Rocky 2011 is an official conference of the International Society for Computational Biology.

Rocky 2011 is supported by the Computational Bioscience Program at the University of Colorado School of Medicine.

Conference Chair
Lawrence Hunter, PhD
University of Colorado Denver
School of Medicine
Dear Rocky 2011 Participant

Welcome to the ninth Rocky Mountain Regional Bioinformatics Meeting.

The organizers hope that you enjoy the program, and find the meeting a productive opportunity to meet researchers, students and industrial users of bioinformatics technology in our region.

We are grateful for your continued interest in the meeting. Despite the hard economic times, we are on track for a record-setting attendance of more than 140 scientists presenting 8 keynote talks, 49 flash presentations and 90 scientific posters.

We are also grateful for the support of our sponsors. The ongoing sponsorship from IBM has made it possible to continue this conference for the past nine years. We also welcome back SomaLogic and Biodesix along with our newest sponsor NEC. We hope to have them support this conference for many years to come. It is only with the help of these sponsors that we can make this meeting as affordable as it is. Please seek out attendees from the sponsoring organizations, and let them know that their participation is important to you!

Finally, the meeting would simply not be possible without organizational help from Stephanie Hagstrom, Suzi Smith from the ISCB team, and Kathy Thomas.

We hope you enjoy the science, the company, and the spectacular scenery of the Rocky Mountains.

Welcome!

Larry Hunter
### THURSDAY – DECEMBER 8, 2011

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00 AM – 1:00 PM</td>
<td>Registration</td>
</tr>
</tbody>
</table>
| 1:00 PM – 1:45 PM | **KEYNOTE 1** • The Future of Research Communication: From Surfing to Deep Diving  
|                  | *JUDITH A. BLAKE, PhD, Associate Professor, The Jackson Laboratory, Bar Harbor, Maine, USA* |
| 1:45 PM – 2:45 PM | **ORAL PRESENTATIONS** 1–6                                             |
| 2:45 PM – 3:00 PM | Break – 15 minutes                                                    |
| 3:00 PM – 4:10 PM | **ORAL PRESENTATIONS** 7–13                                           |
| 4:10 PM – 4:25 PM | Break – 15 minutes                                                    |
| 4:25 PM – 5:10 PM | **KEYNOTE 2** • Building Cell Maps: Status and Challenges  
|                  | *EMEK DEMIR, PhD, Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, New York, USA* |
| 5:10 PM – 6:20 PM | **ORAL PRESENTATIONS** 14–20                                          |
| 7:00 PM – 9:00 PM | Banquet *Il Poggio Restaurant, Snowmass Village*                       |

### FRIDAY – DECEMBER 9, 2011

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 9:00 AM – 9:45 AM | **KEYNOTE 3** • Project Halo: Constructing and Exploiting a Formal Representation of a Biology Textbook to Understand and Answer Users’ Questions  
|                  | *PETER CLARK, PhD, Vulcan Inc., Seattle, Washington, USA*            |
| 9:45 AM – 10:45 AM | **ORAL PRESENTATIONS** 21–26                                         |
| 10:45 AM – 11:00 AM | Break – 15 minutes                                                  |
| 11:00 AM – 12:00 PM | **ORAL PRESENTATIONS** 27–32                                        |
| 12:00 PM – 12:30 PM | **KEYNOTE 4** • Advances in Protein Biomarker Discovery: Controlling for Sample Handling Artifacts and Confounding Effects  
|                  | *Mike Mehan, PhD, SomaLogic, Boulder, Colorado, USA*                |
| 12:30 PM – 4:00 PM | Break                                                               |

**continued on next page…**

**Session Locations:** *All oral presentations will be held at the Silvertree Hotel lower level*

**Poster Session Location:** *Snowmass Conference Center across the street from Silvertree Hotel*
**AGENDA AT-A-GLANCE**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00 PM – 4:45 PM</td>
<td>KEYNOTE 5 • Deciphering of Human Protein Interactome using Structural Complexes</td>
</tr>
<tr>
<td></td>
<td><strong>ANNA PANCHENKO, PhD, Associate Investigator, National Center for Biotechnology Information, NLM, NIH, Maryland, USA</strong></td>
</tr>
<tr>
<td>4:45 PM – 5:45 PM</td>
<td>ORAL PRESENTATIONS 33–38</td>
</tr>
</tbody>
</table>
| 5:45 PM – 8:00 PM  | RECEPTION & POSTER SESSION  
Sinclair Room, Snowmass Conference Center                                                                  |

**SATURDAY – DECEMBER 10, 2011**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 AM – 9:45 AM</td>
<td>KEYNOTE 6 • A New Day, A New High Performance Computer, What Does It Mean for the Life Sciences?</td>
</tr>
<tr>
<td></td>
<td><strong>KIRK E. JORDAN, PhD, Emerging Solutions Executive &amp; Associate Program Director, Computational Science Center IBM TJ. Watson Research, IBM Academy of Technology, Massachusetts, USA</strong></td>
</tr>
<tr>
<td>9:45 AM – 10:35 AM</td>
<td>ORAL PRESENTATIONS 39–43</td>
</tr>
<tr>
<td>10:35 AM – 11:05 AM</td>
<td>KEYNOTE 7 • SOMAMER-BASED PROTEOMIC ANALYSIS WITH MACHINE LEARNING</td>
</tr>
<tr>
<td></td>
<td><strong>HANS PETER GRAF, PhD, NEC Laboratories America, Princeton, New Jersey, USA</strong></td>
</tr>
</tbody>
</table>
| 11:05 AM – 12:30 PM| POSTER SESSION  
Sinclair Room, Snowmass Conference Center                                                                  |
| 12:30 PM – 4:00 PM | Break                                                                                         |
| 4:00 PM – 5:00 PM  | ORAL PRESENTATIONS 44–49                                                                        |
| 5:00 PM – 5:45 PM  | KEYNOTE 8 • Transfer Function Analysis of Signal Transduction — The PSF System                |
|                    | **GABRIELE SCHELER, PhD, Mountain View, California, USA**                                       |
| 5:45 PM            | ROCKY 2011 CLOSING COMMENTS                                                                     |

**Session Locations:** All oral presentations will be held at the Silvertree Hotel lower level  
* Poster Session Location: Snowmass Conference Center across the street from Silvertree Hotel
## AGENDA

**THURSDAY - DECEMBER 08, 2011**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00 AM – 1:00 PM</td>
<td>Registration</td>
</tr>
<tr>
<td>1:00 PM – 1:45 PM</td>
<td><strong>KEYNOTE 1 • The Future of Research Communication: From Surfing to Deep Diving</strong>&lt;br&gt;<strong>JUDITH A. BLAKE, PhD, Associate Professor, The Jackson Laboratory, Bar Harbor, Maine, USA</strong></td>
</tr>
<tr>
<td>1:45 PM – 1:55 PM</td>
<td><strong>ORAL PRESENTATION 1 • The Rat Genome Curation: RGD Automated Data Integration Pipelines Maximize Coverage</strong>&lt;br&gt;Presenter: MAREK TUTAJ, Medical College of Wisconsin&lt;br&gt;Authors: Marek Tutaj, Mary Shimoyama, Elizabeth Worthey, Jennifer Smith, Howard Jacob</td>
</tr>
<tr>
<td>1:55 PM – 2:05 PM</td>
<td><strong>ORAL PRESENTATION 2 • A Sensemaking Model for the Explorative Analysis of Large Gene Lists</strong>&lt;br&gt;Presenter: CARSTEN GÖRG, University of Colorado Denver&lt;br&gt;Authors: Carsten Görg, Barbara Mirel, Hannah Tipney</td>
</tr>
<tr>
<td>2:05 PM – 2:15 PM</td>
<td><strong>ORAL PRESENTATION 3 • The Role of p53 in Oligodendrocyte UPR</strong>&lt;br&gt;Presenter: HASAN JAMIL, Wayne State University&lt;br&gt;Authors: Hasan Jamil</td>
</tr>
<tr>
<td>2:15 PM – 2:25 PM</td>
<td><strong>ORAL PRESENTATION 4 • Quantitative Data: Where are they Hidden in Biomedical Literature?</strong>&lt;br&gt;Presenter: KOMANDUR RAVIKUMAR, University of Colorado School of Medicine&lt;br&gt;Authors: K.E. Ravikumar, Meenakshi Narayanaswamy, S.V. Ramanan</td>
</tr>
<tr>
<td>2:25 PM – 2:35 PM</td>
<td><strong>ORAL PRESENTATION 5 • Multiscale Patient-Specific Blood Systems Biology</strong>&lt;br&gt;Presenter: SCOTT DIAMOND, University of Pennsylvania</td>
</tr>
<tr>
<td>2:35 PM – 2:45 PM</td>
<td><strong>ORAL PRESENTATION 6 • A Resource for the Rational Selection of Drug Target Proteins and Leads for the Malaria Parasite, Plasmodium Falciparum</strong>&lt;br&gt;Presenter: FOURIE JOUBERT, University of Pretoria&lt;br&gt;Authors: Jeanre Smit, Phele Mpangase, Michal Szolkiewicz, Misha le Grange, Fourie Joubert</td>
</tr>
<tr>
<td>2:45 PM – 3:00 PM</td>
<td>Break (15 minutes)</td>
</tr>
</tbody>
</table>
### ORAL PRESENTATION 7 • Identifying Treatment Relevant Colorectal Cancer Subtypes Using Iterative Non-negative Matrix Factorization

**Presenter:** ANDREAS SCHLICKER, Netherlands Cancer Institute  
**Authors:** Andreas Schlicker, Garry Beran, Christine M Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, Sara Davenport, Kerry Heathcote, Denis Alvarez Castro, George Orphanides, Tim French, Lodewyk FA Wessels

### ORAL PRESENTATION 8 • The Origin of Mammalian Placentation Correlates with Protein Functional Shift and the Emergence of New Control Mechanisms

**Presenter:** MARY O’CONNELL, Dublin City University  
**Authors:** Thomas A. Walsh, Kieran Holohan, Anna O’Brien, Robert Carton, Elinor Velasquez, Claire C. Morgan, Noeleen B. Loughran, Mary J. O’Connell

### ORAL PRESENTATION 9 • Exploring and Profiling Long Non-coding RNA in T Cell Development Using Next Generation Sequencing and Bioinformatics Approaches

**Presenter:** TZU L. PHANG, University of Colorado Denver  
**Authors:** Tzu L. Phang, Ping Yao Zeng, Edwin F. de Zoeten

### ORAL PRESENTATION 10 • Building an Interactome to Identify Signaling Components

**Presenter:** SARAH WYATT, Ohio University  
**Authors:** Sarah Waytt, Kaiyu Shen

### ORAL PRESENTATION 11 • geneSmash: a RESTful Web Service for Gene Annotations

**Presenter:** GANIRAJU MANYAM, University of Texas Anderson Cancer Center  
**Authors:** Michelle Payton, Ganiraju Manyam, Chris Wakefield, Jack Roth, Lynne Abruzzo, Kevin Coombes

### ORAL PRESENTATION 12 • Integrating Curated Databases and Text Mining Output into a Biomedical Knowledge Base

**Presenter:** KEVIN LIVINGSTON, University of Colorado Denver  
**Authors:** Kevin M. Livingston, Michael Boda, William A. Baumgartner Jr, Yuriy Malenkiy, Lawrence E. Hunter
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 4:00 PM – 4:10 PM | ORAL PRESENTATION 13 • Using Evolving Protein Networks to Inform the Graph Coloring Problem  
Presenter: TODD GIBSON, University of Colorado Denver  
Authors: Todd A. Gibson, Debra S. Goldberg |
| 4:10 PM – 4:25 PM | Break (15 minutes)                                                      |
| 4:25 PM – 5:10 PM | KEYNOTE 2 • Building Cell Maps: Status and Challenges  
EMEK DEMIR, PhD, Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, New York, USA |
| 5:10 PM – 5:20 PM | ORAL PRESENTATION 14 • Machine Approaches to Recognition of Trypanosomal Variant Surface Glycoprotein Sequences  
Presenter: JON WILKES, Wellcome Trust Centre for Molecular Parasitology  
Authors: Jon Wilkes |
| 5:20 PM – 5:30 PM | ORAL PRESENTATION 15 • Compact Encoding for Gene Therapy  
Presenter: ROGER HALL, University of Arkansas at Little Rock  
Authors: Roger Hall |
| 5:30 PM – 5:40 PM | ORAL PRESENTATION 16 • Microbial Monitoring and Space Exploration  
Presenter: SVEN BILKE, NCI  
Authors: Sven Bilke, Verena Starke |
| 5:40 PM – 5:50 PM | ORAL PRESENTATION 17 • CodingMotif: Exact Determination of Overrepresented Nucleotide Motifs in Coding Sequences  
Presenter: JEFFREY CHUANG, Boston College  
Authors: Yang Ding, Andy Lorenz, Jeffrey Chuang |
| 5:50 PM – 6:00 PM | ORAL PRESENTATION 18 • GO Classes or Biclusters? Alternative Approaches for Exploring Microarray Probeset Subsets  
Presenter: GEORGE ACQUAHA-MENSAH, Massachusetts College of Pharmacy and Health Sciences  
Authors: George Acquaah-Mensah |
| 6:00 PM – 6:10 PM | ORAL PRESENTATION 19 • Accurate Inferring Transcription Regulation from a Compendium of Expression Profiles  
Presenter: XUELING LI, University of Texas Medical Branch  
Authors: Xueling Li, Dirar Homouz, Andrzei Kudlicki |
| 6:10 PM – 6:20 PM | ORAL PRESENTATION 20 • A Proposed Algorithm for Epistasis Detection using Frequent Item-sets and Mutual Information  
Presenter: JAMES RUDD, North Carolina University  
Authors: James Rudd, Assefa Tesfay, ClarLynda Williams-DeVane, Gaolin Zheng |
<p>| 7:00 PM – 9:00 PM | BANQUET • Il Poggio Restaurant, Snowmass Village                        |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Presenter/Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 AM – 9:45 AM</td>
<td><strong>KEYNOTE 3 • Project Halo: Constructing and Exploiting a Formal Representation of a Biology Textbook to Understand and Answer Users’ Questions</strong>&lt;br&gt;<strong>PETER CLARK, PhD, Vulcan Inc., Seattle, Washington, USA</strong></td>
<td></td>
</tr>
<tr>
<td>9:45 AM – 9:55 AM</td>
<td><strong>ORAL PRESENTATION 21 • Molecular Insights to the Drug Resistance of V32I &amp; M46L HIV-1 Protease Mutant to Inhibitor TMC114: Free Energy Calculation and Molecular Dynamics Simulations</strong>&lt;br&gt;<strong>Presenter: BISWA MEHER, Albany State University</strong>&lt;br&gt;<strong>Authors: Biswa Ranjan Meher, Yixuan Wang</strong></td>
<td></td>
</tr>
<tr>
<td>9:55 AM – 10:05 AM</td>
<td><strong>ORAL PRESENTATION 22 • Testing the Ortholog Conjecture with Functional Data from Several Pairs of Closely Related Organisms</strong>&lt;br&gt;<strong>Presenter: WYATT CLARK, Indiana University</strong>&lt;br&gt;<strong>Authors: Wyatt T. Clark, Predrag Radivojac, Matthew Hahn</strong></td>
<td></td>
</tr>
<tr>
<td>10:05 AM – 10:15 AM</td>
<td><strong>ORAL PRESENTATION 23 • Functional Profiling of Pharmacogenetic Non-synonymous SNPs</strong>&lt;br&gt;<strong>Presenter: CHET SELIGMAN, Buck Institute for Research on Aging</strong>&lt;br&gt;<strong>Authors: Chet Seligman, Janita Thusberg, Jackson Miller, Emidio Capriotti, Jim Auer, Michelle Whirl-Carrillo, Teri Klein, Sean Mooney</strong></td>
<td></td>
</tr>
<tr>
<td>10:15 AM – 10:25 AM</td>
<td><strong>ORAL PRESENTATION 24 • Limitations of Automated Annotation Software When Used with Draft Genomic Assemblies</strong>&lt;br&gt;<strong>Presenter and Author: ROBERT NORGREN, Jr. University of Nebraska Medical Center</strong></td>
<td></td>
</tr>
<tr>
<td>10:25 AM – 10:35 AM</td>
<td><strong>ORAL PRESENTATION 25 • Mixed Primary-Secondary Structure Alignment of ncRNA Covariance Models for Model Clustering and Combination</strong>&lt;br&gt;<strong>Presenter and Author: JENNIFER SMITH, Boise State University</strong></td>
<td></td>
</tr>
<tr>
<td>10:35 AM – 10:45 AM</td>
<td><strong>ORAL PRESENTATION 26 • In Silico Gene Expression Based Analysis on Claudin Family Members Association with Human Thyroid Cancer</strong>&lt;br&gt;<strong>Presenter: SHAUKAT MALIK, Mohammad Ali Jinnah University</strong>&lt;br&gt;<strong>Authors: Shaukat Malik, S. Sameen, Z. Khalid</strong></td>
<td></td>
</tr>
<tr>
<td>10:45 AM – 11:00 AM</td>
<td><strong>Break (15 minutes)</strong></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Presentation Title</td>
<td>Presenter(s)</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>11:00 AM – 11:10 AM</td>
<td>ORAL PRESENTATION 27 • Data Management and Analysis Solutions for Meta-Analysis of Multi-Domain Data</td>
<td>CLARLYNDA WILLIAMS-DEVANE, North Carolina Central University</td>
</tr>
<tr>
<td>11:20 AM – 11:30 AM</td>
<td>ORAL PRESENTATION 29 • DAVID-WS: A Stateful Web Service to Facilitate Gene/Protein List Analysis</td>
<td>XIAOLI JIAO, NIH/NIAID</td>
</tr>
<tr>
<td>11:30 AM – 11:40 AM</td>
<td>ORAL PRESENTATION 30 • Predicting Transcription Factor Binding Sites with Hidden Markov Models by using ChIP-Seq Data</td>
<td>ANTHONY MATHELIER, University of British Columbia</td>
</tr>
<tr>
<td>11:40 AM – 11:50 AM</td>
<td>ORAL PRESENTATION 31 • The Critical Assessment of Function Annotation Experiment: A Community-wide Effort Towards a Better Functional Annotation of Genes and Genomes</td>
<td>PREDRAG RADIVOJAC, Indiana University</td>
</tr>
<tr>
<td>11:50 AM – 12:00 PM</td>
<td>ORAL PRESENTATION 32 • In Silico Rational Drug Design and Modeling Studies of Novel Inhibitors for Multi-target Inhibition in Pseudomonas Aeruginosa</td>
<td>JAYARAMAN PREMKUMAR, Nanyang Technological University</td>
</tr>
<tr>
<td>12:00 PM – 12:30 PM</td>
<td>KEYNOTE 4 • Advances in Protein Biomarker Discovery: Controlling for Sample Handling Artifacts and Confounding Effects</td>
<td>MIKE MEHAN, PhD, SomaLogic, Boulder, Colorado, USA</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------------------------------------------------------------</td>
<td></td>
</tr>
</tbody>
</table>
| 4:00 PM – 4:45 PM | **KEYNOTE 5 • Deciphering of Human Protein Interactome using Structural Complexes**  
**ANNA PANCHENKO, PhD, Associate Investigator, National Center for Biotechnology Information, NLM, NIH, Bethesda, Maryland, USA** |
| 4:45 PM – 4:55 PM | **ORAL PRESENTATION 33 • PathCMap: Development of Pathway Signature System for Identifying Druggable Partners of Synthetic Lethal Genes in Cancer**  
**Presenter: JIHYE KIM, University of Colorado Denver School of Medicine**  
**Authors: Jihye Kim, Carlos H C Cano, Aik Choon Tan** |
| 4:55 PM – 5:05 PM | **ORAL PRESENTATION 34 • Identifying Single Copy Orthologs in Metazoa**  
**Presenter: CHRISTOPHER CREEVEY, Teagasc**  
**Authors: Chris Creevey, Jean Muller, Tobias Doerks, Julie D. Thompson, Detlev Arendt, Peer Bork** |
| 5:05 PM – 5:15 PM | **ORAL PRESENTATION 35 • Mining Genomes to Understand the Origin of Vision in Metazoa**  
**Presenter: DAVIDE PISANI, National University of Ireland, Maynooth**  
**Authors: Roberto Feuda, Davide Pisani** |
| 5:15 PM – 5:25 PM | **ORAL PRESENTATION 36 • Evolutionary Pattern Embedded in the Lengths of Proteins and Their Structural Units**  
**Presenter: MINGLEI WANG, University of Illinois at Urbana-Champaign**  
**Authors: Minglei Wang, Cedric Debes, Frauke Gräter, Gustavo Caetano-Anollés** |
| 5:25 PM – 5:35 PM | **ORAL PRESENTATION 37 • A Multiple-template Approach for Protein Threading**  
**Presenter: JIAN PENG, Toyota Technological Institute at Chicago**  
**Authors: Jian Peng, Jinbo Xu** |
| 5:35 PM – 5:45 PM | **ORAL PRESENTATION 38 • A New Probabilistic Model in Predictive Microbiology (NPMPM)**  
**Presenter: NADINE SCHOENE, Goethe University Frankfurt**  
**Authors: Nadine Schoene, Alexander Bockmayr, Bernd Appel, Annmarie Kaesbohrer** |
| 5:45 PM – 8:00 PM | **RECEPTION AND POSTER SESSION**  
Snowmass Conference Center (across street from Silvertree), Sinclair Room |
AGENDA

SATURDAY – DECEMBER 10, 2011

9:00 AM – 9:45 AM  KEYNOTE 6 • A New Day, A New High Performance Computer, What Does It Mean for the Life Sciences?
KIRK E. JORDAN, PhD, Emerging Solutions Executive & Associate Program Director, Computational Science Center, IBM T.J. Watson Research, Member, IBM Academy of Technology, Massachusetts, USA

9:45 AM – 9:55 AM  ORAL PRESENTATION 39 • Rediscovery of the p53 Transcriptome
Presenter: MARY ALLEN, University of Colorado
Authors: Mary Allen, Robin Dowell, Joaquin Espinosa

9:55 AM – 10:05 AM  ORAL PRESENTATION 40 • Developing Maximum Likelihood and Bayesian Supertrees
Presenter: WASIU AKANNI, National University of Ireland, Maynooth

10:05 AM – 10:15 AM  ORAL PRESENTATION 41 • Prediction of Operons in Microbial Genomes by Integrating Diverse Information Sources
Presenter: ANIS KARIMPOUR-FARD, University of Colorado Anschutz Medical Campus
Authors: Anis Karimpour-Fard, Lawrence E. Hunter

10:15 AM – 10:25 AM  ORAL PRESENTATION 42 • Gene Language Model
Presenter: SIAMAK REZAEI, Talai
Authors: Siamak Rezaei

10:25 AM – 10:35 AM  ORAL PRESENTATION 43 • Live and Let Die
Presenter: CHRISTIAN FORST, University of Texas Southwestern Medical Center
Authors: Nassim Sohaee, Christian V. Forst

10:35 AM – 11:05 AM  KEYNOTE 7 • SOMAMER-BASED PROTEOMIC ANALYSIS WITH MACHINE LEARNING
HANS PETER GRAF, PhD, NEC Laboratories America, Princeton, New Jersey, USA

11:05 AM – 12:30 PM  POSTER SESSION • Sinclair Room, Snowmass Conference Center (across street from Silvertree)

12:30 PM – 4:00 PM  Break
<table>
<thead>
<tr>
<th>Time</th>
<th>Presentation Title</th>
<th>Presenter and Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00 PM</td>
<td>ORAL PRESENTATION 44 • Broad Semantic Class Assignment for</td>
<td>KEVIN BRETONNEL COHEN, University of Colorado School of</td>
</tr>
<tr>
<td></td>
<td>Biomedical Text</td>
<td>Medicine</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Authors: K. Bretonnel Cohen, Thomas Christiansen, Lawrence E.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hunter</td>
</tr>
<tr>
<td>4:10 PM</td>
<td>ORAL PRESENTATION 45 • On the Accuracy of Protein Tertiary</td>
<td>ALEKSANDAR POLEKSIC, University of Northern Iowa</td>
</tr>
<tr>
<td></td>
<td>Structure Comparison</td>
<td>Authors: Aleksandar Polezik, Mauricio Arriagada</td>
</tr>
<tr>
<td>4:20 PM</td>
<td>ORAL PRESENTATION 46 • A Distributed Framework for</td>
<td>CHRISTOPHE ROEDER, University of Colorado, Anschutz Medical</td>
</tr>
<tr>
<td></td>
<td>Computation on the Results of Large Scale NLP</td>
<td>Campus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Authors: Christophe Roeder, William Baumgartner Jr, Kevin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Livingston</td>
</tr>
<tr>
<td>4:30 PM</td>
<td>ORAL PRESENTATION 47 • Probabilistic Search Frameworks for</td>
<td>AMARDA SHEHU, George Mason University</td>
</tr>
<tr>
<td></td>
<td>Protein Conformational Spaces</td>
<td></td>
</tr>
<tr>
<td>4:40 PM</td>
<td>ORAL PRESENTATION 48 • Predicting HIV-1, Human Protein</td>
<td>OZNUR TASTAN, Microsoft Research New England</td>
</tr>
<tr>
<td></td>
<td>Interactome Through Indirect and Direct Evidences</td>
<td>Authors: Oznur Tastan, Jaime Carbonell, Judith Klein</td>
</tr>
<tr>
<td>4:50 PM</td>
<td>ORAL PRESENTATION 49 • Using Information Theory to Map Reads</td>
<td>JOHN CONERY, University of Oregon</td>
</tr>
<tr>
<td></td>
<td>to a Reference Sequence</td>
<td></td>
</tr>
<tr>
<td>5:00 PM</td>
<td>KEYNOTE 8 • Transfer Function Analysis of Signal Transduction—</td>
<td>GABRIELE SCHELER, PhD, Mountain View, California, USA</td>
</tr>
<tr>
<td></td>
<td>The PSF System</td>
<td></td>
</tr>
<tr>
<td>5:45 PM</td>
<td>ROCKY 2011 CLOSING COMMENTS</td>
<td></td>
</tr>
</tbody>
</table>
KEYNOTE SPEAKERS

JUDITH A. BLAKE, PhD
Associate Professor, The Jackson Laboratory, Bar Harbor, Maine, USA
http://research.jax.org/faculty/judith_blake.html

THE FUTURE OF RESEARCH COMMUNICATION: FROM SURFING TO DEEP DIVING

ABSTRACT: The Internet has revolutionized the transmission of data and of knowledge. Scientific journal publishers are reeling under the impact of on-line publishing, the inability to handle very large datasets for peer review, and the emergence of download tracking and commentary post-publication as impact metrics. Electronic management of research data and results provide a mechanism for open access to electronic project directories. Very large data sets are increasingly common, and testing reproducibility of reported results is increasingly difficult to undertake. Nonetheless, ontologies, semantic and accession ID mapping, and author-tagging provide immediate points for on-line integration of scientific results, both experimental and inferred, although the distinction is often missing. For scientists associated with large institutions, access to copyrighted biomedical literature continues unabated although the function of the university library is radically changing. Meanwhile, scientists and citizens without access to journal subscriptions have limited access to publically financed scientific results although plenty of scientific discourse is available on the Web. Open access publication, digital data repositories, and electronic journals will help in the dissemination of scientific research results. I will discuss these topics and the impact of these changes on the scientific enterprise.

