribution of the University of Nebraska Agricultural Research Division, supported in part by funds provided through the Hatch Act. Additional support was provided by NIH grants DK063945, DK077816, DK082476 and ES015206, USDA CSREES grant 2006-35200-17138, and by NSF EPS 0701892.)
SUMO (small-ubiquitin-like modifier in subcellular fractions of Chlamydomonas reinhardtii reveals potential functions
Amy R. Knobbe, Kempton Horken, Stephanie Matejka, Don Weeks and Cheryl P. Bailey
Department of Biochemistry, University of Nebraska-Lincoln

SUMO (Small ubiquitin-like modifier), is a member of the family of ubiquitin-like polypeptides that become covalently attached to intracellular target proteins. SUMOylation can alter the function, location, and/or half-life of target proteins. We have used a bioinformatic approach to identify components of the SUMOylation system in C. reinhardtii. We have identified three SUMO genes and three SUMO-like genes in the genome of C. reinhardtii. The three SUMO-like proteins are novel fusion proteins with C-terminal SUMO-like sequence. Expression levels of the various SUMO and SUMO-like genes vary, although expression of individual genes does not change in response to stress conditions, despite SUMOylation being implicated in stress response. Western analysis shows a shift in proteins recognized by C. reinhardtii SUMO antibodies in response to changes in environmental conditions, such as an increase in temperature or increased salinity. This SUMOylation is rapid and transient.

One of the SUMO proteins identified, SUMO96 appears to localize predominantly to the nucleus, similar to SUMO proteins in other organisms, and in accordance with the fact that one known function of SUMOylation is to direct proteins to the nucleus. One of the other SUMO proteins, however, appears to localize, at least in part, to the flagella. Previous proteomic studies of flagella and cilia have failed to identify SUMO as a component of the flagella, suggesting that this protein is present at low levels in this organelle. In addition, it appears as though SUMOylation in the flagella/cilia is not limited to C. reinhardtii, as cilia fractions from Tetrahymena have revealed SUMOylation as well. We are currently working to identify the proteins that are targets for SUMOylation in the flagella in order to begin to understand the function of SUMOylation in this organelle.
Establishing Influenza A viral mutation database through literature and data mining

Ambreen Kedwaii¹, Pavan Kumar Attaluri², and Guoqing Lu³

¹Department of Pathology and Microbiology, University of Nebraska Medical Center
²Department of Computer Science and Engineering, University of Nebraska at Omaha
³Department of Biology, University of Nebraska at Omaha

Understanding virus evolution is critical for influenza surveillance and vaccine strain selection. Genetic mutation and genome re-assortment are two major evolutionary mechanisms which result in new viral strains with distinct genotypes. The goal for this project is to identify biologically significant mutations that will assist in annotating influenza A sequences. Two approaches were used, including information retrieval and data mining. Informative mutation sites were collected manually from literature as compared with those predicted by the decision tree approach.

In this preliminary study, around 250 sites were found in hemagglutinin of human influenza A viruses from about 50 journal papers, whereas 68 positions were discovered by decision tree, with a total of 36 mutations common in two datasets. A relational database was built to retrieve, manage and use this valuable information. In addition, these mutation sites were used to build a Hidden Markov Model (HMM) model and a web tool was developed to use this model to identify informative mutations of a given sequence. Biological meanings of the predicted mutations will be provided as well. We are in the process to collect more mutation sites to enrich our database and extend web functions for the annotation of influenza A genotypes.

The database and web functions are available at http://glee.ist.unomaha.edu/~pattaluri/mutation/.
Micromechanics Based Constitutive Relationships for Collagenous Tissue
Ranganathan Parthasarthy\textsuperscript{1}, Anil Misra\textsuperscript{2}, Viraj Singh\textsuperscript{3} and Paulette Spencer\textsuperscript{4}
\textsuperscript{1}Bioengineering Graduate Program, University of Kansas
\textsuperscript{2}Department of Civil, Environmental and Architectural Engineering, University of Kansas
\textsuperscript{3}Department of Mechanical Engineering, University of Kansas
\textsuperscript{4}Department of Mechanical Engineering, University of Kansas

Structurally important components of soft tissues like the knee meniscus and articular cartilage are considered to be hydrated collagen fiber network, hydrated proteoglycan (PG) gel and free water. Consequently, we seek constitutive relationships of the type used for pressured fluid held in a porous, pre-tensed medium. Under physiological conditions, the fixed charged density of the proteoglycan gel creates an electrochemical potential difference between the tissue and the bath in which it is submerged. This potential difference creates an osmotic pressure in the fluid. Under no external load, the tissue is in a free swollen state with the tense fibrillar network resisting the osmotic pressure. Under external load, compressive stress is distributed between the fibrillar network, the interstitial fluid and to the PG gel. The overall stress in this poroelastic body is obtained from application of Signorini’s theorem of stress means on the composite of the fibrillar network, interstitial fluid and non-fibrillar PG gel. The stress tensor of the continuous fibrillar network is derived using a surface integral involving all the fiber tractions in a spherical unit cell. Osmotic pressure in the interstitial fluid is calculated using a modified Donnan’s model that concurs with data from existing literature. Stress on the non-fibrillar PG gel is estimated using a modified neo-Hookean model. The pre-tension under the initial free swollen condition is estimated using the equilibrium equation under zero external loads. Expressions for pre-strain in the collagen network under free swollen conditions have been obtained using the derived constitutive law for isotropic as well as anisotropic fiber network. Pre-strain magnitudes are critical in updating the constitutive relation consistent with the buckled microstructure of the fibrillar network. The constitutive relationship together with the pre-strain values is used to perform a parametric study of overall tissue behavior under uniaxial, drained loading conditions that are typically used in experimental testing. Current literature does not explain all mechanical phenomena observed during soft tissue loading using a single constitutive approach. The relationship developed using our microstructural modeling approach simultaneously explains confined and unconfined uniaxial behavior, variation in stress-strain behavior with microstructure, analysis of degenerated cartilage and variation in Poisson’s ratio of the tissue material under compression. Stress-strain behavior under strain controlled uniaxial compression is found to be significantly affected by all the microstructural parameters such as fiber content, fiber stiffness, fiber anisotropy, fixed charge density and external bath concentration. The behavior under uniaxial compression can be delineated into linear, softening and hardening zones with different moduli. This is consistent with observations in the literature that indicate different linear and non-linear zones and varying moduli at different strain ranges during cartilage testing. For large pre-tension, the overall behavior can show softening at intermediate strain-levels. For small pre-tension and fiber stiffness, the overall behavior is dominated by the variation of osmotic pressure in the fluid. The decrease in Poisson’s ratio with applied compressive strain is explained by the model with reference to successive buckling of fibers in the collagen network. All these phenomena follow naturally from the microstructural model without the need for any additional parameters apart from the mechanical properties of each of the constituents.
The unicellular green algae model organism *Chlamydomonas reinhardtii* has been a very useful tool for the study of processes such as photosynthesis and flagellar function over the years but is not quite as amenable as some other model organisms. In yeast for example the phenomenon of homologous recombination (HR) is easily exploited, allows for immense control over the genome and greatly facilitates functional and loss of function studies. In *Chlamydomonas*, HR mediated genomic manipulation is possible but extremely inefficient and is not currently a feasible research tool in this organism. The bacteriophage lambda Beta protein is a recombinase that when expressed in *E. coli* greatly enhances HR mediated by small ≥70 bp ssDNA modifying oligos. In this project, Beta, along with the *E. coli* RecA protein, another well known recombinase, are being expressed either transiently or constitutively in *Chlamydomonas* cells in an attempt to enhance the efficiency of HR mediated by modifying oligos. So far, attempts to repair a point mutation in a marker gene of *Chlamydomonas* using this concept have been unsuccessful. Optimization of the methods being used is currently underway.
An SVM-based Method for Linear B-cell Epitope Prediction by Introducing a new Tri-peptides Kernel  
Bo Yao and Chi Zhang  
School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln  

The identification of B-cell epitopes as antigenic determinants is important in many immunological processes. Therefore, the prediction of B-cell epitopes is highly desirable for various immunological applications, such as vaccine design, immunodiagnostic tests, and antibody production. Previous studies gave rise to some methods for predicting both linear and non-linear B-cell epitopes, such as PREDITOP (Kolaskar and Tongaonkar, 1990), PEOPLE (Alix, 1999), BEPITOPE (Odorico and Pellequer, 2003), and BCPRed (Yasser et al, 2008). However, the limited success of these existing methods does not match the importance of the B-cell epitope prediction. In this work, we designed a new method for linear B-cell epitope prediction, which utilizes the Support Vector Machine (SVM) method with a kernel based on high-scoring pairs of tri-peptides. This kernel has been successfully used in the prediction of protein folds and subcellular localization; hence, we believe that this kernel will be beneficial for the predictive performance in linear B-cell epitopes. According to the leave-one-out cross-validation, our preliminary result showed a predict precision of 70.8% and recall-rate of 68.4%. Further independent test is currently under way.
Pattern Detection in the Motor Protein Prestin: A Collaborative Approach

K. Dempsey\textsuperscript{1,2}, B. Currall\textsuperscript{3}, D. Bastola\textsuperscript{2}, R. Hallworth\textsuperscript{3}, and H. Ali\textsuperscript{1,2}

\textsuperscript{1}Department of Pathology and Microbiology, University of Nebraska Medical Center
\textsuperscript{2}College of Information Science and Technology, University of Nebraska at Omaha
\textsuperscript{3}Department of Biosciences, Creighton University

The advent of high-throughput proteomics has resulted in massive data returns, allowing for the application of experimental and \textit{in silico} techniques to deduce novel structure and function. Often, these analyses include understanding how protein function reflects conserved and unique structures within homologous proteins. One such class of proteins is the solute carrier 26 (Slc26) family of proteins. These proteins play a central role in anion transport and are associated with diseases such as Pendred syndrome and dystrophic dysplasia. Of particular interest is the mammalian Slc26a5 (prestin) protein, which acts not as an ion transporter but as a voltage-dependent motor protein. Prestin shares a high sequence homology with its Slc26 family members and orthologs of other organisms; however, its varied function provides an ideal opportunity to examine the relationship between sequence and structure-function.

Collaboration between traditional and computational research laboratories presents a new and unique opportunity to combine techniques from both areas that result in a more focused search field which in turn leads to more powerful results. Application of Bioinformatics techniques enables us to combine established \textit{in silico} methods such as sequence alignment, pattern detection, and visualization, to predict potentially important motifs in protein structure. In this work, we develop a Bioinformatics approach that utilizes motif detection programs which allows us to further define short specific signals among the discovered motifs obtained from the Gibbs and MEME software tools. These short signals become ideal for focus in mutagenesis experiments, which is made possible through the application of traditional lab techniques in this joint effort. As a result of this collaborative approach, we discovered a set of signals within prestin and its homologs with potential functional impact. With the ability to examine these key patterns from both the \textit{in silico} and \textit{in vitro} vantage points, we take steps toward a quicker and more concentrated identification of critical structure-function relationships within protein families.
A Web-Based Tool for Rigidity Analysis of Proteins
Naomi Fox$^1$, Filip Jagodzinski$^1$, Yang Li$^2$, Ileana Streinu$^{1,2}$

$^1$Department of Computer Science, University of Massachusetts Amherst
$^2$Smith College

Rigidity analysis is a computational method for studying the flexibility of proteins and other biomolecules. A protein is first modeled as a special type of constrained structure called the body-bar-hinge framework, then the pebble game algorithm is run in order to compute the rigid components[2]. We have developed a software library that can process the molecular information into a form suitable for running the pebble game. This library has been written in C++ and can be used in any program. To make the software easily available to biologists, we have also developed a web interface.

Our web-based tool for exploring the rigidity of proteins is available as a link from the Linkage Lab website (http://linkage.cs.umass.edu), starting in October 2009. Visitors to the website can upload a pdb file, run rigidity analysis, download data files, and visualize the protein with rigidity information. Default options are preset for users who want to get started quickly. Advanced users may choose to customize the modeling. For example, there are check-boxes for running the Reduce[4] and Hbplus[3] programs as modules for placing hydrogen bonds. Using drop-down menus, one can customize how the hydrogen bonds and hydrophobic interactions are modeled.

The visualization feature, implemented with Jmol[1], displays the protein with rigid regions. The user can view all, the largest, or just selected rigid regions. The bonds along which torsional motion is expected (the “hinges”) can be highlighted and analyzed. A user can fine-tune their rigidity results by iteratively modifying the modeling parameters and displaying the rigidity visualization. By increasing or decreasing the constraints in the system, the protein will be shown as more rigid or more flexible. Figure 1 shows how the rigid regions of HIV-1 protease are depicted. The left figure shows a cartoon, and on the right, the same protein is shown as a single chain with its rigid regions highlighted.

![Figure 1: (a) HIV-1 protease. (b) The rigid regions of 1HHP (half of hiv-1 protease)](image)

Coevolution between \textit{PKD1} and \textit{PKD2}
Wenhu Guo\textsuperscript{1,} and Chun Ye\textsuperscript{1,2}
\textsuperscript{1}School of Life Science and Engineering, Southwest Jiaotong University
\textsuperscript{2}State Key Laboratory of Biotherapy, West China Hospital, West China Medical School and College of Life Science, Sichuan University

Autosomal Dominant Polycystic Kidney Disease (ADPKD) is one of the most common inherited kidney disease affecting estimated 1 in 400 to 1 in 1000 individuals of the world-wide population. The disease is characterized by progressive development and enlargement of fluid-filled cysts in the kidney and other target organs. In a majority of cases with ADPKD (~85%), the gene involved is \textit{PKD1}, and in the remaining ones (~15%), the disease is attributed to mutations in \textit{PKD2} gene.

The protein products of \textit{PKD1} and \textit{PKD2}, polycystin-1 and polycystin-2, interact with each other through their C-terminal coiled-coil domain and coexpress in many human tissues. Polycystin-1 functions as a membrane receptor that regulates the Ca\textsuperscript{2+} channel formed by polycystin-2 in signaling pathways. Physically interacting proteins or parts of proteins are expected to evolve in a coordinated manner that preserves proper interactions, therefore, the coevolutionary analyses of \textit{PKD1} gene and \textit{PKD2} gene are of great importance for fully understanding ADPKD.

In this study, the evolutionary patterns of \textit{PKD1} gene and \textit{PKD2} gene from nine mammalian species have been analyzed and evidence of coevolution between \textit{PKD1} gene and \textit{PKD2} gene in mammals were provided. Correlated evolution test suggested \textit{PKD1} gene and \textit{PKD2} gene have a high degree of coevolution represented by Pearson’s correlation coefficient, especially between their transmembrane domains. And there is also coevolution intra \textit{PKD1} gene and \textit{PKD2} gene, and the degree of coevolution is relatively high intra \textit{PKD1} gene. For intra \textit{PKD2} gene, coevolution is only be detected between transmembrane domain and C-terminus. The findings are very suggestive when exploring the potential interacting relations between the members of \textit{PKD1} and \textit{PKD2} gene family.
A Software Tool for Surveying the Rigidity Properties of Protein Families
Filip Jagodzinski\textsuperscript{1}, Diana Jaunzeikare\textsuperscript{2}, Naomi Fox\textsuperscript{1}, Ileana Streinu\textsuperscript{1,2}
\textsuperscript{1}Department of Computer Science, University of Massachusetts Amherst
\textsuperscript{2}Smith College

Proteins are the molecular machines that are responsible for almost all of the biological processes in living organisms. In order to perform their functions, proteins flex, bend, contract, and expand. During these internal motions, certain regions of a protein may be flexible, or entire regions may be rigid. Characterizing which regions in a protein are rigid and which are flexible may be the key in understanding the possible motions a protein undertakes to perform its function.

We have written a software library and suite of programs that use a pebble-game algorithm that allows us to quickly and easily determine the rigid and flexible regions of a protein. Our software extends the functionality of FIRST by allowing a user to customize how hydrogen bonds and hydrophobic interactions are modeled. After a rigidity analysis of a protein has been performed, data files are made available to the user, and a Jmol graphical interface offers options to explore the rigid and flexible regions of the protein.

In addition to the rigidity analysis software, we have developed tools for surveying rigidity properties of proteins families. We are interested in determining if protein classes, grouped by structural features as defined by the SCOP database classification scheme, exhibit unique rigidity properties. In this study, we would like to ascertain if rigidity properties of a group of similar proteins can be used to distinguish one set of proteins from another.

Users invoke the surveying feature of RigDynWeb by supplying the SCOP classification identifier of a certain class of proteins. Our program then retrieves the relevant protein coordinate files from the PDB database and performs a rigidity analysis on each protein file. For each group of proteins, we then calculate a set of metrics, including the largest rigid cluster normalized over the number of atoms, the average cluster size, and the number of rigid clusters that contain more than 10 atoms. We use these metrics to characterize the group of proteins, and these metrics are then used to distinguish one protein class from another.
A New Approach to Compare Biological Sequences based on Motif Alphabets

Ramez Mina, Dhundy Bastola and Hesham Ali
College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE 68116

Sequence comparison is one of the major topics in Bioinformatics in general and in sequence analysis in particular. By comparing sequences we identify their similarities, structures and functionalities. While sequence alignment continues to be the dominant tool to compare biological sequences, recent attempts have been made to look into alternative approaches for sequence comparison. This is mainly due to the high time complexity of sequence alignment as well as its inaccuracies in comparing sequences with a low degree of similarity. In addition, being a fine grain approach, alignment is highly sensitivity to sequence errors and other potential imperfections associated with the input sequences.