PETER CLARK, PhD
Vulcan Inc., Seattle, Washington, USA
CV: www.cs.utexas.edu/users/pclark

PROJECT HALO: CONSTRUCTING AND EXPLOITING A FORMAL REPRESENTATION OF A BIOLOGY TEXTBOOK TO UNDERSTAND AND ANSWER USERS’ QUESTIONS

ABSTRACT: As part of Project Halo at Vulcan Inc, we are building a large-scale, broad-coverage, formal (logic-based) knowledge-base (KB) that represents a substantial portion of the knowledge in an AP-level biology textbook, and supports question interpretation, reasoning, and question answering. Because users pose questions in English, we are not spared the huge challenge of interpreting natural language; however, the KB does provide significant advantages for this task, in particular creating knowledge-based expectations of what would be coherent to ask, that can be used to coerce a user’s question into something meaningful to the computer.

In this talk I will describe the project, our approach and progress in constructing the KB, and our successes and challenges in interpreting and answering users’
biology questions with it. I will then speculate on the longer-term picture of using the knowledge base to guide interpretation of (parts of) biology texts themselves, with the potential to further expand the KB semi-automatically and ultimately create more knowledgeable machines.

EMEK DEMIR, PhD  
Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, New York, USA  
http://cbio.mskcc.org/~demir/cv.html

BUILDING CELL MAPS: STATUS AND CHALLENGES

ABSTRACT: Advances in molecular technologies have led to rapid generation of data and information about cellular processes at an increasing rate. Current means of knowledge representation and scientific communication in biology cannot adequately deal with the complexity and volume of this information — a serious bottleneck for developing a causal, predictive understanding of the cell.

To address this problem we have developed BioPAX (Biological Pathway Exchange), a standard language for representing and exchanging pathway information. Recently released BioPAX level 3 can represent signaling and metabolic pathways, gene regulation and molecular and genetic interactions in great detail.

The latest version of Pathway Commons, our pathway data integration and aggregation server aims to provide a “merged” network of publicly available databases that support BioPAX. Through an iterative process of aggregation, alignment, matching, and merging - not very different conceptually than putting pieces of a puzzle together — we are building a cell map.

Pathway alignment, or finding similar and/or equivalent portions of two pathways, is a crucial step for this goal. PATCH is an algorithm that can align pathways even when there is missing or omitted knowledge. We have successfully applied Patch to find similar pathways between Reactome and NCI/PID, and detected several curation errors in both databases during this process.

As we build pathway resources and infrastructure it is becoming increasing possible to utilize a knowledge driven approach to biological problems. With a mechanistic understanding of the cellular events, we can better subtype diseases, predict drug responses, and choose combinations of drugs to optimally interfere with the disease. This would be revolutionary, especially for complex, multi-causal diseases such as cancer.
ABSTRACT: SOMAmer technology opens the possibility for a low-cost analysis of thousands of plasma proteins from a small sample of blood. Finding reliable biomarkers among such a large number of proteins is challenging since the plasma proteome has a complex composition and conditions can change quickly, affecting the concentrations of proteins we try to interpret.

Finding reliable biomarkers is a problem of feature selection, where we search for combinations of proteins that consistently indicate certain conditions, such as a disease, with high sensitivity and specificity. Typical approaches are univariate algorithms or algorithms that greedily search for groups of proteins. Recently, research has made considerable progress in feature learning where we try learning automatically features that are more indicative of the underlying phenomena. We applied methods based on deep learning and kernel techniques successfully in various image, text and protein analysis applications.

Here we demonstrate several multivariate feature selection algorithms based on L1 regularization and kernel techniques for the detection of biomarkers in data from three different cancer studies. Each data set consists of several hundred samples with between 800 and 1,000 different proteins. Due to the small sample sizes, individual runs of feature learning can be unstable, with variations in the selected proteins from one run to the next. Yet with statistical sampling, groups of proteins are identified reproducibly that provide high sensitivity and selectivity. These results are very encouraging since the ability to analyze large numbers of proteins holds great promise for a wide range of proteomic analysis applications.
of our experience on this new system including bring up Rosetta, a code that predicts protein structures from amino acid sequences in DNA. In addition, while related to our HPC work, I will describe work continuing to make HPC accessible to a wider audience and eventually targeting the healthcare and life science practitioner directly and explain why this work is of importance.

MICHAEL R. MEHAN, PhD
SomaLogic Inc., Boulder, Colorado, USA

ADVANCES IN PROTEIN BIOMARKER DISCOVERY: CONTROLLING FOR SAMPLE HANDLING ARTIFACTS AND CONFOUNDING EFFECTS
Collaborators: Rachel Ostroff, Alex Stewart, Glenn Sanders, Dom Zichi, Ed Brody, Steve Williams

ABSTRACT: Many biomarker discovery studies may fail to validate because the clinical population does not represent the intended clinical use or because hidden preanalytic variability in the discovery samples contaminates the apparent disease specific information in the biomarkers. This preanalytic variability can arise from differences in blood sample processing between study sites, or worse, introduce case/control bias in samples collected differently at the same study site. To better understand the effect of different blood sample processing procedures, we evaluated protein measurement bias in a large multi-center lung cancer study. These analyses revealed that perturbations in serum protocols result in changes to many proteins in a coordinated fashion.

Using the SomaLogic SOMAscan platform we developed protein biomarker signatures of processes such as cell lysis, platelet activation, and complement activation and assembled these preanalytic signatures into quantitative multi-dimensional Sample Mapping Vectors (SMV). The underlying platform technology uses SOMAmers (Slow Off-rate Modified Aptamers) as affinity reagents to quantify approximately 850 proteins. The SMV score provides critical evaluation of both the quality of every blood sample used in discovery, and also enables the evaluation of candidate protein biomarkers for resistance to preanalytic variability.
DECIPHERING OF HUMAN PROTEIN INTERACTOME USING STRUCTURAL COMPLEXES

ABSTRACT: Proteins function by interacting with other biomolecules and knowledge of the entire set of interactions combined with the properties of protein binding sites is essential for our understanding of cellular functions and the origins of many diseases. Recently we developed a method (IBIS) which analyzes and predicts interaction partners and locations of binding sites in proteins based on the evolutionary conservation of binding sites in homologous structural complexes. IBIS imposes a number of rigorous criteria in order to increase the reliability of homology-based inference of interactions and provides binding site annotations for five different types of interaction partners (proteins, small molecules, nucleic acids, peptides and ions). It facilitates the mapping of the entire biomolecular interaction network for a given organism and we use this framework to map the human protein interactome and analyze its properties. We show that structurally inferred interaction network is highly modular and has small-world characteristics. Moreover it is more functionally coherent and reliable compared to high-throughput interaction networks. Since structurally inferred interaction network provides the details of binding interfaces, we analyze the effect of cancer associated point mutations on protein-protein binding. We show that cancer related mutations can either destabilize or make the complex more stable and lead to excessive activation or inactivation.

Cellular regulatory mechanisms provide a sensitive and specific response to external stimuli and such dynamic regulation can be achieved through reversible covalent modifications. We study the effect of phosphorylation on protein binding and function for different types of complexes from the human proteome. Our analysis of molecular mechanisms of phosphorylation shows that phosphorylation may modulate the binding affinity and trigger the transitions between different conformer and oligomeric states. We also show that phosphorylation sites are not only more likely to be evolutionary conserved than surface residues but even more so than the binding interface.
ABSTRACT: We present a new approach towards a modular and systematic analysis of biochemical reaction models using a modified steady-state assumption.

An ordinary differential equation (ODE) system for both complex formation and enzymatic reactions is automatically transformed into a set of signal-response transfer functions, called protein signaling functions (PSF). Elementary PSFs represent individual biochemical reactions out of context, systemic PSFs are the transformation of the elementary PSF in the context of a system of equations.

The use of systemic PSFs reduces the complexity of biological signal systems to manageable chunks which allow modular parameter adjustment.

The poster uses two published moderate-sized ODE models on striatal neural plasticity to present the analysis. The models use a different selection of proteins and interactions, but aim to model the same biological system. They are derived from essentially the same experimental data by standard methods of parameter tuning.

Re-analysis of the ODEs as PSF systems allows to directly compare shared components between the systems, such as the centrally important cAMP-PKA-DARPP32 connection. The results show that individual pathway components have become tuned to radically different quantitative transfer functions and concentration ranges, due to the influence of the different system embedding. By tuning transfer functions directly in the PSF system, current methods of system construction can be improved and cross-model consistency becomes achievable.
OP1: THE RAT GENOME CURATION: RGD AUTOMATED DATA INTEGRATION PIPELINES MAXIMIZE COVERAGE

Presenter: MAREK TUTAJ, Medical College of Wisconsin
Authors: Marek Tutaj, Mary Shimoyama, Elizabeth Worthey, Jennifer Smith, Howard Jacob

ABSTRACT: In model organism curation, relying solely on manual curation of the literature is impractical and would result in functional information for only a small portion of the genome. RGD uses a combination of targeted literature curation and a network of automated pipelines to provide comprehensive functional coverage of the rat genome. Quality control processes employed in the pipelines help identify conflicts, omissions and questionable relationships among data originating at other sources and data already in RGD. Conflict reports are sent to curators to resolve data issues. Through these pipelines, RGD 1) integrates and matches genes and gene models from multiple sources with links to sequence data, 2) identifies and loads orthologs and creates ortholog relationships, 3) identifies genes requiring nomenclature review and provides provisional nomenclature, 4) adds and updates genomic positions for rat, human and mouse genes, 5) updates multiple ontologies and identifies obsoleted terms and annotations, 6) provides experimentally determined human and mouse ortholog Gene Ontology annotations for rat genes, 7) provides identifiers and links for major protein information. RGD’s pipelines use a variety of data loading strategies, including drop-and-reload and incremental updates. Multiple quality control processes run in parallel on multicore machines allow dramatic reduction of total running time of the pipelines while providing the desired level of quality checking and high throughput of the data processed. This sophisticated data pipeline network allows RGD to provide comprehensive genome-wide functional and structural information with weekly updates.

OP2: A SENSEMAKING MODEL FOR THE EXPLORATIVE ANALYSIS OF LARGE GENE LISTS

Presenter: CARSTEN GÖRG, University of Colorado Denver
Authors: Carsten Görg, Barbara Mirel, Hannah Tipney

ABSTRACT: The cognitive science and visual analytics community have developed well-supported models of the stages analysts go through over the course of an investigation. These models have proven useful for understanding the analyst’s cognitive tasks and deriving design guidelines to develop user-centered tools that can facilitate and enhance these tasks. Specific models for the analysis of gene lists do not yet exist, particularly for the analysis of genotype relationships to generate hypotheses about mechanisms of phenotype-level diseases. To better understand and model these analytical processes we have conducted an in-depth case study of a biomedical researcher. The researcher’s
goal was to uncover genome-level interactions and events that might explain why some heart failure patients did not respond to beta-blocker treatment while others did. Over a six-month period we observed the researcher’s real-time interactions and think aloud protocols with visualization tools using screen and audio capture. The visualization tools included Cytoscape, Hanalyzer, String, Reactome, and tag clouds; the researcher also accessed the Genetic Association Database, the Pharmacogenomics Knowledge Base, and the biomedical literature. Additionally, we interviewed the researcher monthly and obtained a copy of the laboratory notebook and numerous annotated articles. We then performed a cognitive task analysis taking the interplay of the different tools into account and developed a preliminary sensemaking model for the explorative analysis of molecular interactions. Our preliminary model adapts and refines Pirolli and Card’s well-known sensemaking model for intelligence analysis. We present design guidelines for visual analytics tools that we derived from our model.

**OP3: THE ROLE OF P53 IN OLIGODENDROCYTE UPR**

*Presenter and Author: HASAN JAMIL, Wayne State University*

**ABSTRACT:** It has been hypothesized that mutant Plp1 genes in oligodendrocyte are responsible for protein misfolding and retention at the endoplasmic reticulum (ER) leading to demyelination via unfolded protein response (UPR). Neurodegeneration due to demyelination has been linked to diseases such as multiple sclerosis, leukodystrophies, Alzheimer’s and Parkinson’s. It is known that PERK plays a significant role in translation attenuation and cell cycle arrest during UPR activation. Recent studies in our laboratory on rsh and msd mutant mice suggest that the crosstalk between AKT and PERK pathways through p53 may be a key in deciphering the mechanism of oligodendrocyte cell deaths. This possibility increases because p53 is known to be involved in cell cycle progression check point, and the interaction between ATF4 and p53 via ATF3 in PERK pathway and in the AKT pathway. In this study, our goal is to develop a computational approach to study the role of p53 and investigate its contributions in oligodendrocyte UPR signalling cascade. We develop an integrated system to explore gene expression data sets to hypothesize about possible gene regulatory networks for p53 in rsh and msd mutant mice, and cross validate the networks with drosophila genome wide protein and transcription factor interactions toward filtering and sanitizing the hypothesized networks. We plan to determine the contributions of p53 by iterative refinement of the PERK and AKT pathways by way of introducing more missing links.
OP4: QUANTITATIVE DATA: WHERE ARE THEY HIDDEN IN BIOMEDICAL LITERATURE?

Presenter: KOMANDUR RAVIKUMAR, University of Colorado School of Medicine
Authors: K.E. Ravikumar, Meenakshi Narayanaswamy, S.V. Ramanan

ABSTRACT: Computational modeling of biological systems has been greatly impeded by the fact that extracting the values of various parameters from the literature has been primarily a manual task. Automated extraction of such quantitative data from the bio-medical literature, a largely underexplored problem would significantly aid the curation of such information in databases such as KEGG, CellML models repository, BioModels database, which are often a good starting point for system biologists involved in modeling biological pathways and systems. Here, we propose a method to automatically extract quantitative data from the literature, specifically in the context of electrophysiology. We used (i) dictionary lookup (e.g. conductance, membrane potential) and regular expressions (e.g., Kd, Ki) to tag electrophysiological parameters and (ii) regular expressions to detect their values (with units, e.g., 20 pS, -80 mV) as they occur in the text. Linguistic rules were developed to pair occurrences of parameters and compatible values within the clause; we used a development corpus of 150 abstracts. We also extended our approach beyond clausal boundaries by considering all possible relationship pairs that co-occur within a sentence, but retaining only compatible parameter-unit pairs. The precision, recall and F-measure for extracting quantitative relationship pairs were 83.2%, 73.20%, and 77.88% respectively, when evaluated on a small test corpus. While the intra-clausal linguistic patterns gave highly precise relations, the extra-clausal pairing mechanisms significantly improved the recall (by 10%) without any drop in precision.

OP5: MULTISCALE PATIENT-SPECIFIC BLOOD SYSTEMS BIOLOGY

Presenter and Author: SCOTT L. DIAMOND, University of Pennsylvania

ABSTRACT: Predicting tissue function based upon an individual’s unique cells requires a multiscale Systems Biology approach to understand the coupling of intracellular signaling with spatiotemporal gradients of extracellular biochemicals controlled by convective-diffusive transport. During thrombotic or hemostatic episodes, platelets bind collagen and release ADP and thromboxane A2 (TXA2) to facilitate the recruitment of additional platelets to a growing deposit that distorts the flow field. Calcium dye-loaded platelets in PPACK and indomethacin-treated plasma (thus lacking thrombin and TXA2) from 3 healthy donors were subjected to Pairwise Agonist Scanning where platelets were exposed to all pairwise combinations of ADP, U46619, and convulxin (at 0, 0.1, 1, 10 x EC50) to activate P2Y1/P2Y12, TP, and GPVI receptors, respectively, in the presence or absence of the IP receptor agonist, iloprost. With 74 calcium responses to train a neural network (NN) model of platelet calcium mobilization for each donor, each NN model was then embedded into a multiscale Monte Carlo/
finite element simulation of donor-specific platelet deposition under flow. For each donor, simulations predicted the measured platelet deposition dynamics and ranked drug sensitivity for PPACK-treated whole blood flowing over collagen at 200 s⁻¹ wall shear rate in the presence of indomethacin, aspirin, MRS-2179 (P2Y1 inhibitor), or iloprost. Consistent with measurement and simulation, one donor displayed larger clots, while another donor presented a indomethacin-resistance and U46619-insensitivity (revealing a novel heterozygote mutation). In silico representations of an individual’s platelet phenotype allowed prediction of blood function under flow, essential to identifying patient-specific cardiovascular risks, drug responses, and novel genotypes.

OP6: DISCOVERY: A RESOURCE FOR THE RATIONAL SELECTION OF DRUG TARGET PROTEINS AND LEADS FOR THE MALARIA PARASITE, PLASMODIUM FALCIPARUM

Presenter: FOURIE JOUBERT, University of Pretoria
Authors: Jeanre Smit, Phele Mpangase, Michal Szolkiewicz, Misha le Grange, Fourie Joubert

ABSTRACT: Few rational approaches have been successfully followed in the selection of promising drug target proteins in the malaria parasite. The emergence of widespread drug resistance, even against current drugs is making the effective selection of new drug targets together with lead compounds essential and urgent, requiring optimal approaches to be put in place for this process. The Discovery project is aimed at providing a publicly available informatics resource where comprehensive information on the parasite and host proteins are stored, together with the results from relevant 3rd-party investigations as well as results from our own high-throughput analysis. The comprehensive data included in the resource is aimed as wide as possible, including protein, gene-ontology, orthology, metabolic, structural, expression and chemoinformatics information. This is combined with a data-mining interface for researchers to perform the selection of putative drug target protein and lead compounds according to their specific highly-flexible criteria. Protein information includes data from the human, mosquito and the various malaria genome projects. Chemical information is from ChEMBL, DrugBank and PDB. Information includes basic annotations, motifs, domains, binding sites, structural features, orthology information, ontology terms, protein-ligand interactions and comparative genomics information. Chemical information includes protein interactions and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties. Searches may be initiated either from the protein or chemical compound as starting point. The system is available at: http://malport.bi.up.ac.za.
OP7: IDENTIFYING TREATMENT RELEVANT COLORECTAL CANCER SUBTYPES USING ITERATIVE NON-NEGATIVE MATRIX FACTORIZATION

Presenter: ANDREAS SCHLICKER, Netherlands Cancer Institute
Authors: Andreas Schlicker, Garry Beran, Christine M Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, Sara Davenport, Kerry Heathcote, Denis Alferez Castro, George Orphanides, Tim French, Lodewyk FA Wessels

ABSTRACT: Genetic and epigenetic features have been used to define colorectal cancer (CRC) subtypes and to take treatment decisions. To develop new targeted drugs, however, it is necessary to gain a better understanding of the molecular differences of CRC subtypes. We developed a new unsupervised approach for stratifying tumor samples using genome-wide mRNA expression data. Our method, iterative non-negative matrix factorization (iNMF), is based on the iterative application of non-negative matrix factorization based on randomly selected sets of genes. In a gene expression dataset consisting of 63 CRC tumors, we identified two dominant subtypes. These subtypes were highly concordant with the classes induced by an epithelial-mesenchymal-transition (EMT) gene expression signature. This finding is consistent with previous results. Further stratification of the tumor samples revealed five subtypes. These subtypes exhibit many differences, most notably differential activation of specific signaling pathways. Importantly, employing the derived subtype gene signatures, we stratified several independent, published datasets, suggesting that the signatures capture disease-relevant intrinsic features of CRC. Furthermore, application of the gene signatures to expression data obtained from cell lines revealed that the tumor subtypes were covered in all panels analyzed. Integrating pharmacological response data allowed us to identify several targeted compounds showing differential response across the subtypes. The CRC stratification obtained with our new method, iNMF, offers valuable insight into the differences between CRC subtypes at a functional level. Most importantly, it captures features of the disease that are highly relevant for the development of new targeted drugs in defined CRC patient sub-populations.

OP8: THE ORIGIN OF MAMMALIAN PLACENTATION CORRELATES WITH PROTEIN FUNCTIONAL SHIFT AND THE EMERGENCE OF NEW CONTROL MECHANISMS

Presenter: MARY O’CONNELL, Dublin City University
Authors: Thomas A. Walsh, Kieran Holohan, Anna O’Brien, Robert Carton, Elinor Velasquez, Claire C. Morgan, Noeleen B. Loughran, Mary J. O’Connell

ABSTRACT: Placenta is an important synapomorphy that defines the mammalian clade. From the fossil record we know that the first placental mammal lived approximately 125 MYA. Using disease databases and literature searches we have identified genes, networks and regulatory elements that are most likely to have played a large role in the evolution of the placenta. We have determined how and when these genomic elements arose and evolved.
over time. In particular, we focused on the respective roles played by gene duplication and loss, positive selection and novel regulation in the origin and evolution of placenta. We inferred the evolutionary history of each placental gene across a number of animal genomes. In doing so, we tested whether the selective pressures acting on some of the genes are interdependent. We have shown that many of the placental specific genes have undergone protein functional shift uniquely in the ancestral placental mammal lineage. We have also shown that the gene expression in this network is tightly regulated by a small subset of placenta-specific microRNAs discovered in this study.

**OP9: EXPLORING AND PROFILING LONG NON-CODING RNA IN T CELL DEVELOPMENT USING NEXT GENERATION SEQUENCING AND BIOINFORMATICS APPROACHES**

**Presenter:** Tzu Phang, University of Colorado Denver  
**Authors:** Tzu L. Phang, Ping Yao Zeng, Edwin F. de Zoeten

**ABSTRACT:** The human genome is estimated to contain ~23,000 protein coding genes, which represents less than 2% of the genome sequence. However, recent expression tiling microarray experiments revealed that up to 90% of the genome is actively transcribe. One of the transcription products is long non-coding RNA (lncRNA). LncRNAs are long (> 200 nucleotides) RNA that do not translate into functional proteins, and have no obvious open reading frame (ORF) for more than 100 amino acid in length. However, many believe that they have a regulatory role in the cell. We theorize that there is an extra layer of regulation that enables fine-tuning, silencing, or activation of the transcription process. Non-coding RNA appears to be a good candidate in fulfilling this role because they are often located and expressed during important developmental stages. The current estimation of ~7000 to ~23,000 lncRNAs implies that they must play a crucial role in maintaining the normal function of the cellular network. We have developed a bioinformatics workflow to concurrently identify and profile whole genome lncRNA and mRNA expressions in the T cell differentiation using the HiSeq next generation sequencing technology. The characterization of these long non-coding RNA species, their functions and clinical applications could be a major advancement of biologic and clinical importance.

**OP10: BUILDING AN INTERACTOME TO IDENTIFY SIGNALING COMPONENTS**

**Presenter:** Sarah Wyatt, Ohio University  
**Authors:** Sarah Waytt, Kaiyu Shen

**ABSTRACT:** A genomic-level analysis was utilized to study the gravitropic signal transduction in Arabidopsis. First, the gravity persistent signal (GPS) treatment was used to isolate the mechanisms of gravitropic signal transduction from those of response mRNA was extracted across a time course during
the GPS treatment and probed against an Agilent Arabidopsis 4X44k gene expression array. Instead of analyzing the data to find the “top statistically significant” genes, we used the raw data to construct a gravitropic signaling-specific interactome. First, an Arabidopsis gene interaction network was built from five databases: PAIR- Experimentally reported interaction, PAIR — High coverage interactions, TAIR, AtPIN, BioGRID. Next, a collection of expression profile data sets, representing potential background parameters, were obtained from TAIR and NASC databases to be used as filters. Pearson Correlation Coefficient (PCC) values were calculated for the GPS microarray data and the background parameters, and the original interactome was then reduced: gene interactions were kept only if there was a significant PCC value in the GPS data but not in the background parameter profiles. To further fine tune and train the interactome, text-mining of the literature was performed and functional annotations were collected and applied. These analyses generated an interactome that was specific for gravitropic signal transduction by adding/deleting interactions as well as assigning probabilities for the interactions using a Bayesian network approach. Thus, the important gravitropic signal transduction genes, along with the functional clusters and crucial pathways, could be selected for biological validation and study.

OP11: GENESMASH: A RESTFUL WEB SERVICE FOR GENE ANNOTATIONS

Presenter: GANIRAJU MANYAM, University of Texas Anderson Cancer Center
Authors: Michelle Payton, Ganiraju Manyam, Chris Wakefield, Jack Roth, Lynne Abruzzo, Kevin Coombes

ABSTRACT: With the proliferation of new technologies that provide different genome-wide overviews of the molecular landscape within cells, the bioinformatics challenge of integrating disparate sources of information continues apace. Integrated analysis of various biological data types has been a recurrent theme in many ongoing large scale initiatives to understand various molecular pathologies. A fundamental need is to integrate these data, which always requires matching probes across platforms, either by the genes they target or by their genomic coordinates. We present a new website and web service, geneSmash, to collate and provide gene-centric annotations. geneSmash is built upon the Apache open-source database platform, CouchDB. It is a document-oriented, schema-free database with a built-in web server. Database queries are processed through HTTP requests, which are handled by the RESTful JSON API. This feature provides universal accessibility to any modern programming language without any customized API. CouchDB also provides native support for incremental database replication. This would enable the users of geneSmash to maintain their local copy with automatic updates. Since geneSmash provides a generic web service for gene annotations, it can serve as a platform for other specialized applications. We have already developed several such applications including DrugBase. DrugBase is a programmer friendly database of drug-target interactions. It can be an essential tool to integrate drug-
ORAL PRESENTATIONS


OP12: INTEGRATING CURATED DATABASES AND TEXT MINING OUTPUT INTO A BIOMEDICAL KNOWLEDGE BASE

Presenter: KEVIN LIVINGSTON, University of Colorado Denver
Authors: Kevin M. Livingston, Michael Bada, William A. Baumgartner Jr, Yuriy Malenkiy, Lawrence E. Hunter

ABSTRACT: Biomedical researchers face a profound challenge in keeping track and making sense of numerous curated databases, literature published at an ever-accelerating rate, and data from high-throughput experiments. All of this information must be semantically integrated and methods to effectively reason over it developed if researchers are to have any hope of efficiently leveraging this flood of information. We have started by building representations of the information sources themselves through an extension of the Information Artifact Ontology (IAO) and are working toward transforming these records into representations in terms of biomedical concepts grounded in the Open Biomedical Ontologies (OBOs). The intermediate record representation allows access to the information while the biomedical representations are still being constructed and will serve as provenance information for the biomedical knowledge. We have produced over 8.35 billion RDF triples from 17 data sources including information about gene and gene products, pathways, diseases, and drugs. The foundation of the biomedical representation is integrating the many disparate identifiers for the genes and gene products in these sources, using the mappings from 12 different data sources. In parallel, we are working on the representation of annotations, also consistent with the IAO, to model the results of our automated text mining systems. Thus far, as a pilot experiment, we have integrated the results from mining 1,500 abstracts consulted by biologists in our lab in a recent discovery of theirs related to heart failure. We are scaling up to include information from MedLine abstracts and the PubMed Central Open Access documents.

OP13: USING EVOLVING PROTEIN NETWORKS TO INFORM THE GRAPH COLORING PROBLEM

Presenter: TODD GIBSON, University of Colorado Denver
Authors: Todd A. Gibson, Debra S. Goldberg

ABSTRACT: Bioinformatics has benefited from algorithms which were first applied in other domains. Examples include the hidden Markov model which was previously applied to speech recognition, and dynamic programming
which has figured prominently in economic analysis and other areas. Here we reciprocate by looking at how evolving networks of protein interactions can inform the graph coloring problem from mathematics and computer science. Using the evolutionary principles of duplication and subfunctionalization, a complete graph with k nodes can be reinterpreted as a protein interaction network which can evolve into any k-chromatic graph. For an extant protein interaction network, its evolutionary history can be inferred from a time-ordered list of duplications by reversing the duplications in the network. Likewise, by reinterpreting any simple graph as a protein interaction network, reversing it’s putative evolutionary history will eventually reduce it to a complete graph of k nodes, thereby providing a chromatic coloring of the original graph. By reinterpreting a graph as a protein interaction network, determining a chromatic coloring of the graph requires determining the time order of duplications. This ordering can be informed by analyzing the topological similarities between paralogous proteins (protein pairs born of a duplication).

**OP14: MACHINE APPROACHES TO RECOGNITION OF TRYPANOSOMAL VARIANT SURFACE GLYCOPROTEIN SEQUENCES**

*Presenter and Author: JON WILKES, Wellcome Trust Centre for Molecular Parasitology*

**ABSTRACT:** Support vector machines (SVMs) attempt classification of two or more populations of elements by encoding properties of each element as multi-dimensional vectors and defining a hyper-plane(s) separating populations. An SVM’s efficacy in classifying samples depends upon appropriate numerical transformations. We have developed software to rapidly prototype methods to codify aspects of sequence architecture (SVMseq) and generate SVMs. African trypanosomes evade host immune responses by expression of variant surface glycoproteins (VSGs), shielding invariant components of the cell surface. Highly immunogenic, they randomly switch between forms enabling sub-populations expressing different VSGs to expand under immune selection. Extensive sub-telomeric arrays of genes and pseudogenes exist and undergo splicing to generate an unlimited range of novel genes, driving antigenic variation. The VSG repertoires of trypanosome populations will, of necessity, show little sequence similarity; novel techniques for identification and classification of putative VSG (pseudo)genes are required. We applied the software tools of SVMseq to create efficient SVM models for the detection of VSG sequence from canonical sequences derived from 9 of the 11 mega-chromosomes of Trypanosoma brucei Treu-927. Canonical sequences from the remaining 2 chromosomes were used to benchmark performance. We developed a 2 stage method involving the identification of VSG arrays, followed by identification of the N- and C-terminal domains of the component genes. ROC analysis of performance produced AUC values of >0.96. Application of the SVM model to unannotated genomic sequence identified arrays of novel genes.
OP15: COMPACT ENCODING FOR GENE THERAPY
Presenter and Author: ROGER HALL, University of Arkansas at Little Rock

ABSTRACT: Synthetic biology is the practice of creating non-evolved organisms through direct manipulation of DNA and other cellular contents. The practice is generally applied to yeast and bacteria, and differs from transgenic or chimeric research. Cellular therapies using synthetic viral vectors have experienced at least one fatality, and continuing research is hampered by the concerns of further Hippocratic dilemmas. The AAV (adeno-associated virus) currently being used is limited to a package capacity of approximately 5000 bases. The Center for Gene Therapy at the University of Iowa School of Medicine reported using dual-vector approaches to deliver genetic information beyond the limit. Current therapies require manipulation of every virus to remove the existing DNA and insert the custom DNA. The vectors are considered safe since they have no viral DNA and are unable to replicate. Overlapping DNA may require higher GC content to accommodate the distinct functional domains, drawing greater mutational pressure, and may create live but fragile vectors with critically dense genomes and calculable generational iterations, enhancing vector safety and reducing cost of therapy. Our approach considers the limits of information encoding in overlapping reading frames using representative protein motifs of both replication and therapeutically significant genes.

OP16: MICROBIAL MONITORING AND SPACE EXPLORATION
Presenter: SVEN BILKE, NCI
Authors: Sven Bilke, Verena Starke

ABSTRACT: Microorganisms interfere with or are crucial to goals and systems for space flight, and, consequently, the development of effective microbial monitoring technologies is critical for mission safety and success. Recognizing the need for early microbial identification of single-cell organisms, a molecular-based technology for microbial monitoring has been developed based on a high-density oligonucleotide resequencing microarray. The custom array design interrogates the 16S or 18S rRNA gene sequence and rpoB (DNA-directed RNA polymerase subunit beta) gene sequences in 157 organisms, probing 527829 base pairs of sequence on a single array. It includes gene sequences with potential relevance to specific areas of space exploration: pathogens, extremophiles, and common microorganisms from the three domains of life—bacteria, archaea, and eukaryota. One of the challenges in the analysis of the resequencing data is based in the high degree of homology of the ribosomal sequences of distinct but related organisms. In this presentation we will describe our algorithm allowing robust discrimination of the re-sequencing signals, enabling us to detect, identify and approximately quantify DNA of the 157 microbial species, even when a mixture of organisms is present in the bio specimen.
OP17: CODINGMOTIF: EXACT DETERMINATION OF OVERREPRESENTED NUCLEOTIDE MOTIFS IN CODING SEQUENCES

Presenter: JEFFREY CHUANG, Boston College
Authors: Yang Ding, Andy Lorenz, Jeffrey Chuang

ABSTRACT: It has been increasingly appreciated that coding sequences harbor regulatory sequence motifs in addition to encoding for protein. These sequence motifs are expected to be overrepresented in nucleotide sequences bound by a common protein or small RNA. However, detecting overrepresented motifs has been difficult because of interference by constraints at the protein level. Sampling-based approaches to solve this problem based on codon-shuffling have been limited to exploring only an infinitesimal fraction of the sequence space and by their use of parametric approximations. We present a novel $O(N \log^2 N)$-time algorithm, CodingMotif, to identify nucleotide-level motifs of unusual copy number in protein-coding regions. Using a new dynamic programming algorithm we exhaustively calculate the distribution of the number of occurrences of a motif over all coding sequences that encode the same protein sequence, given a background for codon usage and dinucleotide biases. Our method takes advantage of the sparseness of loci where the motif can occur, greatly speeding up the required convolution calculations. Knowledge of the distribution allows one to assess the exact non-parametric p-value of whether a given motif is over- or under-represented. We demonstrate that CodingMotif identifies known functional motifs more accurately than sampling-based approaches in a variety of datasets, including ChIP-seq data for the transcription factors NRSF and GABP. CodingMotif provides a theoretically and empirically-demonstrated advance for the detection of motifs overrepresented in coding sequences. We expect CodingMotif to be useful for analyzing functions such as DNA-protein binding, RNA-protein binding, or microRNA-RNA binding within coding regions.

OP18: GO CLASSES OR BICLUSTERS? ALTERNATIVE APPROACHES FOR EXPLORING MICROARRAY PROBESET SUBSETS

Presenter and Author: GEORGE ACQUAAH-MENSAH, Massachusetts College of Pharmacy and Health Sciences

ABSTRACT: There are a number of emerging algorithms for exploring regulatory relationships among genes. Investigations of regulatory relationships on all possible probeset pairs in large compendia of gene expression data present at least two difficulties: First, are the memory costs. Second is the increased likelihood of selecting large numbers of false regulatory relationships. In this study, two approaches to focusing on subsets of probesets were examined. In the first, probesets associated with classes of Gene Ontology (GO) biological processes relevant to Chronic Obstructive Pulmonary Disease (COPD) were selected from a compendium of lung epithelium cell microarray
data, following processing via Robust Multi-Array Analysis. In the second, probesets associated with biclusters within the dataset were selected. The Context Likelihood of Relatedness (CLR) algorithm was used to learn regulatory relationships from the entire dataset. It was then also used on the selected subsets for each approach. The first approach, which in this case selected probesets associated with apoptosis; response to inflammation; and response to oxidative stress, has a tendency to bias the otherwise unsupervised learning of interactions that CLR affords. Key interactions involving probesets outside of the selected scope of GO classes thus are missed. Varying the number of biclusters used, the Factor Analysis for Bicluster Acquisition (FABIA) algorithm was used in the second approach. The second approach had the benefit of avoiding bias and involving probesets outside the selected scope of GO classes. Known participants in the COPD etiology, such as the Nfe2l2 product, were highlighted only in the second approach.

**OP19: ACCURATE INFERRING TRANSCRIPTION REGULATION FROM A COMPLEMENT OF EXPRESSION PROFILES**

*Presenter: XUELING LI, University of Texas Medical Branch  
Authors: Xueling Li, Dirar Homouz, Andrzej Kudlicki*

**ABSTRACT:** We present a new method for gene regulatory network inference, which combines support vector machines (SVM) and Copula transform to integrate multiple experiments. We illustrate the potential of the methodology on 1755 profiles of genome-wide gene expression of saccharomyces cerevisiae obtained with GPL90 Affymetrix Platform. The Copula-SVM model was validated on YEASTRACT regulation list, which include 48082 interactions between 183 transcription regulators and 6403 target genes. We generate similar number of negative data samples by including only the gene interactions between target genes (TGs). Results shown that copula-SVM has achieved a high accuracy of 95% on 4000 random positive data points and 4000 negatives data points. We further reduced the feature dimension by Wavelet Package Transform (WPT). Results show that feature reduction by WPT has achieved a satisfactory accuracy and have a significant improvement compared with using only one of profile series, yeast metabolic cycle profiles as gene features. Besides much higher performance in Receiver Operating Characteristic (ROC), our copula-SVM method is very general: it successfully predicted regulations by TFs not included in the training set (area under ROC is 0.81) while most methods can only predict new target genes of a known transcription factor given a couple of its known target genes, e.g. SIRENE. It can also predict the directionality of gene association while ARACNe and CLR, etc., can only infer if there is an association between genes. Also, copula-SVM can directly assign probability score to each gene pair to mine the possible false positives and false negatives.
**OP20: A PROPOSED ALGORITHM FOR EPISTASIS DETECTION USING FREQUENT ITEM-SETS AND MUTUAL INFORMATION**

Presenter: JAMES RUDD, North Carolina University  
Authors: James Rudd, Assefa Tesfay, ClarLynda Williams-DeVane, Gaolin Zheng

**ABSTRACT:** Increasing evidence highlights the role of epistasis in disease prognosis and progression. Thus, detection and interpretation of epistatic interactions has become a bioinformatic priority. In this project, we propose a novel algorithm for the detection of epistasis using frequent item-sets and mutual information. First, frequent item-sets up to size k are constructed in a fashion similar to association rule mining. The identified item-sets are then evaluated relative to the status variable using mutual information. The highest scoring item-sets are then assigned a p-value using the log likelihood statistic which can be approximated by a χ² distribution. The item-sets can then be ordered by their adjusted p-values and those above a significance threshold can be plotted as a network graph after multidimensional scaling. When compared to Multifactor Dimensionality Reduction (MDR), our algorithm has three distinguishing characteristics. First, we enable the modeling of binary and categorical status variables such as disease progression or severity. This enables a broader range of study designs. Second, we filter out multi-locus genotypes that occur at low frequency in the dataset; the relationship between these low-frequency variables and the outcome is uninterpretable. This dimension reduction accelerates computation. Third, we estimate model significance statistically rather than determining it empirically through permutation testing. This represents a significant time saving relative to the time complexity of the standard MDR 1,000 permutations. Overall, we expect this algorithm to demonstrate power comparable to or exceeding that of MDR while decreasing computation time.

**OP21: MOLECULAR INSIGHTS TO THE DRUG RESISTANCE OF V32I & M46L HIV-1 PROTEASE MUTANT TO INHIBITOR TMC114: FREE ENERGY CALCULATION AND MOLECULAR DYNAMICS SIMULATIONS**

Presenter: BISWA MEHER, Albany State University  
Authors: Biswa Ranjan Meher, Yixuan Wang

**ABSTRACT:** Mutations V32I and M46L are considered as two of the most key multi-drug resistant mutations of HIV-1 protease (HIV-pr). Inhibitors generally bind to the active site of the protease dimer. However, TMC114 can bind at two distinct sites, one in the active site cavity and the second on the groove of one of the flexible flaps. Surprisingly, TMC114 binds at these two sites simultaneously in two diastereomers related by the inversion of the sulfonamide nitrogen.
Existence of the second binding site, suggests the higher efficiency of this drug against HIV-pr. Nevertheless, even so the drug resistance in HIV-pr due to mutations cannot be ignored, which prompts us to investigate the molecular insights to the drug resistance and behavior of doubly bound TMC114 to the HIV-pr structure. In the present work, the conformational dynamics of HIV-pr and the binding of TMC114 to the WT, V32I and M46L mutant were investigated by MD simulations. The 10 ns MD simulation study shows many intriguing effects of inhibitor binding to the WT and mutant proteases. MM-PBSA calculations explain the binding free energies favorable for M46L mutant as compared to V32I and WT. The increased $\Delta G_{\text{vdw}}$ and $\Delta G_{\text{ele}}$ energies are mainly responsible for the higher binding affinity of TMC114 to M46L mutant, while the less binding affinity for V32I can be attributed to the higher entropic contribution. Furthermore, the mutations have differential effect on the residual behavior in the flap region and flap opening events in M46L is more stable than in WT and V32I.

**OP22: TESTING THE ORTHOLOG CONJECTURE WITH FUNCTIONAL DATA FROM SEVERAL PAIRS OF CLOSELY RELATED ORGANISMS**

*Presenter: WYATT CLARK, Indiana University*

*Authors: Wyatt T. Clark, Predrag Radivojac, Matthew Hahn*

**ABSTRACT:** A common assumption in comparative genomics is that orthologous genes share greater functional similarity than do paralogous genes (the ortholog conjecture). While this assumption is largely untested, many methods and databases developed for the inference of protein function explicitly ignore paralogs. In order to test this assumption we carried out the first large-scale test of the ortholog conjecture using comparative functional genomic data from human and mouse. We used experimentally derived functions of more than 8,900 genes, as well as an independent microarray dataset, to directly assess our ability to predict function using both orthologs and paralogs. Both datasets show that paralogs are often a much better predictor of function than are orthologs, even at lower sequence identities. Among paralogs, those found within the same species are consistently more functionally similar than those found in a different species. Our results, which were published in PlosCB, and featured in Nature Genetics, have implications for the computational prediction of protein function, and shed light on the relationship between sequence divergence and functional divergence. Here we extend our study of the ortholog conjecture to other pairs of closely related organisms in order to determine whether the same patterns hold across a wider range of organisms.
OP23: FUNCTIONAL PROFILING OF PHARMACOGENETIC NON-SYNONYMOUS SNPS

Presenter: CHET SEILINGMAN, Buck Institute for Research on Aging
Authors: Chet Seligman, Janita Thusberg, Jackson Miller, Emidio Capriott, Jim Auer, Michelle Whirl-Carrillo, Teri Klein, Sean Mooney

ABSTRACT: Functional profiling of pharmacogenetic non-synonymous SNPs
Little is known about the nature of pharmacogenetic (PGx) variants as compared
to disease-causing genetic variants and neutral genetic polymorphisms. Similarly
to disease-causing variants, the pharmacogenetic SNP may become invaluable
for personal genetic applications. We are employing bioinformatic methods,
to annotate the protein level consequences of pharmacodynamic (PD) and
pharmacokinetic (PK) variants in the PharmGKB® database to with the goal of
predicting candidate PGx variants. We are working towards a pharmacogenetic
fingerprint of features that describe protein variants. Our research involves two
separate strategies to separate PGx from neutral and PD variants from PK. The
first involves gene and protein attributes determined from GO, KEGG, Reactome,
HRPD and co-expression networks where known PGx entities are substantially
enriched relative to the human genome. These include such annotations as
Glycosylation (PD, PK), Metal ion binding (PD, PK), Intracellular signaling (PD)
and Oxidation-reduction (PK), using DAVID and GeneMANIA. The second
strategy operates at the variant level wherein we employ machine-learning
classifiers, such as Random Forest and Support Vector Machines, to known PGx
variants, in order to evaluate features that will differentiate these from neutral
SNPs or disease-causing mutations. Our dataset includes 143 known PGx
variants and over 26,000 known and presumed neutral ones. The imbalance in
this dataset results in a well known challenge of having an excellent ROC-AUC
along with relatively low precision. We will summarize our findings and describe
how a PGx variant compares to the molecular attributes that describe disease-
associated and neutral variants.

OP24: LIMITATIONS OF AUTOMATED ANNOTATION SOFTWARE
WHEN USED WITH DRAFT GENOMIC ASSEMBLIES

Presenter and Author: ROBERT NORGREN, Jr. University of Nebraska Medical Center

ABSTRACT: The genomes of many organisms have recently been sequenced.
This process has accelerated with the advent of NextGen sequencing. However,
almost all of the vertebrate genome assemblies are draft quality, not finished
assemblies. Among vertebrates, only the human and mouse genomes can
be considered finished (and even these are not complete). The distinction
between finished and draft genome assemblies is important when considering
the quality of annotations. Although automated annotation software appears
to work relatively well with finished assemblies, it often produces substantial
errors when applied to draft genome assemblies. For example, approximately
50% of the protein coding sequences were incorrectly annotated in the rhesus
macaque, for which a typical draft genome assembly is available. These errors
are due to: misassemblies of contigs, gaps in coding sequence and errors in
coding sequence. Even an error involving a single nucleotide can result in
spurious results. In some cases, no gene was annotated by the automated
annotator used by NCBI, Gnomon, even though the location of the gene was
obvious with manual inspection. In other cases, protein-coding genes were
incorrectly annotated as pseudogenes. A particularly troubling class of errors
occurred when Gnomon “invented” false exons and false protein sequences
because the real exon had not been sequenced. For nonhuman primates like
rhesus macaques, algorithms which incorporate synteny and orthology to
human genes perform better than ab initio programs. A modest amount of
NextGen sequencing can also greatly improve draft genomic assemblies and
their subsequent annotation.

**OP25: MIXED PRIMARY-SECONDARY STRUCTURE ALIGNMENT OF NCRNA COVARIANCE MODELS FOR MODEL CLUSTERING AND COMBINATION**

*Presenter and Author: JENNIFER SMITH, Boise State University*

**ABSTRACT:** One way to reduce the enormous computational burden of non-
coding RNA gene finding using covariance models is to cluster models for similar
families and find combined models that work adequately well for each family in
the cluster. When annotating new genomic data by searching with all models,
this reduces computation time by a factor equal to the number of clusters
divided by the number of families. Model clustering and combination relies
crucially on the metric used to represent model distance. Jiang and Weise used
a model distance based on base pair conflict count. An alternative is proposed
here based on multiple alignments using a mixed primary-secondary structure
alphabet where model consensus structures are used. The resulting clusters
are not the same as Rfam clans since the clusters generated here are designed
solely to reduce computation time, regardless of whether the members have any
functional relationship. Also, the Rfam grouping into clans gives no information
about how to go about constructing combined models. However, Rfam clans
can be used as a sanity check, since clan members should tend to end up in the
same clusters.
OP26: IN SILICO GENE EXPRESSION BASED ANALYSIS ON CLAUDIN FAMILY MEMBERS ASSOCIATION WITH HUMAN THYROID CANCER

Presenter: SHAUKAT MALIK, Mohammad Ali Jinnah University
Authors: Shaukat Malik, S. Sameen, Z. Khalid

ABSTRACT: Thyroid cancer is one of the major types of cancers worldwide. A large amount of effort has been put in order to find the genetic basis of this cancer. Many genes have been reported before out of which claudin family is one of them. The claudin protein family consists of more than 20 members which are made up of key structural rudiments inside the tight junction. The association of claudin-1 with the thyroid cancer has already been predicted experimentally. In order to investigate the genetic reason of human thyroid cancer computationally, the systematic analysis of claudin gene family has been carried out. To fulfill this task the bioinformatics methodologies are combined for the assessment of claudin gene family expressions. Results obtained showed and verified the association of CLDN1 member with the thyroid cancer.

OP27: DATA MANAGEMENT AND ANALYSIS SOLUTIONS FOR META-ANALYSIS OF MULTI-DOMAIN DATA

Presenter: CLARLYNDA WILLIAMS-DEVAN, North Carolina Central University
Authors: ClarLynda Williams-DeVane, Archana Radadia, Nina Rountree, Stephen Edwards

ABSTRACT: The ability to cost-effectively use high content technologies to elucidate the biological etiology of complex diseases such as asthma and the environmental factors contributing to them has greatly increased. More frequent use of high content technologies has led to an exponential increase in data that require more sophisticated data management and in silico analysis solutions. To address this need, we developed a Wiki-based Data Management System (WikiLIMS) for the management of data and information associated with toxicogenomic experiments. WikiLIMS currently includes experimental plans, procedures, and details as well as gene expression, clinical chemistry, histopathology, and observational data. WikiLIMS exhibits a unique organizational structure allowing efficient data management, collaborative work between experimentalist and data analyst, and comparison and/or meta-analysis across multiple individual experiments. Two case studies will be presented: 1) a cardiopulmonary cross-laboratory meta-analysis project and 2) the Mechanisms of Childhood Asthma Study, to better explain the data management and analysis solutions necessary to capitalize on the influx of data. Case Study 1, a meta-analysis of multiple studies designed to better understand the impact of air pollutants on cardiopulmonary disease will be presented to illustrate the value of the WikiLIMS System in this context. Case Study 2 will be presented to provide an overview of new data analysis methodologies currently in development for multi-domain in silico data analysis. The ability to capitalize on new and existing data is paramount to scientific progress, and the WikiLIMS data management system greatly enhances our ability to leverage
ABSTRACT: One of the most common outcomes of high-throughput biological experiments is a list of genes or proteins of interest. In order to explain the observed changes of these specific genes and to create new hypotheses one needs to understand the functions and roles of the genes in the lists under the condition studied in the experiment. Here we present two novel applications that facilitate the extraction of functional content of lists of genes. With STOP we overcome the limitations of manually annotated ontologies (like GO) and use automatic annotation techniques using descriptive text as an annotation source. This resource utilizes the National Center for Biomedical Ontology’s database, which includes over 200 biomedically related ontologies. These automatic annotations are then processed and used for enrichment analyses against submitted gene lists. In recent studies, we demonstrated how accurate these automatic annotations can be. STOP ranked 7th in a comparison of functional annotation tools when trying to predict novel annotations in the biological process category of GO. Our second application, DEFOG, eases the functional analysis of gene sets by hierarchically organizing them into functional related groups using data fusion of high-throughput experimental data. The underlying computational pipeline utilizes the state-of-the-art applications GeneMANIA, Transitivity Clustering, and Ontologizer for gene set specific network fusion, non-agglomerative hierarchical clustering, and GO term enrichment respectively. DEFOG allows for a novel visual analysis of gene sets that aids in the discovery of potentially important biological mechanisms and assists in the generation of new hypotheses from gene lists.

ABSTRACT: DAVID (the Database for Annotation, Visualization and Integrated Discovery) is a free web-based online bioinformatics resource that aims to provide tools for functional interpretation of large gene/protein lists. It has been used by researchers from more than 5,000 institutes world-wide with a daily submission rate of over 1,200 lists with over 3,500 citations since its first release in 2003. However, the usage of DAVID has been limited to the current web interface which does not support programmatic access and a URL-based API which is limited to defaults due to its size limit and stateless nature. DAVID-WS is developed to provide stateful web services for users to interact with DAVID programmatically and allows users to change background populations, reset functional parameters, and select species and categories.
for analysis, as well as to provide the ability to query all tools within the same session and format output as desired. Our performance testing shows that it took about 6 to 9 seconds to generate the output for computationally intensive client tasks such as Gene Functional Classification or Functional Annotation Clustering with 2,000 genes. The client code provided by DAVID-WS can be easily integrated into programs, work flows and interactive analysis tools as computational components.

**OP30: PREDICTING TRANSCRIPTION FACTOR BINDING SITES WITH HIDDEN MARKOV MODELS BY USING CHIP-SEQ DATA**

*Presenter: ANTHONY MATHELIER, University of British Columbia*  
*Authors: Anthony Mathelier, Wyeth W. Wasserman*

**ABSTRACT:** Transcription factors (TFs) are proteins implicated in transcriptional regulation by activating or repressing genes. Finding where those proteins bind to DNA is of key importance to decipher gene regulation at a transcriptional level. As TFs bind to DNA in part through sequence specificity, computational biology has become of great interest to predict transcription factor binding sites (TFBSs). Classically, computational prediction of TFBSs is based on position weight matrices (PWMs). Such models do not allow spacers or flexible length motifs and make the strong assumption that each nucleotide within a TFBS participates independently in the corresponding DNA-protein interaction. We propose to use Hidden Markov Models (HMMs) for the prediction of TFBSs. HMMs are flexible and can model position interdependence within TFBSs as well as spacers and variable length motifs. Constructed HMMs are able to model binding DNA sequences (corresponding to “matching states”) with their surrounding flanking regions (corresponding to “background states”). The availability of thousands of experimentally validated DNA-TF interaction sequences coming from ChIP-Seq, hopefully representing all the different properties of the associated TFBSs, allows us to construct and train HMMs to reflect the TFBS properties observed in experimental data. HMMs capture deep properties of binding sites that can be discovered in ChIP-Seq data and can be used to predict potential TFBSs accurately. Using ROC curves, the HMMs have been assessed on several ChIP-Seq data sets and we found that they obtain better results than PWMs in discriminating motifs within ChIP-Seq sequences from background sequences.

**OP31: THE CRITICAL ASSESSMENT OF FUNCTION ANNOTATION EXPERIMENT: A COMMUNITY-WIDE EFFORT TOWARDS A BETTER FUNCTIONAL ANNOTATION OF GENES AND GENOMES**

*Presenter: PREDRAG RADIVOJAC, Indiana University*  
*Authors: Predrag Radivojac, Sean Mooney, Iddo Friedberg*

**ABSTRACT:** A major challenge of the post-genomic era is understanding gene and protein function so that genes can be placed in the proper biological
context. However, due to the rapid growth of the number of sequenced genes and relatively slow functional characterization in the lab, protein function can only be predicted for the vast majority of the sequenced genes. Here we present the results of the first Critical Assessment of Function Annotations (CAFA) held during 2010-2011. Thirty-four research groups worldwide have participated in this experiment, with over 50 function annotation algorithms. We provided a list of 48,298 targets from Swiss-Prot, taken from several model organisms, for the participants to annotate. 594 of these targets have accumulated experimentally verified annotations after the prediction process ended, and were thus fit to be used as a blind benchmark for evaluating the accuracy of the annotation algorithms. The prediction methods were assessed using ROC curves, precision/recall curves, and variations on semantic similarity as applied to the Gene Ontology. I will present the scientific and technical challenges involved in setting up such a large scale community experiment. The results of the experiment will be shown including the top ranking methods according to different accuracy measurements. Other findings that will be discussed are: which terms are best predicted? Which proteins are best annotated? Which classes of methods perform best under different assessment measures? Finally I will discuss the impact CAFA is expected to have on the computational functional annotations, and future plans for this effort.