In this project, we present a new alignment-free method based on converting the input sequences from their single position (nucleotide or amino acid) alphabet to a new alphabet consists of motifs of fixed length. Similarity scores of the original sequences can then be obtained by comparing the motif-based sequences using longest common subsequence and Lempel-Ziv complexity in a pair-wise fashion. The main hypothesis is that the obtained similarity scores provide an accurate measure of sequence comparision as compared to scores obtained from standard multiple sequence alignment. We conducted a set of experiments to validate the proposed hypothesis. We tested the motif alphabet approach using various sets of motifs to classify different datasets. Phylogenetic trees are obtained, and compared to trees obtained from alignment approaches and gold standards trees. Obtained results show that in most of the experiments, the trees obtained from the proposed approach are closer to gold standard trees as compare to the trees obtained from multiple sequence alignments. The results also show that the quality of the trees depends on the properties of the set of motifs selected to construct the new alphabet.
Molecular phylogeny and evolution of pepsinogen subfamily
Michael Wachholtz¹, Jin-Liang Zhao², Guoqing Lu¹
¹Department of Biology, University of Nebraska at Omaha, Omaha, NE, 68182-0040
²Key Laboratory of Aquatic Genetic Resources and Utilization, the Ministry of Agriculture
Shanghai Ocean University, Lingang, Shanghai, China 201306

Pepsinogens are precursors of pepsins belonging to the aspartic proteinase family. These gastric
zymogens are found mainly in vertebrates, but have been characterized mainly in mammals with
respect to their biochemical properties and molecular phylogeny. Due to the lack of sequence data from
non-mammals, evolutionary relationship among vertebrate pepsinogens remains unresolved. To
address this issue, we sequenced five pepsinogens from mandarin fish and Chinese perch and conducted
bioinformatics data mining of pepsinogen sequences from public repositories, and performed a thorough
phylogenetic analysis of over 100 vertebrate pepsinogen protein sequences. The pepsinogen genes consist
mainly of nine exons intervened by eight introns, with the same gene structure in vertebrates.
Comparative analysis of PGC sequences revealed two important motifs respectively in the activation
segment and in the pepsin moiety, indicating different activation mechanisms may be involved in the
pepsins of homeotherms (mammals and avians) and poikilotherms (amphibians and fish). Further
investigation of genomic localization of pepsinogens and their linkage with adjacent genes on genomes of
model fish suggests gene duplication be the major mechanism for the divergence of pepsinogen
subfamily, although genome duplication appears evident in the formation of PGC and PGA.
Lipid-A, a complex glycolipid, is the highly immunoreactive endotoxic center of lipopolysaccharide (LPS). It anchors the LPS into the outer membrane of most gram-negative bacteria. Lipid A can be recognized by animal cells, triggers some defense-related responses, and causes gram-negative sepsis. The lipid-A biosynthesis pathway consists of nine enzymatic steps. The enzymes involved are: LpxA, LpxC, LpxD, LpxH, LpxB, LpxK, WaaA, LpxL, and LpxM. Traditionally bacteria are classified into two major groups as gram-positive and gram-negative based on Gram-staining procedure. Gram-negative bacteria have LPS-containing outer membranes; hence all these bacteria are expected to possess all genes encoding lipid-A biosynthesis-pathway enzymes. These genes, on the other hand, are likely to be missing from gram-positive bacteria. In order to elucidate how the lipid-A biosynthesis pathway has evolved in the bacterial kingdom, we have examined the distribution of the lipid-A biosynthesis enzymes across 62 bacteria genomes (from 16 phyla including five species of gram-positive bacteria as control). Similarity search was performed using BLASTP and TBLASTN, and also profile hidden Markov models. We started our search using the nine enzyme sequences obtained from the \textit{Escherichia coli} K12 genome as queries against the other 61 bacterial genomes. We repeated reciprocal searches against all genomes using each of newly identified protein sequences.

As we expected, none of the five gram-positive bacteria had genes encoding the lipid A biosynthesis enzymes. On the other hand, we found surprisingly a wide range of presence/absence patterns with these genes among 57 gram-negative bacteria we studied. Only one group of gammaproteobacteria (we call it as the group I) had all nine genes required for lipid A biosynthesis. \textit{E. coli} K12 and \textit{Vibrio cholera} are the representatives from this group. The other group of gammaproteobacteria (the group II; e.g., \textit{Pseudomonas syringae}) as well as betaproteobacteria (e.g., \textit{Bordetella parapertussis}, but \textit{Shewanella sp}. MR-4 as one exception) had all except \textit{lpxM} gene. All other gram-negative bacteria are missing both of \textit{lpxH} and \textit{lpxM} genes. This implies that the lipid-A biosynthesis pathway consisting of the nine enzymes is not canonical, but rather a specialized, derived form found only in \textit{E. coli} K12 and closely related group I gammaproteobacteria. The LpxM protein, found only in this group, shared a high similarity with the LpxL protein, and appeared to be a product of duplication. We also found two surprising exceptions among proteobacteria. \textit{Nitrosomonas europaea} (betabacteria) had only one (\textit{lpxK}) of the nine, and two \textit{Walbachia} species (\textit{W. pipiens} and \textit{W. strain TRS of Brugia malayi}) had none of the nine genes. Seven other gram-negative bacteria lacked all of the nine genes although they are classified as gram-negative. Based on the complete lack of the genes required for lipid-A biosynthesis, these gram-negative bacteria could not be distinguished from gram-positive bacteria. Based on phylogenetic analysis using concatenated proteins of lipid A enzymes, \textit{Dictyoglomi} and cyanobacteria were the outmost groups among the gram-negative bacteria. Interestingly, all these bacteria had only four of the nine genes: \textit{lpxA}, \textit{lpxC}, \textit{lpxD}, and \textit{lpxB}. These genes encode the first four enzymes of the lipid-A biosynthesis up to the point of producing the lipid A disaccharide. It indicates that these bacteria may have a different type of Lipid A, which is probably the ancestral status. Another interesting observation was that although \textit{lpxH} gene is considered to be essential in lipid-A biosynthesis, it is missing from the majority of the bacteria we examined. \textit{lpxH} gene, which encodes pyrophosphatase, appeared to have arisen from a duplication of \textit{lpxH2} gene after beta/gammaproteobacteria were diverged from other proteobacteria. \textit{lpxH2} gene encodes phosphatase, and is not involved with lipid-A biosynthesis. Thus this new gene creation was followed by a functional change from pyrophosphatase of LpxH to phosphatase of LpxH2. It is still not clear what enzyme replaces the LpxH activity in bacteria where no \textit{lpxH} gene was found. This study clearly showed that the currently known nine-enzyme pathway for lipid A biosynthesis, which has been mainly studied in \textit{E. coli} K12 and related bacteria, should not be considered as representative nor ancestral to all bacteria.
Chromosome specific detection of NOCTH protein binding sites on Mouse genome using Chip-Seq
Zhijie Jiang¹, Prathibha Ranganathan², Anthony J. Capobianco², and Nicholas F. Tsinoremas¹³

¹ Center for Computational Science, University of Miami
² Sylvester Comprehensive Cancer Center, University of Miami
³ Departments of Medicine, Miller School of Medicine, University of Miami

Chip-Seq is gradually replacing the Chip-Chip on detecting protein binding sites owning to the advantage of next generation sequencing technology. We conducted a Chip-Seq experiment on mouse genome to identify the Notch protein binding sites. Due to the technical artifact of sequencing and the variable binding affinity of Notch protein on different chromosomes, the distribution of reads coverage on different chromosomes is noticeably different. We designed a novel approach to take into account the distinctive distribution of reads coverage on mouse chromosomes. A large amount of peaks identified by this approach are overlapped with Notch binding motif, and also fall in the 5kb promoter regions of numerous genes. Furthermore pathway analysis shows that these genes are involved in many cancer and cancer related pathways. Comparison with other published methods shows that the novel approach detects more peaks within the promoter regions of genes.
Comparative Genome Hybridization Of Maize Inbreds And Their Wild Ancestor, Teosinte

Kai Ying2,3*, Nathan M. Springer1*, Yan Fu4,5, Tieming Ji6, Cheng-Ting Yeh4,7, Yi Jia8, Wei Wu4,7, Todd Richmond9, Jacob Kitzman9, Heidi Rosenbaum9, A. Leonardo Iniguez9, W. Brad Barbazuk10, Jeffrey A. Jeddeloh9, Dan Nettleton6, Patrick S. Schnable2,3,4,5,7,8**