OP32: IN SILICO RATIONAL DRUG DESIGN AND MODELING STUDIES OF NOVEL INHIBITORS FOR MULTI-TARGET INHIBITION IN PSEUDOMONAS AERUGINOSA

Presenter: JAYARAMAN PREMKUMAR, Nanyang Technological University
Authors: Premkumar Jayaraman, Lim Chu Sing Daniel, Meena K Sakharkar

ABSTRACT: Pseudomonas aeruginosa is a major nosocomial and opportunistic pathogen, with the ability to develop multi-drug resistance which significantly reduces the efficacy of many commercially available antibiotics. Hence, there is an urgent need to develop novel therapeutic strategies to combat the development of resistance. On the sequence of our previous work on drug combination assays and common pharmacophoric features data, we have designed and evaluated new drug scaffolds, containing 6-methoxypyrimidine-2,4-diamine analogue conjugated with benzenesulfonamide derivatives, an alternative novel drug-design strategy for inhibition of quadri-enzymes (dihydropteroate synthase (DHPS), dihydrofolate reductase (DHFR), DNA gyrase subunit B and topoisomerase IV subunit parE (topoIV)) of two different pathways. To validate the hypothesis that these novel scaffolds could act as multi-target inhibitors, we have used computational techniques such as molecular docking study, molecular and electronic properties analysis and dynamics simulations. The docking model and dynamics study predicts that these inhibitors have favorable binding to all the four quadri-enzymes, forming strong hydrogen and hydrophobic interactions with key active site residues. The predicted structure-activity relationship of the proposed hybrid drugs with respect to physico-chemical properties and stereo-electronic
properties such as HOMO, LUMO and molecular electrostatic potential maps calculated using quantum chemical methods were found to be well correlated with the common pharmacophoric features required for the multi-target inhibition. The key innovative aspects of this study is to provide novel insights on preventing the emergence of drug resistance based on a rational multi-target drug design and can serve as a prospective lead in the development of anti-pseudomonal drug development.

**OP33: PATHCMAP: DEVELOPMENT OF PATHWAY SIGNATURE SYSTEM FOR IDENTIFYING DRUGGABLE PARTNERS OF SYNTHETIC LETHAL GENES IN CANCER**

Presenter: JIHYE KIM, University of Colorado Denver School of Medicine
Authors: Jihye Kim, Carlos H C Cano, Aik Choon Tan

**ABSTRACT:** While targeted therapies have shown clinical promise in treating solid tumors that “addicted” to oncogenic pathways, these therapies are rarely curative for advanced cancers. As most cancers are typically diagnosed at advanced stages and acquired resistance mechanisms that can protect the cells from these targeted therapies. Therefore, the discovery of pathways that mediate these compensatory survival mechanisms could reveal novel therapeutic targets for cancer cells “addicted” to these pathways. High throughput methods to identify small molecules that target synthetic lethal genes and pathways are critical for facilitating rational combination with targeted therapies of interest, and may have translational potential in clinical cancer trials. Here, we developed a novel bioinformatics tool “Pathway Connectivity Map (PathCMap)” that systematically connects the most similar pathway expression profile from a reference profiles database and extrapolates the most effective drug for individual query pathway signatures. The PathCMap approach is based on the connectivity map concept. It is based on 1) gene expression can be measured accurately and has shown promise as the “universal language” in disease characterization and prognostication; 2) gene expression can be used to connect different biological states and systems; 3) biological pathways drive disease phenotypes and the connectable traits as well. Using PathCMap, small molecules can be easily identified in combination with the targeted therapy which is analyzed and interpreted by our recently developed bioinformatics pipeline (BiNGS!SL-seq). This approach can be applied to query various cancer pathways and can be easily translated into a drug discovery platform.

**OP34: IDENTIFYING SINGLE COPY ORTHOLOGS IN METAZOAN**

Presenter: CHRISTOPHER CREEVEY, Teagasc
Authors: Chris Creevey, Jean Muller, Tobias Doerks, Julie D. Thompson, Detlev Arendt, Peer Bork

**ABSTRACT:** The identification of single copy (1-to-1) orthologs in any group of organisms is important for functional classification and phylogenetic studies. The Metazoa are no exception, but only recently has there been a
wide-enough distribution of taxa with sufficiently high quality sequenced genomes to gain confidence in the wide-spread single-copy status of a gene. Here, we present a phylogenetic approach for identifying overlooked single-copy orthologs from multigene families and apply it to the Metazoa. Using 18 sequenced metazoan genomes of high quality we identified a robust set of 1,126 orthologous groups that have been retained in single-copy since the last common ancestor of Metazoa. We found that the use of the phylogenetic procedure increased the number of single-copy orthologs found by over a third more than standard taxon-count approaches. The orthologs represented a wide range of functional categories, expression profiles and levels of divergence. To demonstrate the value of our set of single-copy orthologs, we used them to assess the completeness of 24 currently published metazoan genomes and 62 EST datasets. We found that the annotated genes in published genomes vary in coverage from 79% (Ciona intestinalis) to 99.8% (human) with an average of 92%, suggesting a value for the underlying error rate in genome annotation, strategies for identifying single copy orthologs in larger datasets. In contrast, the vast majority of EST datasets with no corresponding genome sequence available are largely under-sampled and probably do not accurately represent the actual genomic complement of the organisms from which they are derived.

**OP35: MINING GENOMES TO UNDERSTAND THE ORIGIN OF VISION IN METAZOA**

*Presenter: DAVIDE PISANI, National University of Ireland, Maynooth*

*Authors: Roberto Feuda, Davide Pisani*

**ABSTRACT:** Understanding the origin of vision in Metazoa is proving difficult. This is mostly because of the low level of similarity between visual Opsins, and proteins in other families belonging to the Class-A GPCR superfamily. Here we present results of a large scale analysis of complete animal genomes, which was performed to better elucidate what is the closest sister group of the Animal Opsins, and hence to better elucidate the origin of vision in Metazoa. In addition we present new analyses of the Opsin family that we performed to elucidate the history of gene duplications and deletions within this family, and we compare our results with those of previous studies. We found that the choice of multiple sequence alignment software used can strongly influence the results obtained for the phylogeny of this protein family; casting a doubt on results obtained in previous studies. We suggest that for the specific case of this protein family the use of the Phylogenetic-aware multiple sequence alignment software “Prank” substantially improve the results obtained. Results obtained using this multiple sequence alignment software imply more parsimonious gene histories (less duplications and deletions necessary to explain the distribution of opsins in extant taxa), and display better-supported nodes.
ORAL PRESENTATIONS

OP36: EVOLUTIONARY PATTERN EMBEDDED IN THE LENGTHS OF PROTEINS AND THEIR STRUCTURAL UNITS

Presenter: MINGLEI WANG, University of Illinois at Urbana-Champaign
Authors: Minglei Wang, Cedric Debes, Frauke Gräter, Gustavo Caetano-Anollés

ABSTRACT: The lengths of proteins are subject to systematic variation that reflects distinct evolutionary patterns in the three superkingdoms and relates to important characters of proteins (e.g. mechanical attributes). Our examination of protein structures in ~1,000 proteomes shows the existence of reductive evolution in proteomes and protein structures. Protein length differences between superkingdoms are due to much shorter non-domain sequences in prokaryotic proteins, while domain lengths tend to be shorter along with the evolutionary timeline. Investigations of protein architectures for mechanical function indicate that relative contact order of domain would decrease in evolution. Therefore, shorter lengths are favored by domains when they are becoming more open and less globular by diminishing long-range and favoring short-range interactions, which leads to faster folding. It indicates that domain length plays an important role during evolution towards dynamic regulation.

OP37: A MULTIPLE-TEMPLATE APPROACH FOR PROTEIN THREADING

Presenter: JIAN PENG, Toyota Technological Institute at Chicago
Authors: Jian Peng, Jinbo Xu

ABSTRACT: Due to the increasing number of solved structures, a protein without solved structure is very likely to have more than one similar template structures. Therefore, a natural question to ask is if we can improve modeling accuracy using multiple templates. This work describes a new multiple-template threading method to answer this question. At the heart of this multiple-template threading method is a novel probabilistic-consistency algorithm that can accurately align a single protein sequence simultaneously to multiple templates. Experimental results indicate that our multiple-template method can improve pairwise sequence-template alignment accuracy and generate models with better quality than single-template models even if they are built from the best single templates while many popular multiple sequence/structure alignment tools fail to do so. The underlying reason is that our probabilistic-consistency algorithm can generate accurate multiple sequence/template alignments. In another word, without an accurate multiple sequence/template alignment the modeling accuracy cannot be improved by simply using multiple templates to increase alignment coverage. According to the CASP9 official evaluation, our method outperforms almost all other CASP9 servers and our method generated the best alignments for the 50 hardest template-based modeling targets. Our method was also voted by the CASP9 community as one of the most innovative and interesting methods. Our method will have greater potential in the near future when many more templates are available due to the increasing number of solved structures. We can further improve alignment accuracy by extending
our algorithm to simultaneously thread multiple homologous sequences to multiple templates.

**OP38: A NEW PROBABILISTIC MODEL IN PREDICTIVE MICROBIOLOGY (NPMPM)**

*Presenter: NADINE SCHOENE, Goethe University Frankfurt*  
*Authors: Nadine Schone, Alexander Bockmayr, Bernd Appel, Annemarie Kaesbohrer*

**ABSTRACT:** Predictive microbiology is a basic component of microbial risk assessment. It can help to prevent foodborne disease outbreaks by detecting probable contamination sites and defining monitoring points. The goal of existing models in predictive microbiology is to gain understanding of population kinetics. The global error is minimized by fitting the parameters with pooled experimental data. But for a risk assessment it is essential to predict the bacterial count at a single time point (at the end of the process chain) as precise as possible, i.e., the local error has to be minimized. The New Probabilistic Model in Predictive Microbiology (NPMPM) is based on a new approach for including variability and uncertainty into modeling of microbial growth. It was developed for risk assessment of bacterial contamination in the food supply chain. The NPMPM was implemented as an R package. Model assumptions are kept simple and include exponential growth as basic form of population kinetics and log-normal distributions for bacterial counts. Internal validation with simulated data that fulfilled all model assumptions showed that the NPMPM is able to reproduce the data it was built with. For certain conditions this was shown theoretically, too.

**OP39: REDISCOVERY OF THE P53 TRANSCRIPTOME**

*Presenter: MARY ALLEN, University of Colorado*  
*Authors: Mary Allen, Robin Dowell, Joaquin Espinosa*

**ABSTRACT:** The guardian of the genome, p53, is a transcription factor that can activate transcription of target genes in response to cellular stress. Thus, p53 plays a central role in the regulation of cellular processes and the suppression of cancer. Some of the most potent tumor suppressors and critical targets for cancer therapies are controlled by p53. However, the order of activation of target genes or even all of the targets of p53 cannot be identified by current methods. We are using a novel technique, Global Nuclear Run On Sequencing (GRO-seq), to determine the earliest targets of p53 transcriptional activation. We have treated cancer cells with a drug that induces p53, and then preformed nuclear-run on followed by next-generation sequencing of only the nascent transcripts. Interestingly, preliminary analysis of the sequencing results indicate many target genes and unannotated non-coding RNAs are transcriptionally up-regulated one hour after p53 activation, much sooner than anticipated. Additionally, comparison of our data with known p53 binding sites indicates transcription overlaps many p53 binding sites.
**OP40: DEVELOPING MAXIMUM LIKELIHOOD AND BAYESIAN SUPERTREES**

*Presenter: WASIU AKANNI, National University of Ireland, Maynooth*

*Authors: Wasiu Akanni, Davide Pisani, Peter Forster, Mark Wilkinson*

**ABSTRACT:** Little work has been done on the development of supertree methods in the Likelihood and Bayesian frameworks. Recently, it has been proposed that Maximum Likelihood (ML) supertrees could be developed by using an exponential distribution to model the probability that the input trees could be erroneous. When the tree-to-tree distances used in the ML computation are calculated using the Symmetric Difference, the ML supertree has been shown to be equivalent to a Majority Rule Consensus Supertree, and hence, exactly as the latter, the ML supertree must have the desirable property of being a median tree with reference to the input set. In addition, the ability to estimate the likelihood of supertrees, will allow the implementation of Bayesian MCMC approaches, which have the advantage of allowing for a natural estimation of the support for the nodes on the recovered supertree. We have developed the first software for the estimation of Maximum Likelihood and Bayesian supertrees. The program is being written in Python and will also exploit the capabilities of already available software, i.e. P4. Here, we present results of reanalyses of the datasets of Holton and Pisani (2010) and Pisani (2007) and present the first Bayesian Genomic Supertrees generated using our new software. In addition, we shall compare these supertrees with those derived using other common supertree methods (e.g. Matrix Representation with Parsimony and Average Consensus). We show that results obtained using our method compared well with those obtained using MRP, and significantly outperforms all other methods.

**OP41: PREDICTION OF OPERONS IN MICROBIAL GENOMES BY INTEGRATING DIVERSE INFORMATION SOURCES**

*Presenter: ANIS KARIMPOUR-FARD, University of Colorado Anschutz Medical Campus*

*Authors: Anis Karimpour-Fard, Lawrence E. Hunter*

**ABSTRACT:** The number of published sequenced genomes has been growing in recent years, and at the present time, about 2000 microbial genomes are fully sequenced. The next step after sequencing is to predict genes and their functions from the sequence. Here we used Gene Cluster information for generating protein interaction networks. We hypothesize that in bacteria, the protein interaction networks derived via Gene Cluster and integration of different information could similarly improve methods for predicting operons. Previous studies have shown that operons tend to have short distances between their genes in bacteria. Unfortunately, predictions based on intergenic distance alone increase false positive so other sources of information must be added to bring the specificity to an acceptable level. Progress has been made toward a more generalized method for operon prediction based on a variety of diverse
information sources, including codon usage statistics, and identification of promoter and terminator sequences. However, very little has been done to examine the relative contribution of these features, individually and in combination, for operon prediction in genomes other than the genome(s) on which a prediction program is trained. We validate predicted Gene cluster pairs against known operons from E. coli K12 and B. subtilis because only these two organisms have a substantial number of experimentally verified operons. We examined whether features based on intergenic distance and protein function prove informative for both genomes. Moreover, we examine how combining interactions predicted by the other methods can improve Gene cluster operon prediction.

**OP42: GENE LANGUAGE MODEL**

*Presenter and Author: SIAMAK REZAEI, Talai*

**ABSTRACT:** There has been research in providing a stochastic model of DNA. Here we introduce a stochastic model of Genome. Past research on gene order has focused on breakpoint analysis of genome (e.g. Blanchette and Sankoff) and the relations among different chunks of genome in different species and the distance in transforming one gene order sequence to another with minimal number of operations. A Gene language model will provide a more detailed picture of the interrelations existing in genome of different species and faces less problems in representing circular genomes. A statistical gene language model assigns a probability to the linear sequence of genes using the ngam notion. A bigram probability of (g1,g2) in a bigram gene language model captures the number of times the specific gene sequence g1.g2 have appeared in one or more set of genomes compared to other binary gene sequences. A trigram probability of (g1,g2,g3) in a trigram gene language model captures the number of times the specific gene sequence g1.g2.g3 has appeared in the genome. In general, an ngram of (g1, ..., gn) in an ngram gene language model represents the number of times that a specific gene sequence g1...gn have appeared in one or more set of genomes. The words bigram and trigram gene language model denote n-gram gene language models with n=2 and n=3 respectively. Additionally, in a gene language model, the ngrams can be specified to be uni-directional, or bi-directional.

**OP43: LIVE AND LET DIE**

*Presenter: CHRISTIAN FORST, University of Texas Southwestern Medical Center  
Authors: Nassim Sohaee, Christian V. Forst*

**ABSTRACT:** Pathogens and host engage in an endless battle trying to out-compete each other. Pathogens (viruses in our case) attempt to hijack host processes for their advantage and for replication, whereas the host tries to defend against such attacks. Together with biochemical network information
we were using two classes of siRNA screens to identify (i) host and resistance factors for viral replication, and (ii) processes for host cell-survival and death. For this purpose we have developed a method to identify dense sub-networks in large networks. The method is based on finding bounded diameter subgraphs around a seed node. With this method we were able to identify distinct cellular factors and processes with respect to four distinct phenotypes: (i) resistance factors required for cell survival, (ii) restriction factors inducing cell death, (iii) host factors required for infection, and (iv) host factors that specify death in response to infection. Thus, we are not only able to pin-point processes of the viral life cycle, but also to identify host-defense pathways with diagnostic and therapeutic potential.

**OP44: BROAD SEMANTIC CLASS ASSIGNMENT FOR BIOMEDICAL TEXT**

*Presenter: KEVIN BRETONNEL COHEN, University of Colorado School of Medicine*

*Authors: K. Bretonnel Cohen, Thomas Christiansen, Lawrence E. Hunter*

**ABSTRACT:** Semantic class assignment is important for a wide variety of biomedical text mining applications, including information extraction, coreference resolution, and document classification. However, previous approaches to semantic class assignment have only attempted a small number of categories, typically three. We present work showing that it is possible to assign as many as twenty semantic classes to biomedical text. We evaluate the work with three methods—a corpus of full-text journal articles, a structured test suite, and a novel method involving using ontologies as input texts. We achieve micro-averaged precision/recall/F-measure of 67.06/78.49/72.32 and macro-averaged precision/recall/F-measure of 69.84/83.12/75.31 on the corpus, and accuracy of 77.12 to 95.73. The structured test suite reveals a number of characteristics of the performance of the approach, and the ontology-as-input method reveals a small number of errors in the implementation.

**OP45: ON THE ACCURACY OF PROTEIN TERTIARY STRUCTURE COMPARISON**

*Presenter: ALEKSANDAR POLEKSIĆ, University of Northern Iowa*

*Authors: Aleksandar Poleksic, Mauricio Arriagada*

**ABSTRACT:** One of the most essential tasks in computational molecular biology is to detect and quantify the similarity of two given protein tertiary structures. Attempts to address this task are faced with two difficulties. First, the similarity of protein structures is a relative notion: two proteins can be very similar in one aspect and less similar (or dissimilar) in another aspect. And second, the protein structure matching problem is computationally hard. To put it differently, in order to find a solution reasonably close to optimum, a typical algorithm for protein structure matching needs to generate and score a large number of different structural orientations of the input proteins. At present,
the only solutions that can be generated in a timely manner are the heuristic solutions. Due to the lack of a widely accepted alignment quality metric and in the absence of a fast algorithm for optimal protein structure alignment, the quality of heuristic solutions generated by current protein structure alignment methods is difficult to assess on a large scale. In this study, we utilize an automated method, capable of rigorously optimizing a basic measure of protein structure similarity, to demonstrate a significant increase in the accuracy of several heuristic protein structure alignment methods obtained by fine sampling of the protein superposition space. Searching the proteins’ superposition space in minute details benefits all methods subjected to our benchmarks, across several different and commonly used protein structure alignment precision metrics.

**OP46: A DISTRIBUTED FRAMEWORK FOR COMPUTATION ON THE RESULTS OF LARGE SCALE NLP**

*Presenter: CHRISTOPHE ROEDER, University of Colorado, Anschutz Medical Campus
Authors: Christophe Roeder, William Baumgartner Jr, Kevin Livingston*

**ABSTRACT:** Biomedical researchers produce new information about genes or gene products at a rate that can overwhelm a scientist attempting to keep up with his or her field of study. Natural language processing can assist researchers in navigating the deluge by helping to summarize this information. We are developing a framework to look at large corpora of documents that will allow us to extract information and identify interesting or emerging trends. We apply NLP processes in parallel to large numbers of documents and produce RDF representations of the document content such as identifying PRO or GO entities. The RDF from each document is collated using Map/Reduce computation to produce summarizations of the data such as entity co-occurrence and publication trends. This framework is implemented using cluster computing resources to process a large corpus we have assembled from Pub Med Central Open Access as well as documents obtained from publishers. Continuous updates keep the corpus up to date. The cluster resources include Colorado’s Janus cluster that is comprised of over 1300 nodes, each with 12 cores and 24GB of memory. We use a combination of batch schedulers such as Torque to generate the data from UIMA processes and Hadoop to run the Map/Reduce computation. We have implemented a prototype system on this framework to identify protein publication trends over time. The trends are made available on a web site charting them in a style similar to the Billboard top 100 chart familiar to US radio listeners.
**OP47: PROBABILISTIC SEARCH FRAMEWORKS FOR PROTEIN CONFORMATIONAL SPACES**

*Presenter and Author: AMARDA SHEHU, George Mason University*

**ABSTRACT:** A fundamental issue in our understanding of biology and treatment of disease concerns elucidating the structures and motions that proteins employ for their biological function. Understanding proteins in silico involves searching a vast high-dimensional conformational space of inherently flexible systems with numerous inter-related degrees of freedom and complex geometry. Our recent work focuses on probabilistic search frameworks for protein conformational spaces. A novel robotics-inspired framework is proposed that employs information on the space it has explored so the search can adaptively focus itself and further resources to relevant regions of the conformational space. Inspired by the use of subdivisions and projections of the robot configurational space in sampling-based motion planning, the framework employs projections of the explored conformational space and its associated energy surface in order to bias a search tree towards low-energy under-sampled regions. Inspired by evolutionary search strategies for search spaces rich in local minima, another framework is proposed that explicitly samples local minima. These two approaches are combined to reveal a powerful probabilistic search framework that enhances sampling of the protein conformational space. Extensive applications and analyses on an extensive list of proteins of different lengths and native topologies suggest the proposed efforts greatly enhance the sampling of the protein conformational space. Comparisons with well-known methods in protein structure prediction show that the frameworks push the state of the art and efficiently recover functionally-relevant conformations. Interesting insight is also obtained on how to tackle the dimensionality challenge both in protein chains and robotic articulated mechanisms.

**OP48: PREDICTING HIV-1, HUMAN PROTEIN INTERACTOME THROUGH INDIRECT AND DIRECT EVIDENCES**

*Presenter: OZNUR TASTAN, Microsoft Research New England*  
*Authors: Oznur Tastan, Jaime Carbonell, Judith Klein-Seetharaman*

**ABSTRACT:** HIV-1 depends on its host for virtually every aspect of its life cycle. Communication between pathogen and host are mediated by physical interactions between the virus and the host proteins. Previously, we built a supervised classifier for predicting HIV-1, host protein protein interactions (PPIs) using 35 features derived from various biological datasets (Tastan et al. (2009) Pac. Symp. Biocomput., p. 516-27). These features carried indirect information about PPIs such as gene expression of host proteins in the presence of HIV-1 infection, domain and sequence pairs, functional annotations of the proteins and relationship of human proteins with one another. Since then, several new biological information sources pertinent to HIV-1, host relation have become available. Amongst these, there have been four siRNA screens...
been published to establish the human proteins important for HIV-1 function. These datasets include i) four genome-wide RNA inference screens aimed at establishing host proteins required for HIV-1 replication, ii) a high-throughput immunoprecipitation pull-down mass spectrometry assay for detecting viral and host PPIs, iii) a set of human proteins reported to interact with viruses’ proteins and iv) knowledge of human proteins detected in budding HIV-1 virions. We utilized our knowledge of the context of the host cellular machinery to incorporate new features into our models based on these datasets. The new model trained with random forest classifier utilizing the richer feature set outperformed the first model by a 13% relative increase in mean average precision. The set of predicted interactions serves as biological hypotheses to test.