1Department of Plant Biology, University of Minnesota, Saint Paul MN 55108
2Interdepartmental Genetics Graduate Program, 3Department of Genetics, Development and Cell Biology, 4Department of Agronomy, 6Department of Statistics, 7Center for Plant Genomics, 5Center for Carbon Capturing Crops, 8Interdepartment Plant Biology, Iowa State University, Ames, Iowa 50011, 9Roche NimbleGen, Madison, WI 53719, 10University of Florida, Gainesville, FL 32610

Whole-genome array-based comparative genomic hybridization (array-CGH) were performed on two maize (Zea mays) inbred lines, B73 and Mo17, and their wild ancestor teosinte, using Roche NimbleGen custom HD2 maize arrays. Array-CGH data revealed highly conserved genomic regions among maize lines that have very little or no structural variation. These highly conserved regions can be tens of megabases long and may be related to selection sweeps. We also find levels of copy number variation (CNV) and presence/absence variation (PAV) that are unprecedented among higher eukaryotes. A detailed analysis shows those PAV sequence contain hundreds of single-copy, expressed genes that may contribute to heterosis and the extraordinary phenotypic diversity of this important crop.
Phylomarker: mining phylogenetic markers through comparative genomics of model species and data mining of public repositories

Thaine W. Rowley¹, Chenhong Li², and Guoqing Lu³

¹Department of Computer Science, University of Nebraska at Omaha
²School of Biological Sciences, University of Nebraska-Lincoln
³Department of Biology, University of Nebraska-Lincoln

Using numerous gene markers to infer phylogeny is an essential way to resolving the tree of life. Genomic DNA as well as cDNA/ESTs techniques have been explored to increase the number of genes for phylogeny inference. However, only a limited number of genes are currently used as phylogenetic markers. To fill this gap, we developed a bioinformatics pipeline for the identification of novel nuclear gene markers that fulfill a set of requirements brought by molecular systematists through whole genome comparison of model species. We conducted a case study with torafugu and zebrafish genome sequences, resulting in 138 candidate markers. Experimental tests of 15 markers on 36 taxa demonstrated the high feasibility of this approach. Since the result was published in BMC Evolutionary Biology 4:77, we have been continuously requested for source code and some new requirements were made.

In this poster, we will summarize the latest development of the phylomarker project. Specifically, we will introduce the design and implementation of the phylomarker program. It takes as input complete genome sequences in GenBank format and output a summary table of candidate markers with information of sequences, identify and coverage. Besides the standalone program and as required by some users, we have made available a web version of phylomarker, which allows running a small set of sequences. Additionally, to explore sequence data available in public repositories for potential phylogenetic markers, we extended functions in phylomarker to retrieve and parse sequences in GenBank and construct a super matrix (taxa versus gene). The genes appearing in the most taxa were considered as potential markers. However, the usefulness of these genes as phylogenetic marker needs to be further evaluated.

Visit the website at http://bioinfo-srv1.awh.unomaha.edu/phylomarker/pipeline.php to learn more about phylomarker.
Genome-wide comparison of G-protein coupled receptors in model fish species

Kanwal Naeem1, Etsuko N. Moriyama2, Guoqing Lu3

1Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE 68198-3135
2School of Biological Sciences and Center for Plant Science Initiative, University of Nebraska-Lincoln, Lincoln, NE 68588-0118
3Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182-0040

The G-protein coupled receptors (GPCRs) are considered as one of the largest and most ancient protein families. These proteins play essential roles in cellular signal transduction as responding to exogenous stimuli (e.g., odor). The availability of in-depth genome sequence data in many species has provided us unique opportunities to conduct genome-wide comparison of GPCRs in the context of molecular evolution and functional diversification. The objectives of this research are to mine, classify and further annotate GPCRs in the genomes of model fish species and to study evolutionary relationships of GPCRs within and among them. We are interested in fish because it is the most speciose vertebrate group with high diversity in ecology and physiology. Understanding evolution of fish GPCRs would help in many studies, such as genome duplication and species phylogeny.

In this preliminary study, three different bioinformatics resources (7TMRmine, Ensemble and gpDB) were explored to predict and validate GPCRs in torafugu and zebrafish. With 7TMRMine using all provided methods, 1538 and 950 GPCRs were predicted in torafugu and zebrafish, respectively. However, there were only a small set of GPCRs found in Ensemble (30 for torafugu and 61 for Zebrafish) and gpDB (20 for torafugu and 31 for zebrafish). Most GPCRs in Ensemble and gpDB are found to be authentic with experimental evidence in literature. On the poster, we will present the classification result of these predicted GPCRs in different classes and subfamilies using the GPCRTree algorithm. The phylogenetic relationship of the GPCRs found in torafugu and zebrafish will be compared. Finally, the evolution of fish GPCRs will be discussed in terms of gene and genome duplication.
Sorting through the Sequence: Making Sense of Microbial Communities with Amplicon Pyrosequencing.

Ryan Legge, Andrew Benson, and Etsuko Moriyama
Core for Applied Genomics and Ecology (CAGE), Department of Food Science and Technology
University of Nebraska-Lincoln

Next-Generation (Next Gen) sequencing platforms have made it possible to measure compositional features of complex microbial communities on a large scale. Although massive amounts of sequence data can now be produced, algorithms for measuring phylogenetic diversity and species abundances are met with several computational bottlenecks. Our studies compare two methods for analysis of microbial communities based on large amounts of deep pyrosequencing data derived from bacterial 16S ribosomal RNA genes. Sequence data is obtained from PCR products of complex mixtures of genomic DNA that are extracted from the communities. The ultimate goal of the analysis is to use the sequences to quantify the relative abundances of different microbial taxa represented by the sequences and to estimate the relative degree of phylogenetic diversity that is represented.

To measure the relative abundances of different taxa in two or more microbial communities, we use a taxonomy-dependent and a taxonomy independent approach. In the taxonomy-driven approach, we capitalize on the Ribosomal Database Project (RDP) CLASSIFIER algorithm. This algorithm assigns 16S rRNA gene sequences to a phylogenetically consistent higher-order bacterial taxonomy hierarchy using a naïve Bayesian rRNA classifier. This approach depends on the known taxonomic relationships of the training set. The algorithm works by comparing frequencies of all eight-base subsequences to the sixty-four thousand possible eight-base subsequences in each of the approximately 880 known genera (Wang, et al.) Once taxonomic status is assigned to a sequence, the numbers of sequences representing each taxon in each sample is determined, followed by identification of statistical analysis for treatment effects on individual taxonomic ranks using ANOVA.

The second abundance-related approach is independent of known taxonomic relationships and is based on the phylogenetic relationships that can be inferred when sequences from all samples of an experiment are compared to one another. Although the most robust approach would be alignment-based, true alignment is impractical for the massive numbers of sequences from pyrosequencing. Therefore, K-mer-based approaches are often substituted. We have employed the CD-HIT-EST strategy which uses K-mers and clustering to establish clusters that meet a defined similarity threshold or sequence identity. The method uses a greedy incremental clustering algorithm which first sorts the sequences in order of decreasing length. The longest one becomes the representative of the first cluster. Then, each remaining sequence is compared to the representatives of existing clusters. If the similarity with any representative is above a given threshold, the sequence is grouped into an existing cluster. Otherwise, a new cluster is defined with that sequence as the representative. (Weizhong L. et al.) The clusters represent taxonomic units (Operational Taxonomic Units or OTUs). Behavior of the OTUs in response to experimental variables is then measured statistically by ANOVA. The identity of OTUs with statistical significance can then be determined by comparing representative sequences with the Small Subunit rRNA database using SeqMatch.

To compare microbial communities on the basis of the relative amount of phylogenetic diversity present, phylogenies are estimated from sequences representing each of the taxonomic units present in the samples. To estimate the degree of phylogenetic diversity that remains, the dereplicated sequences are then aligned (Clustal W), followed by distance calculation and Neighbor Joining analysis in the ARB package. ARB allows for visualization of dense phylogenies and provides functionality for visualizing relative distributions of taxa from different samples in the same dendrogram (Ludwig W. et al.) Variation between microbial communities can also be measured based on branch-length can also be measured and visualized using the UniFrac analysis which computes differences between microbial communities based on phylogenetic information by measuring the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both. (Lozupone C. et al.)

Reference:
Proteomics reconstruction of chemotaxis and acrosome reaction signaling pathways in sea star sperm
Yuyan Shelly Xue¹ and Hideaki Moriyama²
¹Department of Computer Science and Engineering, University of Nebraska-Lincoln
²School of Biological Sciences and Center for Biotechnology, University of Nebraska-Lincoln

Although the biological features of spermatozoa are complex, only a limited number of sperm proteins have been identified in human sperm. Our preliminary proteomic study also identified only 75 proteins from sea star sperm tail. Furthermore, fertilization process is known to proceed without transcription of mRNAs. Therefore, the roles of sperm proteins during fertilization cannot be investigated simply by examining the mRNA profile in sperm (e.g., by Expressed Sequence Tag or EST sequencing). Alternatively, proteomics analysis should provide us more direct data on this complicated process. Mass spectrometry is one of the few methods that allow such comprehensive and direct analysis of proteins existing in sperm.