**OP49: USING INFORMATION THEORY TO MAP READS TO A REFERENCE SEQUENCE**

**Presenter and Author:** JOHN CONERY, University of Oregon

**ABSTRACT:** I am developing a new algorithm for mapping short reads to a reference sequence. The goal is to use the information content (derived from the quality scores generated by the base caller) to more accurately trim and align the reads. For each position in a read we determine the probability distribution for the nucleotides that might occur in that position. The distribution can then be used to calculate bit scores for mismatches in the alignment, or to compute the entropy of a position. One example of how entropy is used is a sliding window algorithm that trims low quality characters from the 3’ end of a read: compute the sum of the entropy values of the positions in the window, and move the window toward the 5’ end until the sum of values in the window falls below a threshold. This approach can potentially save high quality bases near the 3’ ends that would otherwise be removed by a filtering algorithm that trims every read to the same fixed length. Another advantage is that the method requires fewer parameters, e.g. the window size and entropy threshold can all be derived automatically from the size of the reference sequence.
**POSTER LIST**

P01: IDENTIFYING NOVEL COMPONENTS OF THE GRAVITROPIC SIGNAL TRANSDUCTION PATHWAY VIA TRANSCRIPTOME ANALYSIS  
Presenter: ZACHARY ABRAMS, Ohio University

P02: DEVELOPING MAXIMUM LIKELIHOOD AND BAYESIAN SUPERTREES  
Presenter: WASIU AKANNNI, National University of Ireland, Maynooth

P03: OPPORTUNITIES AND CHALLENGES ASSOCIATED WITH COMPUTATIONAL RNASEQ DATA ANALYSIS  
Presenter: DAVID ASTLING, University of Colorado School of Medicine

P04: JIKITOU A QUESTION ANSWERING SYSTEM: THE NEXT STAGE IN THE EVOLUTION OF INFORMATION RETRIEVAL  
Presenter: MICHAEL BAUER, University of Arkansas at Little Rock

P05: PHYLOGENETIC SIGNAL DISSECTION AND THE PHYLOGENY OF THE ECDYSOZOA  
Presenter: LAHCEN CAMPBELL, The National University of Ireland, Maynooth

P06: A NETWORK-BASED SURVIVAL PREDICTOR AND ITS APPLICATION TO CYTOGENETICALLY-NORMAL ACUTE MYELOID LEUKEMIA  
Presenter: YONGKEE CHO, Washington University in St. Louis

P07: CODINGMOTIF: EXACT DETERMINATION OF OVERREPRESENTED NUCLEOTIDE MOTIFS IN CODING SEQUENCES  
Presenter: JEFFREY CHUANG, Boston College

P08: TESTING THE ORTHOLOG CONJECTURE WITH FUNCTIONAL DATA FROM SEVERAL PAIRS OF CLOSELY RELATED ORGANISMS  
Presenter: WYATT CLARK, Indiana University

P09: NOVEL INFORMATION THEORY BASED METRICS FOR THE EVALUATION OF GO TERM ANNOTATIONS  
Presenter: WYATT CLARK, Indiana University

P10: MULTISCALE PATIENT-SPECIFIC BLOOD SYSTEMS BIOLOGY  
Presenter: SCOTT DIAMOND, University of Pennsylvania

P11: SEARCHING FOR POPULATION VARIATION OF SMALL TANDEM SEGMENTAL DUPLICATIONS USING THE 1000 GENOMES PROJECT SEQUENCING DATA  
Presenter: MICHAEL DICKENS, University of Colorado

P12: GENE DUPLICABILITY IN THE MAMMALIAN PROTEIN-PROTEIN INTERACTION NETWORK  
Presenter: AOIFE DOHERTY, National University of Ireland, Maynooth

P13: MIXTURE MODELS VS. SUPERVISED LEARNING METHODS FOR INTEGRATIVE GENOMIC ANALYSIS  
Presenter: DANIEL DVORKIN, University of Colorado, Denver

P14: INFERRING PROTEIN BACKBONE FLEXIBILITY FROM CONFORMATIONAL DIFFERENCES  
Presenter: ELIZABETH ESKOW, University of Colorado Boulder

P15: IMPROVING AUTOMATED PROTEIN FUNCTION PREDICTION BY INTEGRATING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING  
Presenter: CHRISTOPHER FUNK, University of Colorado
P16: A SENSEMAKING MODEL FOR THE EXPLORATIVE ANALYSIS OF LARGE GENE LISTS
Presenter: CARSTEN GÖRG, University of Colorado Denver

P17: MATHEMATICAL MODELING OF METASTATIC GROWTH IN LYMPHATIC VESSELS
Presenter: RUTH GRISWOLD, Mount Sinai School of Medicine

P18: CONSENSUS NETWORK INFERENCE USING MULTIPLE DATASETS TO ELUCIDATE COMMON AND CHEMICAL-SPECIFIC EFFECTS ON STEOROIDOGENESIS
Presenter: TANWIR HABIB, US Army ERDC

P19: A NETWORK-BASED APPROACH TO FINDING FUSION GENES
Presenter: LEANNE HAGGERTY, The National University of Ireland, Maynooth

P20: COMPACT ENCODING FOR GENE THERAPY
Presenter: ROGER HALL, University of Arkansas at Little Rock

Presenter: SINÉAD HAMILTON, The National University of Ireland, Maynooth

P22: A RANDOM FOREST MODEL FOR PHOSPHORYLATION SITE LOCALIZATION USING MASS SPECTROMETRIC DATA
Presenter: XIN HE, Indiana University Bloomington

P23: CHARACTERIZING THE PHENOTYPE OF BETA-BLOCKER RESPONSIVENESS IN CHF PATIENTS
Presenter: MICHAEL HINTERBERG, University of Colorado-Denver

P24: CTCF CAN FORM UNIDIRECTIONAL REGULATORY ELEMENTS BY COMBINING WITH OTHER TRANSCRIPTION FACTORS
Presenter: CHIH-HAO HSU, NIH

P25: METABOLIC NETWORK ANALYSIS OF APICOMPLEXAN PARASITES TO IDENTIFY NOVEL DRUG TARGETS
Presenter: STACY HUNG, University of Toronto

P26: IMPROVEMENT OF FOLD RECOGNITION BY INCORPORATING BAYESIAN FRAMEWORK ON ENCODING PROTEIN FRAGMENTS OF UNKNOWN STRUCTURES FOR STRUCTURE ALPHABET-BASED ALIGNMENT
Presenter: KENNETH HUNG, National Taiwan University

P27: THE ROLE OF P53 IN OLIGODENDROCYTE UPR
Presenter: HASAN JAMIL, Wayne State University

P28: EXTENDING THE VALUE OF SPECTRAL LIBRARIES: A NEIGHBOR-BASED APPROACH TO PREDICTING INTENSITIES OF PEPTIDE FRAGMENTATION SPECTRA
Presenter: CHAO JI, Indiana University, Bloomington

P29: DAVID-WS: A STATEFUL WEB SERVICE TO FACILITATE GENE/PROTEIN LIST ANALYSIS
Presenter: XIAOLI JIAO, NIH/NIAID

P30: MOBYLE: A WEB BASED FRAME-WORK FOR THE INTEGRATION OF BIOINFORMATICS SOFTWARE AND DATABANKS
Presenter: LOUIS JONES, Institut Pasteur

P31: DISCOVERY: A RESOURCE FOR THE RATIONAL SELECTION OF DRUG TARGET PROTEINS AND LEADS FOR THE MALARIA PARASITE, PLASMODIUM FALCIPARUM
Presenter: FOURIE JOUBERT, University of Pretoria
P32: FINDING A SURROGATE METRIC OF CANCER ANTIGEN 125 WITH THOUSAND OF PROTEINS
Presenter: SHINTARO KATO, NEC Corporation

P33: AN INFORMATION-THEORETIC MEASURE OF TFBS MOTIF PALINDROMICITY
Presenter: ALASTAIR KILPATRICK, University of Edinburgh

P34: PATHMAP: DEVELOPMENT OF PATHWAY SIGNATURE SYSTEM FOR IDENTIFYING DRUGGABLE PARTNERS OF SYNTHETIC LETHAL GENES IN CANCER
Presenter: JIHYE KIM, University of Colorado Denver School of Medicine

P35: AUTOMATED REACTION MAPPING FOR ISOMERS
Presenter: TINA KOURI, Colorado School of Mines

P36: A DATA INTEGRATION APPROACH REVEALS SPECIES-SPECIFIC PHARMACOLOGY BETWEEN HUMAN AND RAT ORTHOLOGS
Presenter: FELIX KRUGER, European Bioinformatics Institute

P37: ACCURATE INFERRING TRANSCRIPTION REGULATION FROM A COMPRENDIUM OF EXPRESSION PROFILES
Presenter: XUELING LI, University of Texas

P38: PHENOPRED 2: PREDICTION OF GENE-DISEASE ASSOCIATIONS BASED ON MULTIPLE NETWORKS
Presenter: BIAO LI, The Buck Institute for Research on Aging

P39: EXPLORING AN APPROXIMATE SUBGRAPH MATCHING APPROACH FOR BIOMEDICAL EVENT EXTRACTION
Presenter: HAIBIN LIU, University of Colorado School of Medicine

P40: CANCELLED

P41: A PHYSIOLOGICAL PATHWAY PORTAL TOOL AT THE RAT GENOME DATABASE
Presenter: WEISONG LIU, The Medical College of Wisconsin

P42: EVOLUTION AND ARCHITECTURE OF THE INNER MEMBRANE COMPLEX IN ASEXUAL AND SEXUAL STAGES OF THE MALARIA PARASITE
Presenter: NOELEEN LOUGHRAN, The Hospital for Sick Children

P43: LABEL AND EDGE MISMATCH GRAPHLET KERNELS FOR FUNCTIONAL RESIDUE PREDICTION IN PROTEIN STRUCTURES
Presenter: JOSE LUGO-MARTINEZ, Indiana University

P44: ROLE OF PYRUVATE DEHYDROGENASE KINASES (PDKS) AND THEIR RESPECTIVE MICRO RNAS IN HUMAN OVARIAN CANCER
Presenter: SHAUKAT MALIK, Mohammad Ali Jinnah University

P45: ANALYSIS OF HUMAN COMPLEX DISORDERS FROM THE SYSTEMS BIOLOGY PERSPECTIVE
Presenter: NATALIA MALTSEV

P46: GENESMASH: A RESTFUL WEB SERVICE FOR GENE ANNOTATIONS
Presenter: GANIRAJU MANYAM, University of Texas Anderson Cancer Center

P47: TRAC: TRANSCRIPTOME ANALYSIS WITH CIRCOS
Presenter: ALEKSANDRA MARKOVETS, University of Arkansas at Little Rock
P48: PREDICTING TRANSCRIPTION FACTOR BINDING SITES WITH HIDDEN MARKOV MODELS BY USING CHIP-SEQ DATA
Presenter: ANTHONY MATHELIER, University of British Columbia

P49: A COMPARATIVE ANALYSIS OF KNOWLEDGE-BASED GENE PRIORITIZATION METHODS: IS THE SUPPLY GREATER THAN THE DEMAND?
Presenter: NEEL MEHTA, Arizona State University

P50: HIERARCHICAL PREDICTION OF ENZYME CLASSES USING ENSEMBLE MACHINE LEARNING APPROACHES
Presenter: AKRAM MOHAMMED, University of Nebraska Medical Center

P51: INFERRING THE MAMMALIAN PHYLOGENY USING A HETEROGENEOUS APPROACH
Presenter: CLAIRE MORGAN, Dublin City University

P52: AN ISSUE OF CAUTION IN THE USE OF MITOCHONDRIAL DATA TO INFERENCE THE MAMMALIAN PHYLOGENY
Presenter: CLAIRE MORGAN, Dublin City University

P53: ENTROPY-BASED ASSESSMENT OF CONDITION-SPECIFIC HISTONE MODIFICATIONS
Presenter: TADASU NOZAKI, Keio University

P54: PREDICTION OF PROTEIN FUNCTION USING ANALYSIS OF INTERACTION MAPS ACROSS SPECIES
Presenter: KYMBERLEIGH PAGEL, Indiana University, Bloomington

P55: DECHIFFERING OF HUMAN PROTEIN INTERACTOME USING STRUCTURAL COMPLEXES
Presenter: ANNA PANCHENKO, National Center for Biotechnology Information, Institutes of Health

P56: REGULATION OF PROTEIN-PROTEIN BINDING BY PHOSPHORYLATION AND OLIGOMERIZATION
Presenter: ANNA PANCHENKO, National Center for Biotechnology Information, Institutes of Health

P57: CYSNODB: A DATABASE OF S-NITROSYLATION SITES
Presenter: SANJIT PANDEY, University of Nebraska Medical Center

P58: A DATABASE FRAMEWORK FOR IDENTIFYING CODING SEQUENCE FROM NEXTGEN MRNA SEQUENCES
Presenter: SANJIT PANDEY, University of Nebraska Medical Center

P59: MODPRED: PREDICTING POST-TRANSLATIONAL MODIFICATION SITES FROM PROTEIN SEQUENCE
Presenter: VIKAS PEJAVER, Indiana University

P60: A MULTIPLE-TEMPLATE APPROACH FOR PROTEIN THREADING
Presenter: JIAN PENG, Toyota Technological Institute at Chicago

P61: A TOOL FOR ASSESSMENT AND VALIDATION OF SEQUENCING-BASED METHODS IN FAMILY DISEASE STUDIES
Presenter: ALEXANDER POOLE, University of Colorado Anschutz Medical Campus

P62: ALTERNATIVE SPLICING DETECTION USING RNASEQ: AN ASSEMBLY APPROACH
Presenter: LIKIT PREEYANON, Michigan State University

P63: IN SILICO RATIONAL DRUG DESIGN AND MODELING STUDIES OF NOVEL INHIBITORS FOR MULTI-TARGET INHIBITION IN PSEUDOMONAS AERUGINOSA
Presenter: JAYARAMAN PREMKUMAR, Nanyang Technological University
<table>
<thead>
<tr>
<th>Poster Number</th>
<th>Title</th>
<th>Presenter(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P64</td>
<td>QUANTITATIVE DATA: WHERE ARE THEY HIDDEN IN BIOMEDICAL LITERATURE?</td>
<td>KOMANDUR RAVIKUMAR, University of Colorado School of Medicine</td>
</tr>
<tr>
<td>P65</td>
<td>TRANSCRIPTOME PROFILE OF BOVINE RESPIRATORY DISEASE PATHOGEN — MANNHEIMIA HAEMOLYTICA PHL213</td>
<td>JOSEPH REDDY, Mississippi State University</td>
</tr>
<tr>
<td>P66</td>
<td>SPACERSEEKER: DETECTING AND IDENTIFYING IMMUNITY CONFERRED BY CRISPR LOCI</td>
<td>ZACHARY ROMER, Loyola University Chicago</td>
</tr>
<tr>
<td>P67</td>
<td>MUTDB: A DATABASE FOR ASSESSING THE IMPACT OF GENETIC VARIANTS</td>
<td>TAL RONNEN ORON, Buck Institute for Research on Aging</td>
</tr>
<tr>
<td>P68</td>
<td>SUPPORT FOR LARGER EPISTATIC MODELING USING AN OPENCL IMPLEMENTATION OF MULTIFACTOR DIMENSIONALITY REDUCTION</td>
<td>JAMES RUDD, North Carolina Central University</td>
</tr>
<tr>
<td>P69</td>
<td>IDENTIFYING TREATMENT RELEVANT COLORECTAL CANCER SUBTYPES USING ITERATIVE NON-NEGATIVE MATRIX FACTORIZATION</td>
<td>ANDREAS SCHLICKER, Netherlands Cancer Institute</td>
</tr>
<tr>
<td>P70</td>
<td>A NEW PROBABILISTIC MODEL IN PREDICTIVE MICROBIOLOGY (NPMPM)</td>
<td>NADINE SCOENE, Goethe University Frankfurt</td>
</tr>
<tr>
<td>P71</td>
<td>SIMPSONS PARADOX AND CORRELATION CANCELLATION IN GENOME-WIDE ASSOCIATION STUDY DATA</td>
<td>RONALD SCHUYLER, University of Colorado Denver</td>
</tr>
<tr>
<td>P72</td>
<td>PREDICTING CHROMOSOME TERRITORIES USING SEQUENCE SIMILARITY</td>
<td>JOSIAH SEAMAN, University of Colorado</td>
</tr>
<tr>
<td>P73</td>
<td>FUNCTIONAL PROFILING OF PHARMACOGENETIC NON-SYNONYMOUS SNPS</td>
<td>CHET SELIGMAN, Buck Institute for Research on Aging</td>
</tr>
<tr>
<td>P74</td>
<td>SEMI-SUPERVISED LEARNING IMPROVES GENE EXPRESSION-BASED PREDICTION OF CANCER RECURRENT</td>
<td>MINGGUANG SHI, Vanderbilt University</td>
</tr>
<tr>
<td>P75</td>
<td>COMPREHENSIVE ANALYSIS OF ASSOCIATIONS BETWEEN ALTERNATIVE SPLICING AND HISTONE MODIFICATIONS</td>
<td>YUKI SHINDO, Keio University, Japan</td>
</tr>
<tr>
<td>P76</td>
<td>DEVELOPING A DECISION SUPPORT SYSTEM FOR STRATEGIC SELECTION OF DRUG TARGETS TO COMBAT TROPICAL DISEASES</td>
<td>FLORIANO SILVA-jr, FIOCRUZ</td>
</tr>
<tr>
<td>P77</td>
<td>CONCATABOMINATIONS: A CHARACTER RECODING STRATEGY IMPLEMENTED IN THE SAFE TAXONOMIC REDUCTION APPROACH FOR IDENTIFYING UNSTABLE TAXA</td>
<td>KAREN SIU-TING SALVATIERRA, National University of Ireland Maynooth</td>
</tr>
<tr>
<td>P78</td>
<td>INVESTIGATION INTO CONSTRUCTION OF MHC LIGAND PREDICTORS AND IMPLICATIONS OF PROTOCOLS USED TO EVALUATE THEM</td>
<td>WERNER SMIDT, University of Pretoria</td>
</tr>
<tr>
<td>P79</td>
<td>MINING FOR CLASS-SPECIFIC MOTIFS IN PROTEIN SEQUENCE CLASSIFICATION</td>
<td>SATISH SRINIVASAN, University of Nebraska Medical Center</td>
</tr>
</tbody>
</table>
P80: RELATIONSHIPS BETWEEN POSITIONING OF SYNONYMOUS CODON AND GENE TRANSLATION EFFICIENCY IN BACTERIA
Presenter: SATOSHI TAMAKI, Keio University

P81: THE RAT GENOME CURATION: RGD AUTOMATED DATA INTEGRATION PIPELINES MAXIMIZE COVERAGE
Presenter: MAREK TUTAJ, Medical College of Wisconsin

P82: USING TRANSPOSABLE ELEMENTS TO UNDERSTAND GENOMIC EVOLUTION
Presenter: VIJETHA VEMULAPALLI, University of Colorado Denver

P83: DETERMINING THE MINIMAL FUNCTIONAL ELEMENTS IN PROTEIN FAMILIES
Presenter: SULEYMAN VURAL, University of Nebraska

P84: MACHINE APPROACHES TO RECOGNITION OF TRYPANOSOMAL VARIANT SURFACE GLYCOPROTEIN SEQUENCES
Presenter: JON WILKES, Wellcome Trust Centre for Molecular Parasitology

P85: STOP AND DEFOG: WEB APPLICATIONS FOR A COMPREHENSIVE FUNCTIONAL GENE SET ANALYSIS
Presenter: TOBIAS WITTkop, Buck Institute for Research on Aging

P86: BUILDING AN INTERACTOME TO IDENTIFY SIGNALING COMPONENTS
Presenter: SARAH WYATT, Ohio University

P87: POST-TRANSLATIONAL MODIFICATIONS INDUCE SIGNIFICANT YET NOT EXTREME CHANGES TO PROTEIN STRUCTURE
Presenter: FUXIAO XIN, Indiana University Bloomington

P88: CO-EVOLUTION OF SP1 AND DNA BINDING SITES
Presenter: KEN YOKOYAMA, University of Colorado Denver

P89: COEXPRESSION OF LINKED GENES IS NOT EXPLAINED BY SIMILAR CHROMATIN ENVIRONMENTS
Presenter: YIQIANG ZHAO, Buck Institute

P90: WHAT’S NEW AT ECOCENE
Presenter: JINDAN ZHOU, University of Miami Miller School of Medicine
P01: IDENTIFYING NOVEL COMPONENTS OF THE GRAVITROPIC SIGNAL TRANSDUCTION PATHWAY VIA TRANSCRIPTOME ANALYSIS

Presenter: ZACHARY ABRAMS, Ohio University
Authors: Zachary Abrams, Avery Tucker, Matthew Payne, Kaiyu Shen, Sarah Wyatt

The gravity persistent signal (GPS) treatment was used to isolate the mechanisms of gravitropic signal transduction from those of response. Arabidopsis inflorescence stems were gravistimulated at 4ºC for 1 hr then returned to vertical at room temperature. mRNA was collected at 2min, 4min, 10min and 30min after gravistimulation. These time points were chosen to best represent the early events of gravitropic signal transduction. The mRNA (treatment vs. non-gravistimulated control) was probed against an Agilent Arabidopsis gene expression array (4X44k) using a dual color platform, four replicates per time point. Raw data were first normalized across all sixteen arrays and filtered, an empirical Bayesian model was applied on the data and 337 differentially expressed genes (8 genes at 2min, 96 genes at 4min, 190 genes at 10min and 65 genes at 30min) were found. Genes were annotated using Gene Ontologies, AraCyc pathways, KEGG pathways and Plant Ontologies information and a hypergeometric statistical analysis was performed on these categories to find the most significantly enriched functional groups and pathways: response to exogenous stimulus, transcriptional activities and flavonoid pathways. These functions/pathways were potentially involved in gravitropic signal transduction. The differentially expressed genes were further filtered based on their expression levels and the “top” differentially expressed genes were selected to validate with qRT-PCR. These analyses could help find putative genes involved in gravitropic signal transduction.

P02: DEVELOPING MAXIMUM LIKELIHOOD AND BAYESIAN SUPERTREES

Presenter: WASIU AKANNI, National University of Ireland, Maynooth
Authors: Wasiu Akanni, Davide Pisani, Peter Forster, Mark Wilkinson

Little work has been done on the development of supertree methods in the Likelihood and Bayesian frameworks. Recently, it has been proposed that Maximum Likelihood (ML) supertrees could be developed by using an exponential distribution to model the probability that the input trees could be erroneous. When the tree-to-tree distances used in the ML computation are calculated using the Symmetric Difference, the ML supertree has been shown to be equivalent to a Majority Rule Consensus Supertree, and hence, exactly as the latter, the ML supertree must have the desirable property of being a median tree with reference to the input set. In addition, the ability to estimate the likelihood
of supertrees, will allow the implementation of Bayesian MCMC approaches, which have the advantage of allowing for a natural estimation of the support for the nodes on the recovered supertree. We have developed the first software for the estimation of Maximum Likelihood and Bayesian supertrees. The program is being written in Python and will also exploit the capabilities of already available software, i.e. P4. Here, we present results of reanalyses of the datasets of Holton and Pisani (2010) and Pisani(2007) and present the first Bayesian Genomic Supertrees generated using our new software. In addition, we shall compare these supertrees with those derived using other common supertree methods (e.g. Matrix Representation with Parsimony and Average Consensus). We show that results obtained using our method compared well with those obtained using MRP, and significantly outperforms all other methods.

**P03: OPPORTUNITIES AND CHALLENGES ASSOCIATED WITH COMPUTATIONAL RNASEQ DATA ANALYSIS**

*Presenter: DAVID ASTLING, University of Colorado School of Medicine*

*Authors: David Astling, Tiffany Chan, Jihye Kim and Aik-Choon Tan*

Next-generation sequencing has revolutionized mRNA expression analysis (RNA-seq) by providing the ability venture beyond traditional microarrays, offering greater sensitivity and the ability to obtain a true transcriptome. In addition, RNA-seq allows researchers to discover coding SNPs, novel splice variants and gene fusions. However, because this is a new technology, the computational data analysis tools that have been developed are still in their infancy. Some of the challenges include mapping reads across splice junctions, assigning reads to transcripts when a reference is incomplete or unavailable, and quantification when coverage is poor. While the number of reads and the overall coverage is important, the accuracy in assigning reads to individual transcripts can also have a big impact on gene-level expression estimates. To address these issues we have examined various computational methods for analyzing RNA-seq data through simulations. We will discuss the influence of the number or reads, read length and paired-ends on the gene and isoform expression measures. We have also compared gene expression estimated from microarrays and different next-generation sequencers. Selected genes have been validated by qPCR to determine which computational methods provide the best estimation for expression. Finally, we highlighted some computational strategies and experimental steps to optimize the analysis of RNA-seq in different applications.

**P04: JIKITOU A QUESTION ANSWERING SYSTEM: THE NEXT STAGE IN THE EVOLUTION OF INFORMATION RETRIEVAL**

*Presenter: MICHAEL BAUER, University of Arkansas at Little Rock*

*Authors: Michael A. Bauer, Daniel Berleant, Robert E. Belford, Roger A. Hall*

There is a need for intelligent information retrieval systems that can summarize relevant textual information while also incorporating multiple sources of
information from reliable sources to satisfy a user’s query. Question answering (QA) is a specialized type of information retrieval with the aim of returning precise short answers to queries posed as natural language questions. We have developed a QA system (www.jikitou.com), named Jikitou, which answers natural language questions with sentences taken from Medline abstracts. As a user types a question the system suggest additional terms that can be added to clarify the query question. The HyperGlossary (www.hyperglossary.org) is a literacy tool, that we developed and integrated with Jikitou, which automates the insertion of hyperlinks into a digital text document and connects them to textual definitions, multimedia content, and in the case of molecules, 2D and 3D representations. Users are able to take advantage of authoritative knowledge sources on the Internet to enhance the understanding of concepts which may be beyond their zone of proximal development. Our system addresses two current requirement gaps in biomedical question answering, namely, incorporating multimedia information and an ability to interact with the user. There is a lack of systems that allow the user to establish context, utilize that information in the search process, and automatically return the appropriate answer. The fusion of the HyperGlossary and Jikitou combine to create a unique system to help researchers navigate the current information deluge.

P05: PHYLOGENETIC SIGNAL DISSECTION AND THE PHYLOGENY OF THE ECYDOSOZOA

Presenter: LAHCEN CAMPBELL, The National University of Ireland, Maynooth
Authors: Lahcen Campbell, Omar Rota-Stabelli, Kevin J. Peterson, Davide Pisani

Morphological data traditionally recovers Onychophora (Velvet worms), Tardigrada (Water bears) and Arthropoda (e.g. crabs, bees, spiders, and centipedes) within the monophyletic group Panarthropoda. However, studies of molecular sequence data provide support for alternative placements of Tardigrada within the Ecdysozoa; most often grouping tardigrades closer to the cycloneuralian clade Nematoda (round worms). This result is suspicious however, as the branches associated with both the tardigrades and the nematodes are very long, suggesting the nematode- tardigrade grouping might represent a long-branch attraction (LBA) artefact. In order to test the hypotheses of tardigrades relationships, we have analysed two independent genomic data sets: (1) a large EST data set and, (2) microRNAs for all relevant taxa, including newly sequenced microRNAs for Tardigrada and Onychophora. Using careful experimental manipulations, such as: comparisons of model fit, taxonomic sampling and signal dissection, we were able to show that support for a Nematoda plus Tardigrada group derives from the phylogenetic artefact of long branch attraction (LBA). Our findings demonstrate the power of implementing ‘signal dissection’ analysis on data sets containing strong conflicting phylogenetic signals. Here in the case of the placement of tardigrades, multiple phylogenetic signals support alternative placements and phylogenetic signal dissection allow their investigation. Finally, to further corroborate the validity of our approach, we
used were able to show that analysis of microRNA support the same results we obtained using Phylogenetic signal dissection.

**P06: A NETWORK-BASED SURVIVAL PREDICTOR AND ITS APPLICATION TO CYTOGENETICALLY-NORMAL ACUTE MYELOID LEUKEMIA**

*Presenter: YONGKEE CHO, Washington University in St. Louis*  
*Authors: Yongkee Cho, Rakesh Nagarajan*

Cytogenetically-normal acute myeloid leukemia (CN-AML), which constitutes up to 45% of all AML, is a heterogeneous disease in terms of clinical outcomes and genetic changes. The combinations of several genetic markers have been widely used to predict outcomes. Recent genome-wide experiments have added a number of novel candidate genes of CN-AML; however, it is still unclear how they involve in its prognosis. This suggests the aid of available genome annotations, including protein-protein interaction databases, transcription factor-target gene databases, co-expression information, and literature-based gene networks. The Bayesian networks provide a statistically robust framework to integrate these heterogeneous annotation sources as well as to score all the relationship between genes in the integrated network. We present a novel algorithm to generate sub-networks from input genes and to predict patient’s outcome using the sub-networks. We validate this network-based model by applying it to the CN-AML dataset with genetic and prognostic variance. Next, we compare this method to other methods not considering network information.