Furthermore, supporting evidence for our signaling pathway hypothesis (Nakachi et al., 2008) can be obtained by investigating the protein content of sea star sperm as functions of time and space. Sea star sperm will be exposed with acrosome inducing materials including acrosome reaction-inducing substance (ARIS), Co-ARIS, and the asteroidal sperm activating peptide (ASTEROSAP). The sperm separated from the media will be then analyzed along a time course.

Preparation of sperm sample. Soluble and membrane fractions are prepared from lysates with no detergent and with more detergents: for example, Triton X-100 (octyl phenol ethoxylate), SM1200 (sucrose monolaurate), and SDS (sodium dodecyl sulfate) at concentrations around three times of CMC (critical micelle concentration).

Tandem mass spectrometry (MS/MS) analysis. The flagella (or head) protein crude extract will be mixed with 1 volume of 2 x sample buffer (4% SDS, 20% glycerol, 10% Bromophenol blue, 10% β-mercaptopoethanol and 125 mM Tris, pH6.8) and denatured by boiling for 5 min at 100 ºC. The sample will be separated by SDS-PAGE and 20 fractions will be cut out from the gel. Each fraction will be digested with trypsin and resulting peptide fragments will be separated by liquid chromatography, then subjected to MS/MS analysis. Observed MS/MS ions will be searched against databases using Mascot software. Significant protein hits will be identified if they match against more than one peptide with the score of 50. The tryptic mass spectrometry will be done in Mass Spectrometry Core Facility (Manager, Dr. Cerny), Center for Biotechnology, University of Nebraska-Lincoln.

Expected outcomes and alternative strategy. After successful completion of this specific aim, we will have complete inventories of proteins existing in different timing and space during acrosome reaction. There has been no such study in the past. It allows us to find how acrosome reaction inducing substances bind to their receptors and lead to acrosome reaction.

Preliminary studies indicated only 20% of protein fragment could be recovered after the tryptic digestion. This is due, in part, to the less exposure of proteins against trypsin. We will put a detergent solubilization process in the gel using octyl-β-D-glucoside or other small detergents to relax the hydrophobic part of the protein. In the mass spectrometry, the molecular concentration ratio of the flying substance defines the signal and noise quality. In case 1 M salt ion and 0.001 M protein (peptide) ion coexist, the protein will produce only 1/1000 signal in the mass spectrometry. Since the acrosome reaction needs to be done in seawater, which contains 0.5 M NaCl, we will have to remove seawater completely. For this purpose, the gel electrophoresis with abundant buffer without sodium chloride needs to be done.

Indigenous medicinal plants containing active and medicinal principles like glycerides, alkaloids, steroids, tannins etc. grow abundantly in Bangladesh. These indigenous medicinal plants are extremely used in both raw and semi-processed forms in the preparation of pharmaceutical, Homeopathic, Unani, and Ayurvedic medicines. Although our country is rich with this vast natural resource but due to lack of knowledge none processes these indigenous medicinal plants or its extracts locally. As a result every year Bangladesh imports a huge quantity of processed indigenous medicinal plants or its extracts from abroad at the cost of our foreign exchange to meet the country’s demand. So, efforts have been made to systematic processing and screening of indigenous medicinal plants as pharmaceutical raw materials. The present paper deals with tested a large number of indigenous medicinal plants from Lawacherra Rain Forest of Bangladesh origin on healthy and streptozotocin induced Type 1 and Type 2 diabetes mellitus model rats at different prandial states; fasting, simultaneously with glucose load and thirty minutes before glucose load. A large number of indigenous medicinal plant extracts were found to possess significant hypoglycemic effects on different models and different prandial states. Hypoglycemic extracts of *Abrus precatorius* [Josti-modhu], *Ocimum gratissimum* [Raam-tulshe], *Withania somnifera* [Choto-chada], and *Aloe vera* [Gritho-kumari] were also found to have chronic effect in controlling serum glucose and dislipidemia on feeding for twenty eight days. Two sub-fractions of active extract of *Abrus precatorius* and *Withania somnifera* showed insulin releasing effect in pancreatic rat islets and perfused whole rat pancreas. Large-scale isolation of one sub-fraction is going on for chemical trial. *Azadirachta indica* [Neem], *Abrus precatorius* [Josti-modhu], *Achyranthes aspera* [Upoth-lenra], *Coccinia cordifolia* [Rakhal-shosha], *Abroma augusta* [Ulot-kombol], *Alpinia galangal* [Baow-ada], *Wedelia chinensis* [Vhingoraaz], *Scoparia dulcis* [Chinegura], *Murraya koenigii* [Norosing], *Trigonella foenum-graceum* [Methe], *Ipomoea mauritiana* [Vhui-kumura], and *Withania somnifera* [Choto-chada] several secondary metabolites including steroids, triterpenes, flavonoids, megastigmanes, benzohydrofurans and their glycosides and phenylethanoid glycosides were isolated. The hypoglycemic extracts/pure compounds of *Azadirachta indica*, *Scoparia dulcis*, *Withania somnifera*, and *Achyranthes aspera* also showed antioxidant, antifungal, and antibacterial activities. Processing and phyto-chemical screening of some of these indigenous medicinal plants are also carried out by our developed technologies. The work will increase the potentiality of indigenous medicinal plants for the production of various pharmaceutical raw materials and new drugs of our country to a large extent.

Keywords: Lawacherra Rain Forest, Unani, Ayurvedic, Homeopathic, indigenous medicinal plants, hypoglycemic effects.
Development and use of the AllergenOnline.org curated protein database for assessing the potential allergenicity of genetically modified organisms and novel food ingredients
John Wise¹, Steve L. Taylor¹, and Richard E. Goodman¹
¹Department of Food Science & Technology, University of Nebraska- Lincoln

Background: New proteins are being introduced into foods through genetic engineering of organisms and by addition through food processing as novel food ingredients. The potential that the transferred proteins might significantly increase the risk of food allergy due to similarities to known allergens must be assessed under regulations in many countries, including the US (Codex Alimentarius Commission, 2003). The primary step is a bioinformatics comparison of the amino acid sequence of the new protein against those of known allergens by BLAST or FASTA to identify potentially cross-reactive proteins, which may be done using the NCBI Entrez Protein database (more than 16,898 “allergen” sequences listed on 17 August, 2009). Alternatively the primary search may be performed using a specific allergen database. We developed the AllergenOnline database in 2004 and have implemented a peer review process (in 2006) to accept or reject the claims of allergenicity for the “allergenic” proteins based on set criteria. A panel of 7 internationally recognized allergy experts review published information and vote on each group of allergens. Version 9.0 (2009), with 1386 sequences listed as proven or putative allergens. In addition, the AllergenOnline.org database provides full-length FASTA, a sliding 80 amino acid window FASTA and a highly conservative 8 amino acid matching test search tools to identify matches based on Codex guidelines.

Objectives: 1) To describe the peer-review process for including protein sequences as allergens, as used to build version 9.0 of AllergenOnline. 2) To evaluate the 3 newly expressed proteins in genetically engineered Golden Rice 2, a high beta-carotene rice intended to reduce vitamin A deficiency in rice-based cultures of Asia and compare sequence identity matches to those known to cause allergic cross-reactivity.

Methods: A candidate list of 1758 “allergens” with some level of documentation was compiled in June 2006 from NCBI using keyword searches, plus entries in the Allergen Nomenclature list of the International Union of Immunological Societies (IUIS) and AllergenOnline version 6. Sequences were grouped based on taxonomy and sequence identities. Peer reviewed articles were identified for each group. Criteria were developed to evaluate the evidence based on in vivo and in vitro tests using subjects with clinically defined, relevant allergic disease. Published data for each group was compiled and triaged to identify data gaps, then classified by primary and secondary reviewers. The full panel of experts reviewed the completed classification set, removing groups with insufficient evidence of allergenicity (no IgE binding or in vivo reactivity, with subjects having at least minimal descriptions of allergies to the source of the protein).

Results: AllergenOnline version 9.0 (January 2009) contains 1386 sequences supported by published data demonstrating at a minimum, evidence of specific IgE binding using sera from subjects having at least minimal allergic symptoms when exposed to the source of the protein. Bioinformatics comparison of the three proteins introduced into rice through genetic engineering to produce Golden Rice 2 did not identify any similarity to known allergens using Codex criteria, while proteins of known cross-reactive activity are efficiently identified using AllergenOnline.