**P07: CODINGMOTIF: EXACT DETERMINATION OF OVERREPRESENTED NUCLEOTIDE MOTIFS IN CODING SEQUENCES**

*Presenter: JEFFREY CHUANG, Boston College*  
*Authors: Yang Ding, Andy Lorenz, Jeffrey Chuang*

It has been increasingly appreciated that coding sequences harbor regulatory sequence motifs in addition to encoding for protein. These sequence motifs are expected to be overrepresented in nucleotide sequences bound by a common protein or small RNA. However, detecting overrepresented motifs has been difficult because of interference by constraints at the protein level. Sampling-based approaches to solve this problem based on codon-shuffling have been limited to exploring only an infinitesimal fraction of the sequence space and by their use of parametric approximations. We present a novel O(N log^2 N)-time algorithm, CodingMotif, to identify nucleotide-level motifs of unusual copy number in protein-coding regions. Using a new dynamic programming algorithm we exhaustively calculate the distribution of the number of occurrences of a motif over all coding sequences that encode the same protein sequence, given a background for codon usage and dinucleotide biases. Our method takes advantage of the sparseness of loci where the motif can occur, greatly speeding
up the required convolution calculations. Knowledge of the distribution allows one to assess the exact non-parametric p-value of whether a given motif is over- or under-represented. We demonstrate that CodingMotif identifies known functional motifs more accurately than sampling-based approaches in a variety of datasets, including ChIP-seq data for the transcription factors NRSF and GABP. CodingMotif provides a theoretically and empirically-demonstrated advance for the detection of motifs overrepresented in coding sequences. We expect CodingMotif to be useful for analyzing functions such as DNA-protein binding, RNA-protein binding, or microRNA-RNA binding within coding regions.

**P08: TESTING THE ORTHOLOG CONJECTURE WITH FUNCTIONAL DATA FROM SEVERAL PAIRS OF CLOSELY RELATED ORGANISMS**

**Presenter:** WYATT CLARK, Indiana University  
**Authors:** Wyatt T. Clark, Predrag Radivojac, Matthew Hahn

A common assumption in comparative genomics is that orthologous genes share greater functional similarity than do paralogous genes (the ortholog conjecture). While this assumption is largely untested, many methods and databases developed for the inference of protein function explicitly ignore paralogs. In order to test this assumption we carried out the first large-scale test of the ortholog conjecture using comparative functional genomic data from human and mouse. We used experimentally derived functions of more than 8,900 genes, as well as an independent microarray dataset, to directly assess our ability to predict function using both orthologs and paralogs. Both datasets show that paralogs are often a much better predictor of function than are orthologs, even at lower sequence identities. Among paralogs, those found within the same species are consistently more functionally similar than those found in a different species. Our results, which were published in PlosCB, and featured in Nature Genetics, have implications for the computational prediction of protein function, and shed light on the relationship between sequence divergence and functional divergence. Here we extend our study of the ortholog conjecture to other pairs of closely related organisms in order to determine whether the same patterns hold across a wider range of organisms.

**P09: NOVEL INFORMATION THEORY BASED METRICS FOR THE EVALUATION OF GO TERMANNOTATIONS**

**Presenter:** WYATT CLARK, Indiana University  
**Authors:** Wyatt T. Clark, Predrag Radivojac

The non-uniform topological structure of the Gene Ontology compounds the already difficult task of performing an evaluation of any set of assumed or predicted gene annotations. Several factors contribute to this difficulty; including large discrepancies in the prior probability of terms at the same level in the ontology, and successive terms that are semantically distinct yet refer to the same set of genes. Several attempts have been made to deal with these
problems by coming up with new metrics for evaluating annotations. Semantic similarity based measures attempt to account for these problems by only considering the information content of the minimum subsumer of a true and assumed term. While semantic similarity is successful in accounting for bias in the ontology due to uneven stratification of terms, due to the fact that it is only one value it is impossible to impart information about a predicted annotation that is analogous to the information retrieval concepts of precision, recall and specificity. We have developed novel information theory based metrics for evaluating a set of GO annotations based on Kullback-Leibler divergence. Our metrics incorporate the advantages of semantic similarity but also have the added benefit of being interpretable as the information theoretic analogs of precision, recall and specificity. We compare and contrast our novel metric with some of the published metrics for evaluating GO annotations, pointing out advantages and disadvantages of each. Finally, we give a critique of evaluating a given metric based on its correlation with pairwise sequence

P10: MULTISCALE PATIENT-SPECIFIC BLOOD SYSTEMS BIOLOGY
Presenter and Author: SCOTT L. DIAMOND, University of Pennsylvania

Predicting tissue function based upon an individual’s unique cells requires a multiscale Systems Biology approach to understand the coupling of intracellular signaling with spatiotemporal gradients of extracellular biochemicals controlled by convective-diffusive transport. During thrombotic or hemostatic episodes, platelets bind collagen and release ADP and thromboxane A2 (TXA2) to facilitate the recruitment of additional platelets to a growing deposit that distorts the flow field. Calcium dye-loaded platelets in PPACK and indomethacin-treated plasma (thus lacking thrombin and TXA2) from 3 healthy donors were subjected to Pairwise Agonist Scanning where platelets were exposed to all pairwise combinations of ADP, U46619, and convulxin (at 0, 0.1, 1, 10 x EC50) to activate P2Y1/P2Y12, TP, and GPVI receptors, respectively, in the presence or absence of the IP receptor agonist, iloprost. With 74 calcium responses to train a neural network (NN) model of platelet calcium mobilization for each donor, each NN model was then embedded into a multiscale Monte Carlo/finite element simulation of donor-specific platelet deposition under flow. For each donor, simulations predicted the measured platelet deposition dynamics and ranked drug sensitivity for PPACK-treated whole blood flowing over collagen at 200 s-1 wall shear rate in the presence of indomethacin, aspirin, MRS-2179 (P2Y1 inhibitor), or iloprost. Consistent with measurement and simulation, one donor displayed larger clots, while another donor presented a indomethacin-resistance and U46619-insensitivity (revealing a novel heterozygote mutation). In silico representations of an individual’s platelet phenotype allowed prediction of blood function under flow, essential to identifying patient-specific cardiovascular risks, drug responses, and novel genotypes.
Small tandem segmental duplications in the human genome may facilitate genetic variation through homologous recombination events that may be expressed as phenotypic variation or result in certain genomic diseases. High throughput sequencing technologies and accurate assembly of repetitive regions would allow efficient measurements of these tandemly duplicated segments. However, Next-Gen sequencing produces millions of short reads that make genomic assembly of these highly duplicated regions challenging. To address this problem, mapping depth was recorded using sequencing data from the 1000 genomes project aligned to selected regions of the human reference assembly to generate population profiles of four major population groups for assessing copy variation in small (200 bases) tandemly duplicated coding regions. Sequence files of individuals from four populations in the Pilot 1 and Pilot 2 projects with low and high coverage sequencing were analyzed. Unique 200 base regions from sequences of ultraconserved elements in addition to unique multi-copy 200 base coding regions were used to normalize sequencing coverage and estimate copy number across individuals. The population profiles from the 1000 genomes project can provide an understanding of variation within the general population in addition to serving as a reference for investigating small tandem segmental duplications in disease groups as their sequencing data becomes available.

While gene duplication events often occur, only a small proportion of gene duplicates become fixed. These gene duplicates are a major source of raw genetic material and may lead to the evolution of more complex molecular pathways. Genes and their encoding proteins generally operate as networks of interacting molecules. Examination of protein-protein interaction networks (PINs) has previously shown that duplicability is linked to the position of genes in the network. Here, we seek to determine if there is an explicit link between the duplicability of genes across the primate phylogeny and the position of the encoded proteins in the PIN. We found that, in general, duplicated genes tend to occupy the most central positions in the network in agreement with previous observations in the human genome. However, genes that duplicated in certain branches of the primate phylogeny occupy less central positions. Additionally, we show that genes duplicated in a given branch tend to interact with each
other, and that genes encoding interacting proteins have similar phylogenetic tree topologies. These observations uncover previously unrecognized ways in which the structure of the mammalian PIN is related to the duplicability of its components.

**P13: MIXTURE MODELS VS. SUPERVISED LEARNING METHODS FOR INTEGRATIVE GENOMIC ANALYSIS**

*Presenter: DANIEL DVORKIN, University of Colorado, Denver*  
*Authors: Daniel Dvorkin, Katerina Kechris*

Gene classification using multiple data sources such as expression and transcription factor binding is an important problem in bioinformatics. Given high-quality training data, supervised machine learning methods such as logistic regression, naive Bayes classifiers, and random forests are generally more powerful than unsupervised methods. However, when training data sets are small or misspecified, unsupervised and semi-supervised mixture models are more robust. Semi-supervised mixture models also perform well compared to fully supervised methods with good training data, and unlike fully supervised methods, can accept incompletely specified training data. We present here a comparison of unsupervised and semi-supervised mixture models to fully supervised machine learning methods, and discuss the circumstances under which the former may be expected to outperform the latter, using identification of critical genes in Drosophila development as an example of a case in which training data may not be of sufficient size, quality, or completeness to justify the use of fully supervised methods.

**P14: INFERRING PROTEIN BACKBONE FLEXIBILITY FROM CONFORMATIONAL DIFFERENCES**

*Presenter: ELIZABETH ESKOW, University of Colorado Boulder*  
*Authors: Elizabeth Eskow, Debra Goldberg, Deanne Sammond*

The dynamic nature of a protein can be inferred by comparisons of the conformational differences between pairs of independent crystal structures of the same protein arising from differences in crystallization environments or binding partners. Flexible regions in the protein are apparent as positional fluctuations between the atoms of the backbone or side-chains of the conformations, and our hypothesis is that the amplitude of these differences reflects the degree of flexibility. Methods for discerning the overall similarity between proteins with different conformational structures abound, and are used for structure prediction, fold classification or database searching for structural homology. However, we are investigating protein backbone flexibility at the residue and secondary structure element level, and discuss the development of a continuous measure of flexibility calculated by comparing corresponding residues and secondary structure elements of pairs of protein conformations. In order to gain insight into which regions of protein structures are flexible we focus on the
relative sizes of structural changes to the backbone by considering metrics such as angle changes, intra-molecular distance differences and distances calculated by superimposing the structures or corresponding parts of the structures. We compare our measurements to descriptions of conformational changes in the relevant literature and to flexibility measurements of appropriate NMR structures.

P15: Improving Automated Protein Function Prediction by Integrating Natural Language Processing and Machine Learning

Presenter: CHRISTOPHER FUNK, University of Colorado
Authors: Christopher Funk, Artem Sokolov, Asa Ben-Hur, Karin Verspoor

Finding the function of unknown proteins is an important area of research in molecular biology. Due to the large amounts of biological data available (sequence, protein-protein interaction, expression, experimental, eg.), many computational approaches have been developed for automatic function prediction. The biomedical literature is another source of data that is not as widely used but provides a wealth of information that is not available in the public databases. This work utilizes a methodology, GOstruct, because it is a multi-view machine learning framework that is able to combine many different types of information. We obtained 500k abstracts from pubmed corresponding with a list of 8k proteins and the mesh term mouse and performed natural language processing to extract two different types of co-occurrence: protein-protein and protein-gene ontology term. Both types of co-occurrences were calculated as counts within a certain span; we examined spans of sentence and document level. Protein-protein co-occurrences act as an approximation to protein-protein interactions and suggest similar function while protein-gene ontology terms act as an approximation of function. These counts were the input features for GOstruct and a dot product kernel was produced. Compared to a baseline of only using protein-protein interactions from databases and sequence similarity (BLAST) we found that integrating literature co-occurrence kernels improved the prediction in a mouse dataset. (0.773 vs 0.815 AUC). Since literature provided useful information we will now begin to examine more sophisticated ways of incorporating it into GOstruct.

P16: A SenseMaking Model for the Explorative Analysis of Large Gene Lists

Presenter: CARSTEN GÖRG, University of Colorado Denver
Authors: Carsten Görg, Barbara Mirel, Hannah Tipney

The cognitive science and visual analytics community have developed well-supported models of the stages analysts go through over the course of an investigation. These models have proven useful for understanding the analyst’s cognitive tasks and deriving design guidelines to develop user-centered tools that can facilitate and enhance these tasks. Specific models for the analysis of
gene lists do not yet exist, particularly for the analysis of genotype relationships to generate hypotheses about mechanisms of phenotype-level diseases. To better understand and model these analytical processes we have conducted an in-depth case study of a biomedical researcher. The researcher’s goal was to uncover genome-level interactions and events that might explain why some heart failure patients did not respond to beta-blocker treatment while others did. Over a six-month period we observed the researcher’s real-time interactions and think aloud protocols with visualization tools using screen and audio capture. The visualization tools included Cytoscape, Hanalyzer, String, Reactome, and tag clouds; the researcher also accessed the Genetic Association Database, the Pharmacogenomics Knowledge Base, and the biomedical literature. Additionally, we interviewed the researcher monthly and obtained a copy of the laboratory notebook and numerous annotated articles. We then performed a cognitive task analysis taking the interplay of the different tools into account and developed a preliminary sensemaking model for the explorative analysis of molecular interactions. Our preliminary model adapts and refines Pirolli and Card’s well-known sensemaking model for intelligence analysis. We present design guidelines for visual analytics tools that we derived from our model.

**P17: MATHEMATICAL MODELING OF METASTATIC GROWTH IN LYMPHATIC VESSELS**

*Presenter: RUTH GRISWOLD, Mount Sinai School of Medicine  
Authors: Ruth Griswold, Simona Podgrabinska, Charles Peskin, Mihaela Skobe*

Mathematical modeling of tumor growth is important for designing and testing cancer treatments. The major cause of death from cancer is metastasis, yet mathematical models of metastasis predominantly examine the probability of metastatic dissemination, and do not investigate growth of metastases in their unique microenvironment. Metastases in the lung can form in the lymphatic vessels or in the tissue parenchyma. Our recent data from a mouse model showed that metastases in the lymphatics grew more rapidly than metastases in the tissue parenchyma, in agreement with the clinical data showing rapid progression of metastatic disease when lung lymphatics are involved. To explain the difference in growth kinetics between metastases in the lymphatics and in the tissue parenchyma, we are developing a model of intralymphatic tumor growth which includes the following distinct characteristics of the lymphatic niche: (i) architecture of the lymphatic vasculature; (ii) density of metastatic cells within lymphatics; and (iii) oxygen and lactic acid distribution. Findings from a two dimensional model suggest that metastases in the lymphatic vessels will not become severely hypoxic, and therefore will not require angiogenesis for growth. We will employ a three-dimensional model to examine how different biophysical characteristics of the lymphatic microenvironment contribute to the rapid growth of metastases. This model is based on deterministic differential equations commonly used to describe avascular tumor growth. The equations and parameters from this model will be adapted to reflect the lymphatic environment.
P18: CONSENSUS NETWORK INFEERENCE USING MULTIPLE DATASETS TO ELUCIDATE COMMON AND CHEMICAL-SPECIFIC EFFECTS ON STEROIDOGENESIS

Presenter: TANWIR HABIB, US Army ERDC
Authors: Tanwir Habib, Edward J Perkins, Lynn Escalon, Daniel Villeneuve, Gerald Ankley, Natalia Garcia-Reyero

Microarrays are being increasingly used in the field of environmental toxicology, with the potential to be applied to ecological risk assessment. Our research aims to explore the use of regulatory network with multiple microarray datasets to build a robust consensus model for inferring gene regulatory interactions, thereby defining the characteristics that describe the overall system. In this study, we used network modeling on microarray data generated from ovaries of fathead minnow (Pimephales promelas) exposed to several chemicals that affect steroidogenesis at different steps. We then inferred gene regulatory networks using CLR and ARACNE. We analyzed the data following two different approaches. First, a “combined” network was reconstructed using all the microarray data, a total of 1,472 samples from 23 different exposures combined together. In the second approach, “individual” networks were reconstructed from a subset of seven exposures. Clusters of highly interacting genes were identified using MCODE and MINE, and interpreted using Ingenuity Pathway Analysis and DAVID database. Common interactions between the combined and individual networks could offer insights about the common mechanism of the effects of chemicals such as stress and inflammation response, whereas interactions specific for “individual” networks could provide information on the specific chemical mode of action such as disrupting estrogen signaling in the body.

P19: A NETWORK-BASED APPROACH TO FINDING FUSION GENES

Presenter: LEANNE HAGGERTY, The National University of Ireland, Maynooth
Authors: Leanne Haggerty, James. O. McInerney

A fusion gene is the result of an event whereby two previously separate genes are joined to encode a single, usually multifunctional, protein. With the accumulation of genome sequence data, it has become obvious that domain, gene and indeed perhaps genome fusion is frequent in nature, though it is still not clear how often fusion events occur and what the main drivers of fusion are likely to be. Current fusion detection algorithms rely on non-overlapping, side-by-side BLAST matches of two genes from a reference genome to a single open reading frame (ORF) in a target genome. This approach is cumbersome and difficult to implement on a large-scale. However, these data can be more efficiently represented and explored using network-based models of homology. We can construct a network from the pattern of hits from exhaustive database searches. Each node on the network is an identified gene and each edge represents a homology statement. The principal theory behind our algorithm is
that a fusion gene is made up of two individual genes (components) and that a group of closely related homologs on a graph will form a maximal clique. The fusion gene node therefore should be connected to all maximal cliques formed by the individual components and their homologs. By this logic a fusion gene will always lie on the path between its components and this path should be of length 3. We implement this logic to enrich a dataset in fusion genes.

P20: COMPACT ENCODING FOR GENE THERAPY

Presenter and Author: ROGER HALL, University of Arkansas at Little Rock

Synthetic biology is the practice of creating non-evolved organisms through direct manipulation of DNA and other cellular contents. The practice is generally applied to yeast and bacteria, and differs from transgenic or chimeric research. Cellular therapies using synthetic viral vectors have experienced at least one fatality, and continuing research is hampered by the concerns of further Hippocratic dilemmas. The AAV (adeno-associated virus) currently being used is limited to a package capacity of approximately 5000 bases. The Center for Gene Therapy at the University of Iowa School of Medicine reported using dual-vector approaches to deliver genetic information beyond the limit. Current therapies require manipulation of every virus to remove the existing DNA and insert the custom DNA. The vectors are considered safe since they have no viral DNA and are unable to replicate. Overlapping DNA may require higher GC content to accommodate the distinct functional domains, drawing greater mutational pressure, and may create live but fragile vectors with critically dense genomes and calculable generational iterations, enhancing vector safety and reducing cost of therapy. Our approach considers the limits of information encoding in overlapping reading frames using representative protein motifs of both replication and therapeutically significant genes.


Presenter: SINÉAD HAMILTON, The National University of Ireland, Maynooth
Authors: Sinéad Hamilton, Roberto Feuda, Stuart Longhorn, Davide Pisani, James McInerney

Opsins are the photoreceptor proteins for visual perception across all of the metazoa. They arose very early in animal evolution and diversified during the Cambrian, thereby allowing animals to detect various wavelengths of light. We wished to analyse the timing of the duplication events in this protein family as it is these events that have resulted in animals perceiving different wavelengths of light. We show that the visual opsins, for both protostomes and deuterostomes, show a pattern of duplication that correlates with the penetration ability of different wavelengths of light through water. This suggests that the duplications did not arise to confer a selective advantage to the organism with colour vision,
but rather to allow it to perceive the wavelength of light that was most available at certain oceanic depths. The vertebrate visual system has two separate but homologous activation pathways. A colour, bright light pathway present in cone cells and a dim light pathway present in rod cells. The colour vision pathway is the more ancestral type with the dim light pathway emerging as a result of duplication followed by mutations in each of the cone pathway proteins. It has previously been assumed that the rod pathway arose by co-duplication rather than co-option, but using phylogenetic tree building and molecular dating methods we can show that although most of the pathway shows a co-duplication pattern, the G-protein(transducin) appears to have been co-opted from a previous function due to a much older date for this duplication event.

P22: A RANDOM FOREST MODEL FOR PHOSPHORYLATION SITE LOCALIZATION USING MASS SPECTROMETRIC DATA

Presenter: XIN HE, Indiana University Bloomington
Authors: Xin He, Randy J. Arnold, Haixu Tang, Predrag Radivojac

In phosphoproteomics analysis, methods for phosphopeptide identification by LC-MS/MS have been well developed; but the identification of phosphorylation sites has been addressed far less. The power of site localization is limited by the quality of peptide fragmentation by collision-induced dissociation (CID). To locate the phosphorylation site among multiple putative sites (S, T, Y) within a peptide, we introduce an automated scoring mechanism. Our approach treats each putative site within a peptide equally, and feature extraction is performed on each candidate separately. A random forest classifier is trained and evaluated on each dataset using cross validation analysis. For each peptide, the difference between prediction scores from the top two candidates is assigned to the best site. The difference is denoted as the Xscore. In our method we predict the mass spectrum of the corresponding unmodified peptide by PeptideART. We present evidence that the addition of correlation coefficients between annotated experimental and unmodified spectrum as a feature moderately increases sensitivity. Additional features are also extracted based on the peak intensities of the MS/MS spectra. We evaluated our method on several datasets and compared the performance with Ascore, which is a probability based score that measures the correct phosphorylation site localization. We found that our method outperforms Ascore.

P23: CHARACTERIZING THE PHENOTYPE OF BETA-BLOCKER RESPONSIVENESS IN CHF PATIENTS

Presenter and Author: MICHAEL HINTERBERG, University of Colorado-Denver

Congestive heart failure (CHF) is a debilitating and costly disease, affecting more than 5 million people in the United States, and responsible for 3.4 million hospital visits annually. Beta-blocker therapy has been shown to be effective for some patients but not for others, and the complex presentation and progression
of the disease makes it difficult to predict which patients will be responsive to therapy. Efficacy of beta-blocker treatment on an individual patient is not evident for months, in which case more invasive surgical options may be suggested. Therefore, predicting which patients are likely to be responsive to beta-blocker therapy is important for ensuring the most appropriate treatment option as soon as possible. In order to characterize the phenotype of responsive patients, we obtained various sources of patient data from electronic medical records, including demographic data, lab tests, echocardiogram measurements and coded observations, nuclear medicine, and catheterization, in a cohort of 45 patients with congestive heart failure, across two time points. We developed a script to combine these data sources, and compared several linear and non-linear models for predicting outcomes, using techniques such as principal components analysis and random forests. Different sets of explanatory phenotype features were found between the linear and non-linear models. Additionally, we compared human-readable diagnosis models developed using J48 trees, which may be helpful in visualization and hypothesis generation, but the predictive capability of a model using this technique was not replicated in a separate cohort.

**P24: CTCF CAN FORM UNIDIRECTIONAL REGULATORY ELEMENTS BY COMBINING WITH OTHER TRANSCRIPTION FACTORS**

*Presenter: CHIH-HAO HSU, NIH*
*Authors: Chih-Hao Hsu, Ivan Ovcharenko*

Even though most regulatory elements are independent of orientation and distance from the transcription start site (TSS) of genes, a complex consisting of a closely spaced CTCF binding site and enhancer could potentially act as a unidirectional enhancer if bound CTCF will be strong enough to block the activity of the enhancer in the direction of the CTCF site. In this study, we found that histone modifications distributed asymmetrically around CTCF binding sites and this kind of asymmetric profiles of histone marks around CTCF binding sites are associated with differential gene expression of flanking genes. Furthermore, a small portion of CTCF binding sites with asymmetric profiles of histone marks (~ 200 CTCF binding sites) include an extra nucleosome depleted region near the binding sites of CTCF suggesting the binding of other transcription factors in these depleted regions. We performed motif enrichment analysis and found that many activators distributed closely to the binding sites of CTCF and form complexes that function as unidirectional enhancers. Genome-wide analysis suggested that these complexes are tissue-specific and widely-distributed in the human genome. All of these studies indicate the existences of unidirectional enhancers and provide new insight into the activity of regulatory elements along the genome.
P25: METABOLIC NETWORK ANALYSIS OF APICOMPLEXAN PARASITES TO IDENTIFY NOVEL DRUG TARGETS

Presenter: STACY HUNG, University of Toronto
Authors: Stacy Hung, James Wasmuth, Michael E. Grigg, John Parkinson

The Apicomplexa is a large phylum of intracellular obligate parasites, many which are a significant burden on human health and economics. To aid drug development programs, global sequencing initiatives are generating increasing numbers of apicomplexan genomes. By exploiting these resources, we have generated a systematic pipeline for reconstructing the metabolic networks for the Apicomplexa. Previous investigations of the enzyme landscape indicated annotation is poor due to the paucity in biochemical data for enzymes and lack of discriminatory power inherent with homology-based methods for enzyme prediction. To overcome these limitations, we developed DETECT, a probabilistic framework for improved enzyme prediction accounting for sequence diversity present across enzyme families. Given disparity among datasets, DETECT was combined with complementary automated tools and curated resources to allow for a more accurate and comprehensive metabolic reconstruction. Mapping these datasets to the global metabolic network has revealed surprising differences entailing further investigation. For instance, species coverage overlaid onto the apicomplexan network highlights a set of highly conserved “core” enzymes suggesting these parasites have evolved different strategies for performing the same core metabolic activities. Moreover, a pathway heatmap representation highlights conservation patterns of pathways across the phylum, including the pantothenate biosynthesis pathway, in which biochemical assays have suggested the viability of the pathway for drug targeting. These and future findings will provide insight into the metabolic adaptations of apicomplexans and with the availability of a high-quality metabolic reconstruction, detailed physiological, biochemical and genetic information can be incorporated to create a predictive model that generates real, testable hypotheses.

P26: IMPROVEMENT OF FOLD RECOGNITION BY INCORPORATING BAYESIAN FRAMEWORK ON ENCODING PROTEIN FRAGMENTS OF UNKNOWN STRUCTURES FOR STRUCTURE ALPHABET-BASED ALIGNMENT

Presenter: KENNETH HUNG, National Taiwan University
Authors: Kenneth Hung, Cheng-Long Chuang, Kun-Nan Tsai, Chung-Ming Chen

Fold recognition is a popular protein structure prediction approach relying on a good-quality alignment of the target and the template. It becomes very challenging when the sequence identity is low between target and template proteins. The key to the success of template-based methods lies in the proper incorporation of sequence, physiochemical and structural information. We
introduced a new idea featuring the Bayesian framework on encoding protein fragments of unknown structure in structure alphabets to achieve better fold recognition alignment. The structural alphabet-based alignment has been previously developed on incorporating the local structure features for improving structural alignment quality and computational efficiency in our group. The framework of the proposed algorithm performs a fold recognition alignment based on the one-dimensional alphabet code sequence containing information of local structural features of target and template protein sequence. To test the alignment accuracy, the SALIGN benchmark was selected as our test set. The selected structure pairs share an average 20% sequence identity and 65% of structurally equivalent Cα atoms superposed with an rmsd of 3.5 Å. The Lindahl benchmark, which was designed to assess the fold recognition sensitivity, was also chosen as the test set. It contains 976 proteins with 555, 434, and 321 pairs of proteins in the same family, superfamily and fold, respectively. The preliminary results showed that the proposed algorithm successfully demonstrated its ability to enhance the alignment quality, computational efficiency and recognition accuracy.

P27: THE ROLE OF P53 IN OLIGODENDROCYTE UPR

Presenter and Author: HASAN JAMIL, Wayne State University

It has been hypothesized that mutant Plp1 genes in oligodendrocyte are responsible for protein misfolding and retention at the endoplasmic reticulum (ER) leading to demyelination via unfolded protein response (UPR). Neurodegeneration due to demyelination has been linked to diseases such as multiple sclerosis, leukodystrophies, Alzheimer’s and Parkinson’s. It is known that PERK plays a significant role in translation attenuation and cell cycle arrest during UPR activation. Recent studies in our laboratory on rsh and msd mutant mice suggest that the crosstalk between AKT and PERK pathways through p53 may be a key in deciphering the mechanism of oligodendrocyte cell deaths. This possibility increases because p53 is known to be involved in cell cycle progression check point, and the interaction between ATF4 and p53 via ATF3 in PERK pathway and in the AKT pathway. In this study, our goal is to develop a computational approach to study the role of p53 and investigate its contributions in oligodendrocyte UPR signalling cascade. We develop an integrated system to explore gene expression data sets to hypothesize about possible gene regulatory networks for p53 in rsh and msd mutant mice, and cross validate the networks with drosophila genome wide protein and transcription factor interactions toward filtering and sanitizing the hypothesized networks. We plan to determine the contributions of p53 by iterative refinement of the PERK and AKT pathways by way of introducing more missing links.
**P28: Extending the Value of Spectral Libraries: A Neighbor-Based Approach to Predicting Intensities of Peptide Fragmentation Spectra**

Presenter: CHAO JI, Indiana University, Bloomington  
Authors: Chao Ji, Randy J. Arnold, Haixu Tang, Predrag Radivojac

Peptide fragmentation is a complex process that is determined by many factors such as the peptides’ physico-chemical properties, the charge state of the precursor ion, and is essentially a stochastic process. Despite the complex nature of peptide fragmentation, the MS/MS spectrum of a peptide generally displays a reproducible pattern, which allows using data-driven approaches for predicting MS/MS spectra. Although the existing approaches have achieved useful accuracy, most of them utilize only features extracted from peptide sequences. Given the observation that some peptides with same length and charge state have highly correlated intensities of b/y ion series despite the difference in amino acid sequence, we gave a similarity measure of fragmentation spectrum of pair of such peptides, and developed a kernel-based approach that can predict spectrum similarity based on only peptide sequences with reasonably good accuracy. Further, we developed a neighbor-based approach for predicting the peak intensities of a target peptide using the spectra of peptides that are predicted to be similar by the proposed model. In addition, our approach provides confidence score for each predicted spectrum that reflects the prediction quality. By restricting prediction to those peptides with high confidence scores, the coverage of existing spectral libraries can be extended significantly.

**P29: DAVID-WS: A Stateful Web Service to Facilitate Gene/Protein List Analysis**

Presenter: XIAMLI JIAO, NIH/NIAID  
Authors: Xiaoli Jiao, Brad T. Sherman, Da Wei Huang, Richard A. Lempicki

DAVID (the Database for Annotation, Visualization and Integrated Discovery) is a free web-based online bioinformatics resource that aims to provide tools for functional interpretation of large gene/protein lists. It has been used by researchers from more than 5,000 institutes world-wide with a daily submission rate of over 1,200 lists with over 3,500 citations since its first release in 2003. However, the usage of DAVID has been limited to the current web interface which does not support programmatic access and a URL-based API which is limited to defaults due to its size limit and stateless nature. DAVID-WS is developed to provide stateful web services for users to interact with DAVID programmatically and allows users to change background populations, reset functional parameters, and select species and categories for analysis, as well as to provide the ability to query all tools within the same session and format output as desired. Our performance testing shows that it took about 6 to 9 seconds to generate the output for computationally intensive client tasks such
Performing bioinformatics analyses requires the selection and combination of tools and data to answer a given scientific question. Many bioinformatics applications are command-line only and researchers are often hesitant to use them based on installation issues and complex command requirements. Mobyle is a framework and web portal specifically aimed at the integration of bioinformatics software and databanks. It allows to run bioanalyses through a web interface without installing anything locally. In addition to command-line tools, the latest release of Mobyle offers the possibility to execute predefined workflows. The data model has been extended to define a workflow as a dataflow-based coordination of programs that run successive and/or parallel tasks to perform an analysis. Similarly to programs, workflows are viewed as services, sharing most of their description with programs, with the exception of the execution, which consists of a coordination of subtasks rather than the generation and execution of a command line. Furthermore, we now offer a solution to overcome the limitations of results pre-visualization in the portal, which is not adapted to potentially large and/or complex text files. Viewers are a way to embed type-dependant visualization components for the data displayed in the Mobyle Portal. As opposed to programs and workflows, viewers are not executed on the server side, but rather rely entirely on browser-embedded code. The Mobyle package is open-source and freely available at: http://projets.pasteur.fr/wiki/mobyle/download. More information about this project is available at http://projets.pasteur.fr/projects/show/mobyle.

P31: DISCOVERY: A RESOURCE FOR THE RATIONAL SELECTION OF DRUG TARGET PROTEINS AND LEADS FOR THE MALARIA PARASITE, PLASMODIUM FALCIPARUM

Presenter: FOURIE JOUBERT, University of Pretoria
Authors: Jeanre Smit, Phele Mpangase, Michal Szolkiewicz, Misha le Grange, Fourie Joubert

Few rational approaches have been successfully followed in the selection of promising drug target proteins in the malaria parasite. The emergence of widespread drug resistance, even against current drugs is making the effective
selection of new drug targets together with lead compounds essential and urgent, requiring optimal approaches to be put in place for this process. The Discovery project is aimed at providing a publicly available informatics resource where comprehensive information on the parasite and host proteins are stored, together with the results from relevant 3rd-party investigations as well as results from our own high-throughput analysis. The comprehensive data included in the resource is aimed as wide as possible, including protein, gene-ontology, orthology, metabolic, structural, expression and chemoinformatics information. This is combined with a data-mining interface for researchers to perform the selection of putative drug target protein and lead compounds according to their specific highly-flexible criteria. Protein information includes data from the human, mosquito and the various malaria genome projects. Chemical information is from ChEMBL, DrugBank and PDB. Information includes basic annotations, motifs, domains, binding sites, structural features, ontology information, ontology terms, protein-ligand interactions and comparative genomics information. Chemical information includes protein interactions and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties. Searches may be initiated either from the protein or chemical compound as starting point. The system is available at: http://malport.bi.up.ac.za.

**P32: FINDING A SURROGATE METRIC OF CANCER ANTIGEN 125 WITH THOUSAND OF PROTEINS**

*Presenter: SHINTARO KATO, NEC Corporation*
*Authors: Shintaro Kato, Alex Stewart, Rachel M. Ostroff, Iwao Waga, Stephen Williams*

Cancer Antigen 125 (CA125) is a well known biomarker for several malignant cancers and benign tumors, especially ovarian cancer. We can currently measure more than one thousand proteins from human blood using SOMAscan, but, unfortunately, not CA125. In this project, we attempt to identify the surrogate metric of CA125 with a combination of the proteins which have been measured in the current assay. We have evaluated the performance of our surrogate metric to determine if we can replace the direct measurement of CA125. In this work we are not developing a cancer classification algorithm, rather we are trying to determine to what extent the information in the CA125 signal is contained within the high dimensional proteomics data measured by SOMAscan. In order to derive a multidimensional proteomic surrogate for CA125, Partial Least Squares (PLS) regression was applied to Ovarian Cancer data, this dataset contains samples from individuals with benign tumors and also individuals with cancer determined at biopsy (stage I to III). As a result, we found a multidimensional combination of protein signals which mimic the signal from CA125. The SOMAscan assay results projected onto this proteomic vector are significantly correlated with corresponding CA125 Elisa results.
P33: AN INFORMATION-THEORETIC MEASURE OF TFBS MOTIF PALINDROMICITY

Presenter: ALASTAIR KILPATRICK, University of Edinburgh
Authors: Alastair Kilpatrick, Stuart Aitken, Bruce Ward

In this work, we propose a novel measure of transcription factor binding site (TFBS) palindromicity based on information theory and show that our measure scores well-conserved palindromic motifs more highly, hence aiding their discovery. Discovery of TFBS motifs is an important task in the wider challenge of understanding the mechanisms of gene expression. Experimental work has shown that TFBS motifs are often (quasi-)palindromic (i.e. the inverse complement is the same as the original sequence): a measure of motif palindromicity may therefore be helpful in improving TFBS motif discovery. While measuring the palindromicity of a given DNA sequence is trivial, no such measures are currently available for position weight matrices (PWMs), the probabilistic models of motifs used in motif discovery algorithms. We propose a novel measure of motif palindromicity based on two factors: firstly, the information-theoretic similarity of a given PWM to its inverse complement and secondly, the average information content per position of the PWM. Together these factors score more palindromic motifs more highly and also filter out less well-conserved or ‘uninformative’ motifs. We have implemented this measure and are incorporating it within an improved TFBS motif discovery algorithm. Besides this application, our measure may also be an indicator of the structure (and therefore to some extent the function) of a transcription factor protein.

P34: PATHCMAP: DEVELOPMENT OF PATHWAY SIGNATURE SYSTEM FOR IDENTIFYING DRUGGABLE PARTNERS OF SYNTHETIC LETHAL GENES IN CANCER

Presenter: JIHYE KIM, University of Colorado Denver School of Medicine
Authors: Jihye Kim, Carlos H C Cano, Aik Choon Tan

While targeted therapies have shown clinical promise in treating solid tumors that “addicted” to oncogenic pathways, these therapies are rarely curative for advanced cancers. As most cancers are typically diagnosed at advanced stages and acquired resistance mechanisms that can protect the cells from these targeted therapies. Therefore, the discovery of pathways that mediate these compensatory survival mechanisms could reveal novel therapeutic targets for cancer cells “addicted” to these pathways. High throughput methods to identify small molecules that target synthetic lethal genes and pathways are critical for facilitating rational combination with targeted therapies of interest, and may have translational potential in clinical cancer trials. Here, we developed a novel bioinformatics tool “Pathway Connectivity Map (PathCMap)” that systematically connects the most similar pathway expression profile from a reference profiles database and extrapolates the most effective drug for individual query pathway signatures. The PathCMap approach is based on the connectivity map
concept. It is based on 1) gene expression can be measured accurately and has shown promise as the “universal language” in disease characterization and prognostication; 2) gene expression can be used to connect different biological states and systems; 3) biological pathways drive disease phenotypes and the connectable traits as well. Using PathCMap, small molecules can be easily identified in combination with the targeted therapy which is analyzed and interpreted by our recently developed bioinformatics pipeline (BiNGS!SL-seq). This approach can be applied to query various cancer pathways and can be easily translated into a drug discovery platform.

**P35: AUTOMATED REACTION MAPPING FOR ISOMERS**

*Presenter: TINA KOURI, Colorado School of Mines*

*Authors: Tina Kouri, Dinesh Mehta*

Significant research has been done to enumerate all isomers for a given set of atoms since the isomers are used in structure elucidation and molecular design [1]. Structure elucidation uses information input by an expert user to determine which isomers are consistent with the input information [1]. Molecular design is used to design molecules, such as drugs, which optimize specific characteristics [1]. Since there may be multiple resulting isomers for each application, it is useful to compare and contrast each of the resulting isomers. Our automated reaction mapping algorithms are an important tool for this comparison since they will determine which bonds break and form for any isomer to transform to any other isomer. Due to the size of each isomer (usually over 30 atoms) and the number of isomers for a given set of atoms (usually over 10,000 isomers) it is impractical to use the automated reaction mapping algorithms that we have developed for validating combustion mechanisms (which usually contain a few thousand reactions which involve small molecules of approximately 10 atoms each) to map all pairs of isomers. In this talk we present our ongoing work to develop an improved automated reaction mapping algorithm for the all-pairs mapping problem. Our new algorithm is able to map all of the isomers simultaneously rather than looking at each pair individually. [1] J.-L. Faulon, D.P. Visco Jr., and D. Roe. Enumerating Molecules. Sandia Report. April 2004

**P36: A DATA INTEGRATION APPROACH REVEALS SPECIES-SPECIFIC PHARMACOLOGY BETWEEN HUMAN AND RAT ORTHOLOGS**

*Presenter: FELIX KRUGER, European Bioinformatics Institute*

*Authors: Felix Kruger, John Overington*

We present a large-scale study of small molecule binding to related protein targets. We integrated small molecule bioactivity data and homology information and compared small molecule potencies for pairs of human to rat orthologs. To account for the sample error in the underlying data set, we further compared binding of the same ligand and target in different assays. The obtained
distribution of inter-assay differences was used to identify pairs of orthologs with significant differences in small molecule binding between the human and rat species. Our results indicate that small molecule binding between the human and rat species is conserved for the large majority of protein targets. However, out of 151 orthologous pairs, twelve exhibited statistically different pharmacology in the two species. Among these twelve orthologous pairs is the Histamine H3 receptor, a seven transmembrane receptor with implications in brain function. A homology model of the Histamine H3 receptor was constructed and used to elucidate the molecular mechanism underlying the observed differences between the human and rat ortholog. We identified a mutation Thr119Ala in the third transmembrane domain as the likely cause for species-specific pharmacology of the Histamine H3 receptor. Our study of the robustness of small molecule binding across species confirms a longstanding critical assumption in the field of medicinal and biological chemistry — that of the utility of model organisms to study human biology.

**P37: Accurate Inferring Transcription Regulation from a Compendium of Expression Profiles**

*Presenter: Xueling Li, University of Texas Medical Branch  
Authors: Xueling Li, Dirar Homouz, Andrzej Kudlicki*

We present a new method for gene regulatory network inference, which combines support vector machines (SVM) and Copula transform to integrate multiple experiments. We illustrate the potential of the methodology on 1755 profiles of genome-wide gene expression of Saccharomyces cerevisiae obtained with GPL90 Affymetrix Platform. The Copula-SVM model was validated on YEASTRACT regulation list, which include 48082 interactions between 183 transcription regulators and 6403 target genes. We generate similar number of negative data samples by including only the gene interactions between target genes (TGs). Results shown that copula-SVM has achieved a high accuracy of 95% on 4000 random positive data points and 4000 negatives data points. We further reduced the feature dimension by Wavelet Package Transform (WPT). Results show that feature reduction by WPT has achieved a satisfactory accuracy and have a significant improvement compared with using only one of profile series, yeast metabolic cycle profiles as gene features. Besides much higher performance in Receiver Operating Characteristic (ROC), our copula-SVM method is very general: it successfully predicted regulations by TFs not included in the training set (area under ROC is 0.81) while most methods can only predict new target genes of a known transcription factor given a couple of its known target genes, e.g. SIRENE. It can also predict the directionality of gene association while ARACNe and CLR, etc., can only infer if there is an association between genes. Also, copula-SVM can directly assign probability score to each gene pair to mine the possible false positives and false negatives.
**P38: PHENOPRED 2: PREDICTION OF GENE-DISEASE ASSOCIATIONS BASED ON MULTIPLE NETWORKS**

**Presenter:** BIAO LI, The Buck Institute for Research on Aging  
**Authors:** Biao Li, Sean D. Mooney, Predrag Radivojac

Current high-throughput biotechnologies generate rich types of data and provide immense information with respect to important biological processes and mechanisms. Mining these different types of data together offers opportunities for answering critical biological problems and thereby improves our understanding of them. Locating genes underlying human disease has been long pursued and a number of computational methods have been developed to integrate multiple sources of information. Here we present a new approach, PhenoPred 2, which is based on the earlier PhenoPred algorithm, to easily combining multiple types of data in the form of network to infer novel gene-disease associations. PhenoPred 2 builds on several network data sets, including protein-protein interaction network, sequence similarity, gene expression, ontologies, co-evolution network and proteomics data. We also expand set of known gene-disease associations from PhenoPred by adding curated results from recent publication in genetics. PhenoPred 2 creates a Markov random field (MRF) for each source of data and embeds all MRFs in a Bayesian framework to form a global likelihood function. By Gibbs sampling from this likelihood function PhenoPred 2 derives candidate gene-disease associations. PhenoPred 2 provides an effective way to combine different types of data and make predictions based on expanded and up-to-date knowledge of gene-disease associations.

**P39: EXPLORING AN APPROXIMATE SUBGRAPH MATCHING APPROACH FOR BIOMEDICAL EVENT EXTRACTION**

**Presenter:** HAIBIN LIU, University of Colorado School of Medicine  
**Authors:** Haibin Liu, Karin Verspoor

In the BioNLP-ST 2009 and 2011, we proposed a subgraph-based approach to automatically extract events in the biomedical literature. The precision achieved by our approach is comparable to state-of-the-art systems. However, lower recall prevents our system from achieving a competitive, overall event extraction performance. We observed that many gold events were missed because they are described in grammatical structures that are not covered by the existing rules. These structures tend to be more complex, involving a long dependency path from event trigger to event arguments in the graphs of sentences. In order to explore the generalization potential of event rules, extending the current restrictive framework of the exact subgraph matching (ESM) approach to allow approximate matching is an appropriate solution. In this work, a penalty-based approximate subgraph matching (ASM) algorithm is developed to extend the
original ESM approach. It measures the distance between graphs of rules and sentences by the penalty summation of node mismatch, edge mismatch and gaps of nodes and edges. When a rule graph cannot find its exact match in the sentence, the algorithm starts the approximate matching process which allows nodes or edges of the sentence graph to be skipped with penalty in order to retrieve the potential matching between graphs. We evaluate our algorithm in the context of event extraction and we further investigate other rule generalization methods, such as including the second shortest path when building event rules to enrich the rule set, and post-processing dependency graphs systematically, to compare their performance.

P40: CANCELLED

P41: A PHYSIOLOGICAL PATHWAY PORTAL TOOL AT THE RAT GENOME DATABASE

Presenter: WEISONG LIU, The Medical College of Wisconsin
Authors: Weisong Liu, Diane Munzenmaier, Mary Shimoyama, Melinda Dwinell, Howard Jacob

The Rat Genome Database has created a web-based tool for authorizing and displaying interactive physiological pathway diagrams. The tool is based on the Adobe Flex/BlazeDS technology and consists of two components: 1. Diagram Designer; 2. Diagram Player. The two components can be run in web browsers with Adobe Flash Player installed. This eliminates the hassle of downloading, installing and updating, which is required by most other diagram authoring tools. Using the Diagram Designer, diagram authors can draw physiological pathway diagrams online. Authors can also share and work on same diagrams with others. By dragging and dropping images in the user-friendly Designer, a diagram can be created in minutes. Authors can add text and URLs to every image in a diagram. A diagram can have many groups and layers of components. Authors can decide what part of the diagram to show according to user inputs. A diagram is ready for viewing online upon its creation, no exporting is needed. Diagram viewers can "play" diagrams stored on the server via the Diagram Player. The Diagram Player is feature-packed. Features like dynamic legends, full-screen display, build-in PDF printer, diagram auto-resizing and pop-up blocker detector, make the viewing process convenient and enjoyable. The software allows authors to publish physiological pathway diagrams of physiology, pathology and pharmacology, in system level and tissue level, as a package and make connections to each other. Most commonly used image formats, even video and audio clips, are supported by the software. The software also supports auto-backup and version control.
**P42: EVOLUTION AND ARCHITECTURE OF THE INNER MEMBRANE COMPLEX IN ASEXUAL AND SEXUAL STAGES OF THE MALARIA PARASITE**

Presenter: NOELEEN LOUGHRAN, The Hospital for Sick Children  
Authors: Maya Kono, Ana Cabrera, Noeleen B. Loughran, Susann Herrmann, Klemens Engelberg, Christine Lehmann, Dipto Sinha, Boris Prinz, Ulrike Ruch, Volker Heussler, Tobias Spielman, John Parkinson, Tim W. Gilberger

The inner membrane complex is a unifying morphological feature of all alveolate organisms. It consists of flattened vesicles underlying the plasma membrane and is interconnected with the cytoskeleton. Depending on the ecological niche of the organisms, the function of the IMC varies from a fundamental role in reinforcement to more specialized roles in motility and cytokinesis. In this study we present a comprehensive evolutionary analysis of IMC components, which exemplifies the adaptive nature of the IMCs protein composition. Focusing on 8 structurally distinct proteins in the most prominent genus of the Alveolata — the malaria parasite Plasmodium — we demonstrate that the level of conservation is reflected in phenotypic characteristics, accentuated in differential spatial-temporal patterns of these proteins in the motile stages of the parasites lifecycle. Co-localization studies with the centromer and the spindle apparatus reveal their discriminative biogenesis. We also reveal that the IMC is an essential structural compartment for the development of the sexual stages of *Plasmodium*, as it seems to drive the morphological changes of the parasite during the long and multi-staged process of sexual differentiation. We further found a Plasmodium specific IMC membrane matrix protein that highlights transversal structures in gametocytes, which could represent a genus-specific structural innovation required by *Plasmodium*. We conclude that the IMC has an additional role during sexual development supporting morphogenesis of the cell, which together with its other functions, highlights the multifunctional nature of the IMC in the *Plasmodium* lifecycle.

**P43: LABEL AND EDGE MISMATCH GRAPHLET KERNELS FOR FUNCTIONAL RESIDUE PREDICTION IN PROTEIN STRUCTURES**

Presenter: JOSE LUGO-MARTINEZ, Indiana University  
Authors: Jose Lugo-Martinez, Predrag Radivojac

Computational prediction of functional residues from protein 3D structure plays a significant role in studying mechanistic aspects of protein function, as well as understanding molecular cause of disease upon mutation. In this study, we propose two novel graph-based kernel methods, referred to as the label and edge mismatch graphlet kernels, for annotation of functional residues in protein structures. First, a protein contact graph is constructed from protein structures deposited in Protein Data Bank, where residues are represented by nodes, and edges represent connections between neighboring residues. Next, our algorithm counts all labeled non-isomorphic subgraphs (or graphlets) that
Pivot around a pre-defined residue and satisfy specific label and edge mismatch measures. Moreover, we incorporate evolutionary conservation information into the graphlets labeling. Finally, similarity between two vertices is determined as the inner product of their respective count vectors, and subsequently used in a supervised learning framework to classify protein residues. We report experiments on four residue-level function prediction datasets: identification of catalytic residues, identification of zinc-binding sites and DNA-binding sites, and phosphorylation site prediction. Our graphlet kernels performed as good as or better than established sequence and structure-based approaches. Additionally, we present evidence that the proposed graphlet-based methods account for structural flexibility while efficiently capturing neighborhood similarities in protein structures.

P44: ROLE OF PYRUVATE DEHYDROGENASE KINASES (PDKS) AND THEIR RESPECTIVE MICRO RNA'S IN HUMAN OVARIAN CANCER

Presenter: SHAUKAT MALIK, Mohammad Ali Jinnah University
Authors: Shaukat Malik, S. Sameen, Z. Khalid

Cancer is a metabolic disorder and in energy metabolism PDKs play a very vital role which make them important candidate for involvement in cancer. The bioinformatics analysis performed on ovarian cancer data sets taken from Gene expression Omnibus proved the up regulation of PDK2 and PDK4. The regulating microRNA’s for these two genes were also predicted computationally which were found down regulated in cancer and hence confirmed the over expression of PDK2 and PDK4. Further investigations on the behavior of PDK’s and their corresponding microRNA’s will provide a major break though in cancer research and investigation.

P45: ANALYSIS OF HUMAN COMPLEX DISORDERS FROM THE SYSTEMS BIOLOGY PERSPECTIVE

Presenter: NATALIA MALTSEV
Authors: Dinannath Sulakhe, Alex Paciorkowski, Sandhya Balasubramanian, Ravinesh Kumar, Bingqing Xie, Chalam Chitturi, Eduardo Berrocal, William Dobyns, Natalia Maltsev, Conrad Gilliam

Understanding of genetic mechanisms underlying common heritable disorders (e.g. autism, schizophrenia, diabetes) is one of the most important challenges of modern biology. We present an approach and a supporting computational platform, GEDI (http://gedi.ci.uchicago.edu/), for analysis of common heritable disorders from the systems biology perspective. The approach is based on large-scale integration of experimental data with diverse types of public databases and with molecular interaction data derived from the literature using advanced text mining. GEDI also contains tools and algorithms for analysis and mining of the data. We will illustrate the approach with a comparative analysis of a group
of developmental brain disorders that often co-occur and which are believed to share at least a subset of causative, or predisposing genes (e.g. autism, agenesis of corpus callosum, mid-hindbrain malformations). Whereas we have witnessed progress in the identification of gene mutations that cause, rare Mendelian disorders, the genetic dissection of multi-gene disorders remains a challenge: the disease targets are often ambiguous, or account for only a very small fraction of genetic attributable risk to disease. By using GEDI to predict how individual genes interact with other genes in the context of a given disease pathology, we outline an approach to identify systems-level interactions that underlie disease susceptibility. We show how the comparative analysis of rare and common disorders that share developmental and phenotypic features can be used in conjunction with large-scale data-mining to focus the identification of genes, and interacting genes networks, related to a set of disorders.

P46: GENESMASH: A RESTFUL WEB SERVICE FOR GENE ANNOTATIONS

Presenter: GANIRAJU MANYAM, University of Texas Anderson Cancer Center
Authors: Michelle Payton, Ganiraju Manyam, Chris Wakefield, Jack Roth, Lynne Abruzzo, Kevin Coombes

With the proliferation of new technologies that provide different genome-wide overviews of the molecular landscape within cells, the bioinformatics challenge of integrating disparate sources of information continues apace. Integrated analysis of various biological data types has been a recurrent theme in many ongoing large scale initiatives to understand various molecular pathologies. A fundamental need is to integrate these data, which always requires matching probes across platforms, either by the genes they target or by their genomic coordinates. We present a new website and web service, geneSmash, to collate and provide gene-centric annotations. geneSmash is built upon the Apache open-source database platform, CouchDB. It is a document-oriented, schema-free database with a built-in web server. Database queries are processed through HTTP requests, which are handled by the RESTful JSON API. This feature provides universal accessibility to any modern programming language without any customized API. CouchDB also provides native support for incremental database replication. This would enable the users of geneSmash to maintain their local copy with automatic updates. Since geneSmash provides a generic web service for gene annotations, it can serve as a platform for other specialized applications. We have already developed several such applications including DrugBase. DrugBase is a programmer friendly database of drug-target interactions. It can be an essential tool to integrate drug-target interaction information in high-throughput genomic analyses. Availability:

geneSmash: http://app1.bioinformatics.mdanderson.org/genesmash/_design/basic/index.html
DrugBase: http://app1.bioinformatics.mdanderson.org/drugbase/_design/basic/index.html
Deep sequencing of DNA molecule is becoming widely popular in the field of genomic study. Reducing the cost of this procedure attracts interest of researchers in small laboratories and allows them to conduct studies, which a few years ago were doable only in major genomic centers. In the meantime high-throughput sequencing technology introduces many challenges into data management and analyses. For example, interpretation of second-generation sequencing data requires expertise in different aspects of data manipulation and analysis such as quality control, read alignment, expression quantitation and visualization of massive volumes of data. Many computational methods have been developed to address these questions. However each of these programs is designed to solve one aspect of complex data processing. Tuning all parameters and programs together in a single robust analysis pipeline is highly non-trivial and may still present a challenge for small sequencing facilities. To meet this limitation we have developed Transcriptome Analysis with Circos (TrAC), a novel tool for comparative analysis and visualization of short reads based on Circos. This automated RNA-Seq pipeline is applicable for pre-processing, expression estimation, visualization, and global transcriptome analysis. With TrAC users can investigate the whole pathways, cycles within pathways or linear segments of genes in any given configuration. TrAC is fully customizable and allows end users to study any gene expression-related biological question.