Conclusion: The curated AllergenOnline database provides an easy to use tool for developers of new genetically modified crops and novel foods to identify potential risks of allergy before investing significant time and money on potential products that would pose an unacceptable risk for consumers.
Species-Specific Duplications of Trace Amine-Associated Receptors in Vertebrate Genomes

Seong-il Eyun¹ and Etsuko N. Moriyama¹, ²
¹School of Biological Sciences, University of Nebraska-Lincoln
²Center for Plant Science Innovation, University of Nebraska-Lincoln

Chemosensory receptors (CRs) are used to detect a wide range of chemicals and are a crucial gateway between perception and environment. Expansion or loss of CRs would reflect the adaptation to the organism's life at the molecular level. Life-history traits such as foraging behavior and type of foods are expected to play a central role in driving variation in the number of CRs. Evolution of CRs may be also correlated with that of other detection functions (e.g., vision and audition). CRs, especially the olfactory receptors (ORs), represent the largest gene family in animal genomes (Nei et al., 2008). Vertebrate ORs are predominantly expressed in the main olfactory epithelium (MOE) in the nasal cavity. Interestingly, Liberles and Buck (2006) demonstrated that mouse Trace Amine-Associated Receptors (TAARs) are also expressed in the MOE and can function as ORs for volatile amines found in urine. TAAR-expressing olfactory sensory neurons could be involved in the detection of social pheromone cues such as β-phenylethylamine, isoamylamine, and trimethylamine that elicit innate behaviors. In order to examine how the number of TAARs varies among vertebrates and elucidate the evolutionary history of these proteins, we identified all TAARs from twenty two vertebrate genomes. We found that the number of TAAR genes is much smaller than the number of ORs. On the other hand, extremely large variation in the number of TAARs among vertebrate species was observed. For example, while the megabat (Pteropus vampyrus) genome carries the largest number of TAARs (26 genes and 10 pseudogenes) among the mammals we studied, the dolphin (Tursiops truncatus) genome has no TAAR gene (3 pseudogenes). The chicken and lizard (Gallus gallus and Anolis carolinensis) genomes have also smaller numbers of TAAR genes (4 and 3, respectively). The numbers of TAARs in this study showed 13-fold difference among the tetrapod mammals, ranging from 2 in dog to 26 in megabat. This range is much larger than that of OR genes, which showed only 4-fold difference, from ~326 in macaques to 1,259 in rats (Nei et al., 2008).

Gene duplication has been counted as a major mechanism in molecular evolution. We found that four TAAR paralogs (TAAR6, TAAR7, TAAR8, and TAAR9) are characterized with recent and extensive tandem duplications. For instance, the gene duplication rate of TAAR7 in rat is 0.043 (duplications per gene per million years), which is 16 to 33-fold higher than their average gene duplication rate (0.0013–0.0026 per gene per million years; Gibbs et al., 2004). Although many of the GPCR families are characterized with gene duplication and divergence, TAARs show the case of extreme. All TAARs identified in this study form a monophyletic clade and are clearly distinguishable from other biogenic amine receptors. Although TAARs are regarded as a second class of ORs, their sequences are similar to classical biogenic amine receptors but not to ORs. Two TAARs are present in the elephant shark genome (Callorhinchus milii). A lamprey, two tunicates, and a lancet genomes do not contain TAARs, although they have biogenic amine (serotonin) receptors. Therefore, the origin of TAARs seems to be traced back to the common ancestor of jawed vertebrates (cartilaginous fishes). The pairwise distances calculated from CRs are normally distributed except for TAARs, whose distribution is bimodal (P < 1 × 10⁻⁶) implying that TAARs are highly conserved among orthologs but diverged between paralogs. While ORs are scattered among many chromosomes, the TAARs and other adjacent genes are located in a gene cluster in vertebrates. The syntenic relationships of these genes are highly conserved among vertebrates, although the order is inverted in mammalian.

References
Discovering SNP association patterns with spatial constraint using data mining approach
Fangrui Ma¹, Liying Jiang¹, Etsuko Moriyama², and Jitender S. Deogun¹

¹Computer Science and Engineering Department, University of Nebraska-Lincoln
Lincoln, NE 68588-0115. Email: fma{ljiang, deogun}@cse.unl.edu
²School of Biological Science and Center for Plant Science Innovation, University of Nebraska-Lincoln. Email: emoriyama2@unl.edu

In this paper we present a very useful tool for discovering SNP association patterns using data mining approach. The algorithm also takes spatial constraint of a pattern into consideration. We first combine the position and character information of a SNP into a new item. Therefore the original matrix of SNP sequences is transformed into a new matrix with an element having both spatial and SNP information. We modified association mining algorithm, Apriori, and applied it to this matrix to find a set of maximum patterns with user specified support. This tool can offer biologist with very useful information for the research in many problems like haplotype blocks, linkage disequilibrium, and phylogeny.

The preliminary experiment suggests that the algorithm, SNP-PAT, can identify the patterns correctly. However, the algorithm requires large storage. We are currently working on improving storage requirements and the algorithm.

Algorithm SNP-PAT
Input: an m x n SNP matrix, minimum support c_min, and a SNP X
Output: SNP pattern collection L

For each sequence i (0 ≤ i ≤ m-1)
   For each character x_{ij} ∈ {A, C, G, T} (0 ≤ j ≤ n-1)
       change x_{ij} into x_{ijj}

Let C_k be set of candidates k-itemsets, and
L_k be set of frequent k-itemsets

L_1 = {frequent 1-itemsets}
For (k = 2; C_k ! empty; k++) do
   begin
      C_k = \{I_{i,j} | \forall i, j, I_{i,j} \in L_k \}
      for each c_k \in C_k do
         begin
            if sup(c_k) > c_min
            L_k = L_k \cup \{c_k \}
         end
   end
For each item I_k \in L_k do
   I_k \in L , If I_k contains SNP X
Return L
A web application for the analysis of differential Histone modification states from ChIP-seq data

Sridhar A Malkaram1, Rong Liu1, Michael Fromm1, Jean-Jack M Riethoven1
1Center for Biotechnology, University of Nebraska-Lincoln

The modifications of histones in chromatin including covalent modifications like methylation, acetylation and phosphorylation are known to be associated with up/down regulation of gene expression1. A comparison of genome-wide differences in the histone modification levels can help in the prediction of the set of genes that are up or down regulated under a biological condition such as water stress. The ChIP-seq technique is now being widely used to find the level of histone modifications across the genome. A crucial part of data analysis comprises of finding the enriched regions and estimating their statistical significance. Several methods have been developed to identify the enriched regions or peaks. Basically, the peaks are found by counting the number of reads in a window and compared with a control2. In some methods, the positive and negative strand reads are counted separately to better locate the peak centers3. Three types of peaks - sharp, broad and mixed are found in the enriched regions and most of the current algorithms are designed to find the sharp peaks which generally represent the protein-DNA binding sites4. Efforts are underway to unify the detection of sharp along with peaks that are associated with histone modifications at transcribed regions or repressed regions4.

We present a methodology wrapped into a user friendly web application, to identify the histone modification regions that are known to be associated with transcription. The peak identification is based on identifying coverage regions above a threshold height and width. The assumptions behind using the threshold height and width are the enrichment of the coverage over the control DNA and the expected minimum width of the nucleosome bound DNA respectively. The coverage data are initially normalized by scaling the mean coverages to a common value. The coverage profiles are then smoothed with the running average method. The False Discovery Rates (FDRs) of peak identification were determined by calculating the number of peaks identified in the control DNA for a specific height and width. The differences in the peak profiles of two different experiments are found by calculating the differences in the coverage of the overlapping peaks. The FDRs of difference identification were determined from the number of differences found among the replicates compared to the number of differences between the experiments. The application annotates the results with genome features such as gene names, transcription starts, exons, introns etc. with hyperlinks to be viewed in the genome browser.

Comparative miRNA analysis in pathogenic fungi

Kanika Jain, B.B. Chattoo

Center for Genome Research, M. S. University of Baroda

**RNA interference (RNAi)** is a system within living cells that helps to control which genes are active and how active they are. Two types of small RNA molecules – microRNA (miRNA) and small interfering RNA (siRNA) – are central to RNA interference.

siRNAs derive from long double stranded RNA precursors and are perfectly complementary to a particular target mRNA and cleaves that target, whereas miRNAs are generated from short hairpin RNA precursors and are partially complementary to the target mRNA.

MicroRNAs (miRNAs) are small single-stranded RNA molecules with a length of 21-23 nucleotides that regulate gene expression post-transcriptionally. They have a regulatory function that is down regulation of gene expression either by degradation of messenger RNA or by inhibition of protein translation.

Both pathways are found in plants and animals. However, in fungi, evidence exists only for siRNA pathway.

Although no experimental evidence for the presence of miRNA pathways in fungi has been published, various developmental processes point to the possibility that miRNAs are also encoded in the genome of fungi. Unfortunately, traditional gene expression techniques like Northern blot, reverse transcription PCR (RT-PCR), SAGE, and DNA microarrays are not always applicable to predict miRNAs due to the lack of sensitivity and specificity. Because of these disadvantages of experimental approach, computational methods have been developed to predict new miRNAs in worms, flies, and humans. This project was undertaken 1) to computationally predict the microRNA sequence patterns in the genome of 22 fungi using miRNA Search Tool, 2) to identify the target genes for predicted miRNAs using target prediction programs and, 3) to perform functional annotation of the target genes using various web resources.

The results obtained from the analysis predicted microRNA sequences which shared sequence homology with some of the plant and animal miRNA. Most of the genes targeted by predicted microRNA were found to be highly conserved in pathogenic fungi and were involved in transcription regulation, DNA binding, ATP binding, protein and zinc ion binding while many other genes were hypothetical. It can be concluded from the analysis that microRNAs are well distributed among fungal genomes and may play an important role in controlling expression of genes which are responsible for transcription regulation, DNA binding, ATP binding and many other important physiological processes.
<table>
<thead>
<tr>
<th>Author</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmmed, Bulbul</td>
<td>115</td>
</tr>
<tr>
<td>Aldape, Kenneth D</td>
<td>13</td>
</tr>
<tr>
<td>Ali, H</td>
<td>102</td>
</tr>
<tr>
<td>Ali, Hesham</td>
<td>106</td>
</tr>
<tr>
<td>Anantharamanan, R</td>
<td>80</td>
</tr>
<tr>
<td>Ashton, Evan</td>
<td>82</td>
</tr>
<tr>
<td>Attaluri, Pavan Kumar</td>
<td>21, 98</td>
</tr>
<tr>
<td>Badruddaza, Md</td>
<td>115</td>
</tr>
<tr>
<td>Bailey, Cheryl P</td>
<td>97</td>
</tr>
<tr>
<td>Barbazuk, W. Brad</td>
<td>110</td>
</tr>
<tr>
<td>Bashyam, M. D</td>
<td>81</td>
</tr>
<tr>
<td>Bastola, D</td>
<td>102</td>
</tr>
<tr>
<td>Bastola, Dhundy</td>
<td>106</td>
</tr>
<tr>
<td>Bednarz, Jane</td>
<td>82</td>
</tr>
<tr>
<td>Benson, Andrew</td>
<td>5, 113</td>
</tr>
<tr>
<td>Bereg, Sergey</td>
<td>91</td>
</tr>
<tr>
<td>Bhattacharjee, Biplab</td>
<td>80</td>
</tr>
<tr>
<td>Broom, Bradley M</td>
<td>13</td>
</tr>
<tr>
<td>Brueggeman, Andrew</td>
<td>87</td>
</tr>
<tr>
<td>Capobianco, Anthony J</td>
<td>109</td>
</tr>
<tr>
<td>Cerutti, Heriberto</td>
<td>79</td>
</tr>
<tr>
<td>Chatterjee, Subhojoyoti</td>
<td>80</td>
</tr>
<tr>
<td>Chattoo, B.B.</td>
<td>120</td>
</tr>
<tr>
<td>Chen, Zhengxin</td>
<td>21</td>
</tr>
<tr>
<td>Choubey, Deepak</td>
<td>80</td>
</tr>
<tr>
<td>Chowdhary, Anirudh</td>
<td>80</td>
</tr>
<tr>
<td>Chowdhury, Anup</td>
<td>115</td>
</tr>
<tr>
<td>Chu, Cheuk-Man</td>
<td>85</td>
</tr>
<tr>
<td>Connare, Sanjay</td>
<td>82</td>
</tr>
<tr>
<td>Cook, Dianne</td>
<td>83, 89</td>
</tr>
<tr>
<td>Currall, B.</td>
<td>102</td>
</tr>
<tr>
<td>Dempsey, K.</td>
<td>102</td>
</tr>
<tr>
<td>Deogun, Jitender S</td>
<td>50, 118</td>
</tr>
<tr>
<td>Dhawan, Manik</td>
<td>28</td>
</tr>
<tr>
<td>Do, Kim-Anh</td>
<td>13</td>
</tr>
<tr>
<td>Du, Lingcheng</td>
<td>84</td>
</tr>
<tr>
<td>Duan, Zhong-Hui</td>
<td>28, 88, 95</td>
</tr>
<tr>
<td>Duax, William L</td>
<td>82</td>
</tr>
<tr>
<td>Dziak, David</td>
<td>82</td>
</tr>
<tr>
<td>Edgerton, Mary E</td>
<td>13</td>
</tr>
<tr>
<td>Ewing, Kailly</td>
<td>41</td>
</tr>
<tr>
<td>Eyun, Seong-il</td>
<td>117</td>
</tr>
<tr>
<td>Fox, Naomi</td>
<td>103, 105</td>
</tr>
<tr>
<td>Fromm, Michael</td>
<td>119</td>
</tr>
<tr>
<td>Fu, Yan</td>
<td>110</td>
</tr>
<tr>
<td>Fulk, Brandon L</td>
<td>73</td>
</tr>
<tr>
<td>George, Daliny A</td>
<td>85</td>
</tr>
<tr>
<td>Gibas, Kevin</td>
<td>82</td>
</tr>
<tr>
<td>Goodman, Richard E</td>
<td>116</td>
</tr>
<tr>
<td>Gu, Wenyi</td>
<td>90</td>
</tr>
<tr>
<td>Guo, Wenhu</td>
<td>104</td>
</tr>
<tr>
<td>Hallworth, R.</td>
<td>102</td>
</tr>
<tr>
<td>Hassan, Azmal Ibna</td>
<td>115</td>
</tr>
<tr>
<td>Hennig, Fiona</td>
<td>82</td>
</tr>
<tr>
<td>Hines, Lisa</td>
<td>41</td>
</tr>
<tr>
<td>Hofmann, Heike</td>
<td>83, 89</td>
</tr>
<tr>
<td>Hogan, Dana</td>
<td>82</td>
</tr>
<tr>
<td>Honavar, Vasant G</td>
<td>4</td>
</tr>
<tr>
<td>Horken, Kempton</td>
<td>97</td>
</tr>
<tr>
<td>Hu, Haiyan</td>
<td>65</td>
</tr>
<tr>
<td>Huang, Yuanyuan</td>
<td>83</td>
</tr>
<tr>
<td>Huether, Robert</td>
<td>82</td>
</tr>
<tr>
<td>Inguez, A. Leonardo</td>
<td>110</td>
</tr>
<tr>
<td>Islam, Md. Torikul</td>
<td>115</td>
</tr>
<tr>
<td>Jagodziński, Filip</td>
<td>103, 105</td>
</tr>
<tr>
<td>Jain, Kanika</td>
<td>120</td>
</tr>
<tr>
<td>Jaunzeikare, Diana</td>
<td>105</td>
</tr>
<tr>
<td>Jayadepa, R. M.</td>
<td>80</td>
</tr>
<tr>
<td>Jeddeloh, Jeffrey A</td>
<td>110</td>
</tr>
<tr>
<td>Ji, Tiemig</td>
<td>110</td>
</tr>
<tr>
<td>Jia, Yi</td>
<td>110</td>
</tr>
<tr>
<td>Jiang, Litying</td>
<td>118</td>
</tr>
<tr>
<td>Jiang, Zhijie</td>
<td>109</td>
</tr>
<tr>
<td>Kalita, Jugal</td>
<td>41</td>
</tr>
<tr>
<td>Kedwai, Ambreen</td>
<td>98</td>
</tr>
<tr>
<td>Kirsch, Tyler</td>
<td>82</td>
</tr>
<tr>
<td>Kitzman, Jacob</td>
<td>110</td>
</tr>
<tr>
<td>Klein, Luke</td>
<td>82</td>
</tr>
<tr>
<td>Knobbe, Amy R</td>
<td>97</td>
</tr>
<tr>
<td>Ladunga, Istvan</td>
<td>87</td>
</tr>
<tr>
<td>Lai, Yinglei</td>
<td>33</td>
</tr>
<tr>
<td>Lane, Charlotte</td>
<td>82</td>
</tr>
<tr>
<td>Lawrence, Michael</td>
<td>89</td>
</tr>
<tr>
<td>Legge, Ryan</td>
<td>113</td>
</tr>
<tr>
<td>Li, Chenhong</td>
<td>111</td>
</tr>
<tr>
<td>Li, Ming</td>
<td>86</td>
</tr>
<tr>
<td>Li, Xiaoman</td>
<td>65</td>
</tr>
<tr>
<td>Li, Yang</td>
<td>103</td>
</tr>
<tr>
<td>Li, Yaoyao</td>
<td>84</td>
</tr>
<tr>
<td>Limbu, Sarita</td>
<td>95</td>
</tr>
<tr>
<td>Liu, Rong</td>
<td>119</td>
</tr>
<tr>
<td>Lou, Lili</td>
<td>84</td>
</tr>
<tr>
<td>Louw, Tobias M.</td>
<td>71</td>
</tr>
<tr>
<td>Lu, Guoqing</td>
<td>21, 98, 107, 111, 112</td>
</tr>
<tr>
<td>Ma, Fangrui</td>
<td>50, 118</td>
</tr>
<tr>
<td>Macrander, Jason</td>
<td>94</td>
</tr>
<tr>
<td>Maiman, Daniel</td>
<td>82</td>
</tr>
<tr>
<td>Malkaram, Sridhar A</td>
<td>119</td>
</tr>
<tr>
<td>Matejka, Stephanie</td>
<td>97</td>
</tr>
<tr>
<td>McEachon, Courtney</td>
<td>82</td>
</tr>
<tr>
<td>McField, Romeo</td>
<td>115</td>
</tr>
<tr>
<td>Author Index</td>
<td>Page(s)</td>
</tr>
<tr>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>McMillan, Nigel A. J.</td>
<td>90, 92</td>
</tr>
<tr>
<td>Meeker, William Q.</td>
<td>86</td>
</tr>
<tr>
<td>Middha, Sushil K.</td>
<td>80</td>
</tr>
<tr>
<td>Mina, Ramez</td>
<td>106</td>
</tr>
<tr>
<td>Misra, Anil</td>
<td>99</td>
</tr>
<tr>
<td>Mollik, Md. Ariful Haque</td>
<td>115</td>
</tr>
<tr>
<td>Moriyama, Hideaki</td>
<td>108, 114</td>
</tr>
<tr>
<td>Muppaneni, Sri Harsha</td>
<td>88</td>
</tr>
<tr>
<td>Naeem, Kanwal</td>
<td>112</td>
</tr>
<tr>
<td>Nawandar, Dhananjay M.</td>
<td>93</td>
</tr>
<tr>
<td>Nettleton, Dan</td>
<td>110</td>
</tr>
<tr>
<td>Opiyo, Stephen O.</td>
<td>73, 108</td>
</tr>
<tr>
<td>Orti, Guillermo</td>
<td>94</td>
</tr>
<tr>
<td>Panner, Angela</td>
<td>96</td>
</tr>
<tr>
<td>Pardy, Rosevelt L.</td>
<td>108</td>
</tr>
<tr>
<td>Parthasarthy, Ranganathan</td>
<td>99</td>
</tr>
<tr>
<td>Pauley, Mark</td>
<td>9</td>
</tr>
<tr>
<td>Peck, Scott</td>
<td>6</td>
</tr>
<tr>
<td>Plucinak, Thomas M.</td>
<td>100</td>
</tr>
<tr>
<td>Polapragada, Teja</td>
<td>88</td>
</tr>
<tr>
<td>Powell, Rasheen</td>
<td>82</td>
</tr>
<tr>
<td>Ranganathan, Prathibha</td>
<td>109</td>
</tr>
<tr>
<td>Richmond, Todd</td>
<td>110</td>
</tr>
<tr>
<td>Riefler, Brian</td>
<td>82</td>
</tr>
<tr>
<td>Riethoven, Jean-Jack M</td>
<td>119</td>
</tr>
<tr>
<td>Robins, Derek</td>
<td>82</td>
</tr>
<tr>
<td>Rosenbaum, Heidi</td>
<td>110</td>
</tr>
<tr>
<td>Rowley, Thaine W</td>
<td>111</td>
</tr>
<tr>
<td>Rutkowski, Emma</td>
<td>82</td>
</tr>
<tr>
<td>Ryan, Patrick</td>
<td>82</td>
</tr>
<tr>
<td>Sayood, Khalid</td>
<td>3</td>
</tr>
<tr>
<td>Schnable, Patrick S.</td>
<td>110</td>
</tr>
<tr>
<td>Schunk, Adam</td>
<td>82</td>
</tr>
<tr>
<td>Scott, Stephen D.</td>
<td>10, 69, 75</td>
</tr>
<tr>
<td>Selvaraja, Sudarshan</td>
<td>28</td>
</tr>
<tr>
<td>Shadique, Md. Saroar</td>
<td>115</td>
</tr>
<tr>
<td>Shaik, Ameer</td>
<td>88</td>
</tr>
<tr>
<td>Singh, Dipika</td>
<td>96</td>
</tr>
<tr>
<td>Singh, Rakesh K</td>
<td>93</td>
</tr>
<tr>
<td>Singh, Viraj</td>
<td>99</td>
</tr>
<tr>
<td>Singhania, Akul</td>
<td>92</td>
</tr>
<tr>
<td>Singhania, Richa</td>
<td>90</td>
</tr>
<tr>
<td>Soh, Leen-Kiat</td>
<td>10</td>
</tr>
<tr>
<td>Spencer, Paulette</td>
<td>99</td>
</tr>
<tr>
<td>Springer, Nathan M</td>
<td>110</td>
</tr>
<tr>
<td>Sriram, Aparna</td>
<td>88</td>
</tr>
<tr>
<td>Stoutenburg, Suzette K</td>
<td>41</td>
</tr>
<tr>
<td>Streinu, Ileana</td>
<td>103, 105</td>
</tr>
<tr>
<td>Strope, Cory L.</td>
<td>69, 75</td>
</tr>
<tr>
<td>Strope, Pooja K</td>
<td>67</td>
</tr>
<tr>
<td>Stuckey, David C.</td>
<td>85</td>
</tr>
<tr>
<td>Sulman, Erik P</td>
<td>13</td>
</tr>
<tr>
<td>Talambedu, Usha</td>
<td>80</td>
</tr>
<tr>
<td>Taylor, Steve L</td>
<td>116</td>
</tr>
<tr>
<td>TerMaat, Joel</td>
<td>71</td>
</tr>
<tr>
<td>Teyssir, Jimmitti</td>
<td>82</td>
</tr>
<tr>
<td>Thapa, Kishore Kumar</td>
<td>115</td>
</tr>
<tr>
<td>Trapp, Joseph</td>
<td>82</td>
</tr>
<tr>
<td>Tsinoremas, Nicholas F</td>
<td>109</td>
</tr>
<tr>
<td>Varney, Michelle L</td>
<td>93</td>
</tr>
<tr>
<td>Vendettuoli, Marie</td>
<td>89</td>
</tr>
<tr>
<td>Venugopal, Poornima</td>
<td>91</td>
</tr>
<tr>
<td>Viljoen, Hendrik</td>
<td>71</td>
</tr>
<tr>
<td>Voshall, Adam</td>
<td>79</td>
</tr>
<tr>
<td>Wachholtz, Michael</td>
<td>107</td>
</tr>
<tr>
<td>Wang, Dong</td>
<td>63</td>
</tr>
<tr>
<td>Weeks, Don</td>
<td>97</td>
</tr>
<tr>
<td>Weeks, Donald P</td>
<td>87, 100</td>
</tr>
<tr>
<td>Whitney, Scott E</td>
<td>71</td>
</tr>
<tr>
<td>Wise, John</td>
<td>116</td>
</tr>
<tr>
<td>Wu, Sherry Y</td>
<td>92</td>
</tr>
<tr>
<td>Wu, Wei</td>
<td>110</td>
</tr>
<tr>
<td>Xie, Yunxuan</td>
<td>84</td>
</tr>
<tr>
<td>Xu, Isabel</td>
<td>82</td>
</tr>
<tr>
<td>Xue, Yuyan Shelly</td>
<td>114</td>
</tr>
<tr>
<td>Yao, Bo</td>
<td>101</td>
</tr>
<tr>
<td>Ye, Chun</td>
<td>104</td>
</tr>
<tr>
<td>Yeh, Cheng-Ting</td>
<td>110</td>
</tr>
<tr>
<td>Ying, Kai</td>
<td>110</td>
</tr>
<tr>
<td>Zempleni, Janos</td>
<td>96</td>
</tr>
<tr>
<td>Zhang, Chi</td>
<td>101</td>
</tr>
<tr>
<td>Zhao, Jin-Liang</td>
<td>107</td>
</tr>
<tr>
<td>Zheng, Ximeng</td>
<td>21</td>
</tr>
</tbody>
</table>
BIOT-2009
Sponsors

LI-COR

National Science Foundation

University of Nebraska-Lincoln Center for Biotechnology

University of Nebraska-Lincoln Center for Plant Science Innovation

University of Nebraska-Lincoln Department of Computer Science and Engineering

University of Nebraska-Lincoln Institute of Agriculture and Natural Resources

University of Nebraska-Lincoln Office of Research

University of Nebraska-Lincoln School of Biological Sciences

Symposium Co-Chairs
University of Nebraska, Lincoln, NE

Etsuko Moriyama
emoriyama2@unl.edu
School of Biological Sciences / Center for Plant Science Innovation

Stephen Scott
sscott@cse.unl.edu
Department of Computer Science & Engineering

Leen-Kiat Soh
lksoh@cse.unl.edu
Department of Computer Science & Engineering