Transcription factors (TFs) are proteins implicated in transcriptional regulation by activating or repressing genes. Finding where those proteins bind to DNA is of key importance to decipher gene regulation at a transcriptional level. As TFs bind to DNA in part through sequence specificity, computational biology has become of great interest to predict transcription factor binding sites (TFBSs). Classically, computational prediction of TFBSs is based on position weight matrices (PWMs). Such models do not allow spacers or flexible length motifs and make the strong assumption that each nucleotide within a TFBS participates independently in the corresponding DNA-protein interaction. We propose to use Hidden Markov Models (HMMs) for the prediction of TFBSs. HMMs are flexible and can model position interdependence within TFBSs as well as spacers and variable length motifs. Constructed HMMs are able to model binding DNA
sequences (corresponding to “matching states”) with their surrounding flanking regions (corresponding to “background states”). The availability of thousands of experimentally validated DNA-TF interaction sequences coming from ChIP-Seq, hopefully representing all the different properties of the associated TFBSs, allows us to construct and train HMMs to reflect the TFBS properties observed in experimental data. HMMs capture deep properties of binding sites that can be discovered in ChIP-Seq data and can be used to predict potential TFBSs accurately. Using ROC curves, the HMMs have been assessed on several ChIP-Seq data sets and we found that they obtain better results than PWMs in discriminating motifs within ChIP-Seq sequences from background sequences.

**P49: A COMPARATIVE ANALYSIS OF KNOWLEDGE-BASED GENE PRIORITIZATION METHODS: IS THE SUPPLY GREATER THAN THE DEMAND?**

*Presenter: NEEL MEHTA, Arizona State University  
Authors: Neel Mehta, Graciela Gonzalez*

A total of 1288 papers were published in the area of automatic gene prioritization in the past four years. As a point of comparison, only 303 appeared between 1983 and 2007. This sudden increase reflects the pressing need for adequate methods to analyze the deluge of genomic data. Gene Prioritization methods are mainly used to find genes that are strongly associated with specific pathologies, or with drug response and metabolism, or other molecular processes. In general, we can classify the published approaches as i) manual ii) semi-automated or iii) fully automated interventions. In general, manual summarization of the literature describing potential genes of interest is nearly impossible to complete, and is used mainly to verify insights obtained by genomic researchers by other methods. Automatic gene prioritization methods vary in approach, with machine-learning based and rule-based systems being common. They usually draw upon different data sources, including i) gene sequence data, ii) functional annotation data, iii) protein interaction data, iv) gene expression data, and v) combination of multiple data sources. Also, they can be based upon Association of genes in a manner such as i) biochemical associations (protein interactions) ii) similarity of biological processes (functions of genes) and iii) regulatory associations (TF regulates the expression of its target genes). The great number of approaches notwithstanding, the actual use of these systems is quite dismal. Only a handful of papers report empirical validation of genes predicted by such tools. In order to advance the computational approaches, researcher buy-in is essential, and the reasons behind this apparent lack of enthusiasm could shed light into ways of improving such methods to make them true catalysts of discovery.
P50: HIERARCHICAL PREDICTION OF ENZYME CLASSES USING ENSEMBLE MACHINE LEARNING APPROACHES

Presenter: AKRAM MOHAMMED, University of Nebraska Medical Center
Authors: Akram Mohammed, Chittibabu Guda

Determining the functional role(s) of enzymes is very important to build the metabolic blueprint of an organism and to identify the potential roles enzymes may play in metabolic and disease pathways. With exponential growth in gene and protein sequence data, it is not feasible to experimentally characterize the function(s) of all enzymes. Therefore, there is a need to develop a computational method that can predict enzyme class from protein sequences. Given the incomplete and unbalanced nature of annotations in biological databases, ensemble methods or methods that bank on a combination of orthogonal features are more desirable than using single-feature-based individual classifiers for achieving higher accuracy and coverage in enzyme classification. We have collected a comprehensive dataset of all the known enzyme sequences in the public domain. Given these labeled enzyme sequences belonging to different enzyme classes, class-specific enzyme features from PFAM, PROSITE, SCOP and Gene Ontology databases are extracted. Using these features with supervised machine learning algorithms such as ensemble learning, Bayesian and support vector machines, we are developing a new method that can hierarchically identify enzymes from non-enzymes and also the sub-classes within a broader enzyme class. We will validate the accuracy of our enzyme prediction method and apply the method to predict the unknown enzymes from sequenced proteomes.

P51: INFERRING THE MAMMALIAN PHYLOGENY USING A HETEROGENEOUS APPROACH

Presenter: CLAIRE MORGAN, Dublin City University
Authors: Claire C. Morgan, Peter G. Foster, Andrew E. Webb, Davide Pisani, James O. McInerney, Mary J. O’Connell

65 million years ago marked a period of rapid speciation and diversification in the mammalia. As a consequence resolving the order in which these species of mammalia arose has proven extremely difficult. There are currently four major competing hypotheses for the topology of mammals each of which have been generated using molecular data and various phylogenetic approaches. From decades of phylogenetic research we know that the quality of data, taxon sampling, and reconstruction methods, impact hugely on the phylogenetic outcome. More recently, the development and implementation of Next Generation Sequencing technologies has resulted in a surge of mammalian genomes being completed to a far higher quality. We are now at an exciting juncture in the understanding of mammalian evolution. With greater availability of higher quality data and novel heterogeneous models of sequence evolution, along with a greater understanding of the evolutionary process, we have addressed the longstanding question of mammalian phylogeny.
**P52: AN ISSUE OF CAUTION IN THE USE OF MITOCHONDRIAL DATA TO INFER THE MAMMALIAN PHYLOGENY**

*Presenter: CLAIRE MORGAN, Dublin City University*

*Authors: Claire C. Morgan, Christopher J. Creevey, Paul Kilroy-Glynn, Mary J. O’Connell.*

The rate at which mutations arise across species and within genes is highly variable. It is known that mtDNA is more susceptible to mutations as compared to nuclear DNA. mtDNA has been used on many occasions to infer phylogenetic relationships between distantly related species, most notably the Mammals. Our study looks at the suitability of mtDNA as a phylogenetic marker. We examine the balance between phylogenetic signal and noise at various phylogenetic depths, for each of the mitochondrial coding genes. We found that none of the mitochondrial genes contained sufficient phylogenetic signal at the base of the placental mammals. However, there is an improvement in the phylogenetic signal of the data as we move from the deepest to the most recent nodes. Using 16 datasets, representing 9 mitochondrial genes across 378 placental mammals, we have successfully resolved the intra order placement of placental mammals through supertree analysis, but the root remains unresolved using mitochondrial DNA.

**P53: ENTROPY-BASED ASSESSMENT OF CONDITION-SPECIFIC HISTONE MODIFICATIONS**

*Presenter: TADASU NOZAKI, Keio University*

*Authors: Tadasu Nozaki, Yuki Shindo, Rintaro Saito, Masaru Tomita*

Histone modification is known as one of the epigenetic regulatory mechanisms. Ubiquitous modifications of histones may cause associated genes to be expressed or repressed ubiquitously whereas condition-specific modifications may cause the genes to be regulated only in specific condition. Such epigenetic regulations have been shown to play a key role in dynamics of cellular processes including differentiation. To evaluate conditional specificities of histone modification patterns, we introduced two measurements (Hs and Hu), based on entropy which have been applied to the analysis of cell-type specificities of gene expression or DNA methylation patterns. In this work, Hs and Hu have been specialized to assess the cell-type specificity and the ubiquitousness of histone modifications respectively. Using histone modification patterns in 7 human cell types obtained by ChIP-Seq, we showed that histones having low Hs values are likely to be modified only in specific cell types whereas those having high Hu values are likely to be modified ubiquitously through several cell types. Furthermore, we found that genes are likely to be expressed in cell-type specific manner if Hs values of the histones which are located around transcription start sites are low, indicating that expressions of these genes might be regulated by histone modifications. We suggest that these two measures, especially when they are used with both histone modification data and gene expression data, will help us to predict genes which participate in epigenetic regulatory networks, which can be targets for further experimental investigations.
P54: PREDICTION OF PROTEIN FUNCTION USING ANALYSIS OF INTERACTION MAPS ACROSS SPECIES

Presenter: KYMBERLEIGH PAGEL, Indiana University, Bloomington
Authors: Kymberleigh Pagel, Wyatt T. Clark, Predrag Radivojac

The prediction of protein function is an important problem that has been approached in many different ways with moderate success. Function prediction is important because there are limited resources that can be devoted to experimentally determining the function of the already large number of uncharacterized genes. We intend to develop a novel method for the prediction of protein function by incorporating non-homologous data sources into a common graph based representation. We plan to integrate several non-homologous data types, including protein-protein interaction (PPI) data, evolutionary relationships based on homology, as well as sequence similarity. In doing so we will be able to analyze existing algorithms that have already been applied to graphs generated solely from PPI data and whose edges have been restricted to existing between pairs of genes within the same organism. We intend to analyze the ability of different combinations of non-homologous data types to infer function and will provide recommendations for the best methods for incorporating different data sources into a common graph based representation. Finally, because our graphs will contain edges between proteins in different organisms, we will perform an analysis of how much these edges contribute, or detract, from the prediction of function. Specifically, we will be interested, given the particular data at hand, how much edges between particular pairs of organisms contribute to the prediction of function.

P55: DECIPHERING OF HUMAN PROTEIN INTERACTOME USING STRUCTURAL COMPLEXES

Presenter: ANNA PANCHEŃKΟ, National Center for Biotechnology Information, Institutes of Health
Authors: Manoj Tyagi, Kasuke Hashimoto, Benjamin Shoemaker, Stefan Wuchty, Anna R. Panchenko

Proteins function by interacting with other biomolecules and knowledge of the entire set of interactions combined with the properties of protein binding sites is essential for our understanding of cellular functions and the origins of many diseases. Recently we developed a method (IBIS-Inferred Biomolecular Interaction server) which analyzes and predicts interaction partners and locations of binding sites in proteins based on the evolutionary conservation of binding sites in homologous structural complexes. IBIS imposes a number of rigorous criteria in order to increase the reliability of homology-based inference of interactions and provides binding site annotations for five different types of interaction partners (proteins, small molecules, nucleic acids, peptides and ions). We show that our method can reach 70-80% sensitivity and specificity for protein-protein and protein-small molecule site annotations. It facilitates the mapping of the entire biomolecular interaction network for a given organism.
and we use this framework to map the human protein interactome and analyze its properties. We show that structurally inferred interaction network is highly modular and has small-world characteristics. Moreover it is more functionally coherent and reliable compared to high-throughput interaction networks.

**P56: REGULATION OF PROTEIN-PROTEIN BINDING BY PHOSPHORYLATION AND OLIGOMERIZATION**

*Presenter: ANNA PANCHENKO, National Center for Biotechnology Information, Institutes of Health*  
*Authors: Hafumi Nishi, Kosuke Hashimoto, Anna R. Panchenko*

Cellular regulatory mechanisms provide a sensitive and specific response to external stimuli and such dynamic regulation can be achieved through reversible covalent modifications. We study the effect of phosphorylation on protein binding and function for different types of complexes from the human proteome. We observe that phosphorylation sites have a tendency to be located on binding interfaces in heterooligomeric and weak transient homooligomeric complexes. Our analysis of molecular mechanisms of phosphorylation shows that phosphorylation may modulate the binding affinity and trigger the transitions between different conformer and oligomeric states. We discuss the possible role of transitions between different homooligomeric states of the same protein in the regulation of protein activity and compile a set of experimental examples with such regulatory mechanisms. Although majority of phosphosites do not show significant stability differences upon phosphorylation, for about one third of all complexes phosphorylation causes relatively large changes in binding energy. We show that phosphorylation sites are not only more likely to be evolutionary conserved than surface residues but even more so than other interfacial residues.

**P57: CYSNO-DB: A DATABASE OF S-NITROSYLATION SITES**

*Presenter: SANJIT PANDEY, University of Nebraska Medical Center*  
*Authors: Sanjit Pandey, Hong Peng, Shi-Jian Ding, Chittibabu Guda*

Protein S-nitrosylation is the covalent redox-related modification of a cysteine sulfhydryl group with nitric oxide (NO). It is increasingly becoming recognized as a ubiquitous regulatory reaction comparable to phosphorylation. S-nitrosylation may function as an important regulatory mechanism for fine-tuning protein activities within diverse cellular processes and biochemical pathways, including signal transduction, DNA repair, ion channel regulation and apoptosis. There is no web-based resource that makes information on the S-nitrosylation sites accessible to the research community. The purpose of this project is to develop a resource containing all the site information currently available and develop a prediction tool to identify potential nitrosylation site(s) in a given protein sequence. We have recently developed a resource, CysNO-DB, that collects all the known S-nitrosylation sites. CysNO-DB is a database containing S-nitrosylation site information for proteins from different species (Mouse and
POSTER PRESENTATIONS

Human). The database currently has data curated from published literature, and proteomic data obtained from Dr. Shi-Jian Ding’s lab at University of Nebraska Medical Center. The database currently has 9993 distinct nitrosylation sites, which is accessible at “http://cysno.unmc.edu”. It currently allows users to search the database using Organism, IPI ID, Gene Name, Keyword or protein sequence. The query output page also allows users to link to the uniprot and KEGG database whenever possible. We are currently working to develop a tool for S-nitrosylation site prediction that will be incorporated with the website. We are currently exploring machine-learning approaches (Bayesian and SVMs-Support Vector Machines) to build the S-nitrosylation prediction tool.

P58: A DATABASE FRAMEWORK FOR IDENTIFYING CODING SEQUENCE FROM NEXTGEN mRNA SEQUENCES

Presenter: SANJIT PANDEY, University of Nebraska Medical Center
Authors: Sanjit Pandey, Chittibabu Guda, Robert B. Norgren, Jr.

NextGen sequencing is an attractive approach for generating de novo transcriptomes. We used the Velvet/Oasis assembler pipeline with Illumina mRNA sequences to generate transcript contigs for rhesus macaques. However, once the contigs were generated, identifying the complete coding sequences (CDS) required additional processing. Although there are many ways in which this could be accomplished, we have chosen to use MySQL to develop a database framework that is used both to extract CDS and to store data associated with the project. Because we are working with a closely related species, we were able to leverage information obtained from the human genome project. Briefly, rhesus RNA contigs were aligned with full-length human mRNA sequences using BLAST. Information from this alignment was imported into the MySQL database. Information regarding CDS locations within human transcripts was also stored in the MySQL database. A suite of MySQL procedures were created to use the human CDS locations to identify the orthologous rhesus CDS. We performed mRNA-seq with several different samples. Although, it is not permissible to assemble contigs across samples, we utilized contigs from different samples to produce model transcripts. This was accomplished with a series of MySQL procedures. Using the pipeline we have developed, we believe it will be possible to identify the majority of rhesus macaque CDS from a relatively small set of mRNA samples. This will be useful for transcriptomics, proteomics and genome annotation.

P59: MODPRED: PREDICTING POST-TRANSLATIONAL MODIFICATION SITES FROM PROTEIN SEQUENCE

Presenter: VIKAS PEJAVER, Indiana University
Authors: Vikas Pejaver, Predrag Radivojac

Post-translational modifications (PTMs) are chemical groups that are covalently attached to amino acids, during or after protein translation. Currently, due to
experimental limitations in the detection of PTM sites, there has been a growing
interest in predicting them computationally. However, recent tools allow for
the prediction of only specific types of PTM sites and are not amenable for
proteome-wide prediction of a variety of modification sites in each protein.
Moreover, in the case of motif-associated modifications such as N-linked
glycosylation and sumoylation, it has been observed that motifs alone are not
predictive of these modifications. We are currently building sequence-based
predictors for 17 types of PTM sites in proteins. These include acetylation,
amidation, N-linked glycosylation, phosphorylation, ubiquitination, among
many others. We have collected experimentally-verified PTM sites from several
databases and literature sources and have extracted features (sequence-based
and evolutionary) associated with these sites. We are currently building random
forest models for each PTM type. In the case of motif-associated PTMs, we build
two separate models for prediction — motif-based and non-motif based. The
eventual goal is to implement ModPred, a tool (web-based and standalone)
that, given a protein sequence, would accurately predict sites for each of the
PTM types. Thus, from a user’s perspective, ModPred would serve as a “unified”
predictor of major PTM sites, and would eliminate the need to repeatedly run
sequences through multiple predictors for individual PTM types.

P60: A MULTIPLE-TEMPLATE APPROACH FOR PROTEIN
THREADING

Presenter: JIAN PENG, Toyota Technological Institute at Chicago
Authors: Jian Peng, Jinbo Xu

Due to the increasing number of solved structures, a protein without solved
structure is very likely to have more than one similar template structures.
Therefore, a natural question to ask is if we can improve modeling accuracy
using multiple templates. This work describes a new multiple-template
threading method to answer this question. At the heart of this multiple-template
threading method is a novel probabilistic-consistency algorithm that can
accurately align a single protein sequence simultaneously to multiple templates.
Experimental results indicate that our multiple-template method can improve
pairwise sequence-template alignment accuracy and generate models with
better quality than single-template models even if they are built from the best
single templates while many popular multiple sequence/structure alignment
tools fail to do so. The underlying reason is that our probabilistic-consistency
algorithm can generate accurate multiple sequence/template alignments. In
another word, without an accurate multiple sequence/template alignment the
modeling accuracy cannot be improved by simply using multiple templates to
increase alignment coverage. According to the CASP9 official evaluation, our
method outperforms almost all other CASP9 servers and our method generated
the best alignments for the 50 hardest template-based modeling targets. Our
method was also voted by the CASP9 community as one of the most innovative
and interesting methods. Our method will have greater potential in the near
future when many more templates are available due to the increasing number
of solved structures. We can further improve alignment accuracy by extending our algorithm to simultaneously thread multiple homologous sequences to multiple templates.

P61: A TOOL FOR ASSESSMENT AND VALIDATION OF SEQUENCING-BASED METHODS IN FAMILY DISEASE STUDIES

Presenter: ALEXANDER POOLE, University of Colorado Anschutz Medical Campus
Authors: Alexander Poole, Nancy Miller, Robin Dowell

With the completion of the Human Genome Project came the expectation that we would quickly discover the specific genetic components responsible for heritable traits and diseases. However, researchers soon discovered that phenotypes are often governed by far more complex interactions than initially thought, with contributions from multiple genetic loci and environmental factors. After many subsequent genome-wide linkage and association studies, and many millions of dollars spent, there has been only limited success in understanding the mechanisms of complex diseases within families. Clearly more sophisticated methods for detecting the genetic basis of complex traits need to be developed. Second-generation sequencing technology has the potential to detect genomic differences at a much higher resolution and is increasingly being used for these comparative genomics studies. Therefore, the tools for evaluating our ability to detect causal variants need to adapt for both the new short read sequencing data and our evolving understanding of multigenic phenotypes. We propose a pedigree simulator for studying complex heritable traits and diseases that accounts for genome-wide variation representative of actual individuals, generates realistic high-throughput sequencing test data sets, provides a way to asses genome-wide accuracy of read mapping, variant calling, phasing and recombination, and to evaluate methods for detecting causative genetic variants.

P62: ALTERNATIVE SPLICING DETECTION USING RNASSEQ: AN ASSEMBLY APPROACH

Presenter and Author: LIKIT PREEYANON, Michigan State University

Recently, RNA sequencing (RNAseq) from Next Generation Sequencing (NGS) technology has been successfully used to study alternative splicing in humans and mice. Methods used in these analyses rely solely on high quality gene models. Consequently, these methods are not suitable for other organisms lacking high quality gene annotations. To overcome this problem, other methods have been developed; for example, one method does not rely on an existing annotation but instead constructs the gene models from sequence reads that are mapped to the genome. However, the method is limited by using only sequence reads that are mapped to the genome and gene models are built based on a computational prediction. We have developed a new method,
employing an assembly approach, to build gene models and identify alternative isoforms. We have been using this method to study alternative splicing in chickens from two inbred lines that are resistance and susceptible to Marek’s disease (MD). The method identified many novel genes and isoforms that are not included in existing gene models. Isoforms were confirmed by reads that mapped across the exons. All kinds of alternative splicing events were detected including cassette exons, alternative splice sites, alternative 5’ and 3’ end, mutually exclusive exons and intron retention. Moreover, the method has successfully detected alternative isoforms between two chicken inbred lines that might contribute to genetic resistance to MD. Furthermore, this method does not rely on existing gene annotations, therefore, it can be applied to study alternative splicing in any organism.

**P63: IN SILICO RATIONAL DRUG DESIGN AND MODELING STUDIES OF NOVEL INHIBITORS FOR MULTI-TARGET INHIBITION IN PSEUDOMONAS AERUGINOSA**

*Presenter: JAYARAMAN PREMKUMAR, Nanyang Technological University*

*Authors: Premkumar Jayaraman, Lim Chu Sing Daniel, Meena K Sakharkar*

Pseudomonas aeruginosa is a major nosocomial and opportunistic pathogen, with the ability to develop multi-drug resistance which significantly reduces the efficacy of many commercially available antibiotics. Hence, there is an urgent need to develop novel therapeutic strategies to combat the development of resistance. On the sequence of our previous work on drug combination assays and common pharmacophoric features data, we have designed and evaluated new drug scaffolds, containing 6-methoxypyrimidine-2,4-diamine analogue conjugated with benzenesulfonamide derivatives, an alternative novel drug-design strategy for inhibition of quadri-enzymes (dihydropteroate synthase (DHPS), dihydrofolate reductase (DHFR), DNA gyrase subunit B and topoisomerase IV subunit parE (topoIV)) of two different pathways. To validate the hypothesis that these novel scaffolds could act as multi-target inhibitors, we have used computational techniques such as molecular docking study, molecular and electronic properties analysis and dynamics simulations. The docking model and dynamics study predicts that these inhibitors have favorable binding to all the four quadri-enzymes, forming strong hydrogen and hydrophobic interactions with key active site residues. The predicted structure-activity relationship of the proposed hybrid drugs with respect to physicochemical properties and stereo-electronic properties such as HOMO, LUMO and molecular electrostatic potential maps calculated using quantum chemical methods were found to be well correlated with the common pharmacophoric features required for the multi-target inhibition. The key innovative aspects of this study is to provide novel insights on preventing the emergence of drug resistance based on a rational multi-target drug design and can serve as a prospective lead in the development of anti-pseudomonal drug development.
P64: QUANTITATIVE DATA: WHERE ARE THEY HIDDEN IN BIOMEDICAL LITERATURE?

Presenter: KOMANDUR RAVIKUMAR, University of Colorado School of Medicine
Authors: K.E. Ravikumar, Meenakshi Narayanaswamy, S.V. Ramanan

Computational modeling of biological systems has been greatly impeded by the fact that extracting the values of various parameters from the literature has been primarily a manual task. Automated extraction of such quantitative data from the bio-medical literature, a largely underexplored problem would significantly aid the curation of such information in databases such as KEGG, CellML models repository, BioModels database, which are often a good starting point for system biologists involved in modeling biological pathways and systems. Here, we propose a method to automatically extract quantitative data from the literature, specifically in the context of electrophysiology. We used (i) dictionary lookup (e.g. conductance, membrane potential) and regular expressions (e.g., Kd, Ki) to tag electrophysiological parameters and (ii) regular expressions to detect their values (with units, e.g., 20 pS, -80 mV) as they occur in the text. Linguistic rules were developed to pair occurrences of parameters and compatible values within the clause; we used a development corpus of 150 abstracts. We also extended our approach beyond clausal boundaries by considering all possible relationship pairs that co-occur within a sentence, but retaining only compatible parameter-unit pairs. The precision, recall and F-measure for extracting quantitative relationship pairs were 83.2%, 73.20%, and 77.88% respectively, when evaluated on a small test corpus. While the intra-clausal linguistic patterns gave highly precise relations, the extra-clausal pairing mechanisms significantly improved the recall (by 10%) without any drop in precision.

P65: TRANSCRIPTOME PROFILE OF BOVINE RESPIRATORY DISEASE PATHOGEN — MANNHEIMIA HAEMOLYTICA PHL213

Presenter: JOSEPH REDDY, Mississippi State University
Authors: Joseph Reddy, Ranjit Kumar, Mark Lawrence, Shane Burgess, Bindu Nanduri

An in-depth understanding of an organism begins with sequencing its genome, identifying expressed regions, and then assigning functions to these expressed regions using experimental as well as in silico methods. Most genomes are annotated using computational algorithms such as GLIMMER or GeneMark.hmm that use markov models generated from statistics of previously annotated genomes. Gene expression data from deep sequencing experiments such as RNA-Seq provide a high resolution transcriptome profile that helps in determining operon structure, identify expressed genes not predicted computationally (especially small RNAs), and helps in accurately quantifying...
gene expression. In this study, RNA-Seq was used to improve the existing genome annotation of Mannheimia haemolytica PHL213, one of the primary pathogens of Bovine Respiratory Disease (BRD) in cattle. Using the Illumina platform, we generated 9,055,826 reads (average length 70nt) that were aligned to the reference genome using BOWTIE. The transcribed regions were analyzed using SAMTOOLS and in-house PERL scripts. Analysis of expressed intergenic regions (EIR) led to the identification of 40 small RNA. This single nucleotide resolution map enabled the identification of 13 genes with alternate/mutated start/stop sites and 3 frameshift mutations within genome. We identified 14 potentially novel protein coding regions missed by the initial annotation. Basal transcription profile indicates that 2617 of the previously annotated 2837 genes were expressed, representing all broad functional gene categories in the genome. Application of RNA-Seq based transcriptome profiling improved the structural annotation of M. haemolytica.

**P66: SPACERSEEKER: DETECTING AND IDENTIFYING IMMUNITY CONFERRED BY CRISPR LOCI**

Presenter: ZACHARY ROMER, Loyola University Chicago
Authors: Zachary Romer, Catherine Putonti

Clustered regularly interspaced short palindromic repeats or CRISPRs have recently been identified in half of all bacterial and nearly all archaeal genomic sequences. CRISPRs provide the bacteria/archaea with “immunity” against viruses and plasmids, recognizing “foreign” DNAs which match spacer sequences located within the CRISPR loci. These 26- to 72 base pair spacer sequences within the microbial genomes are identical to the corresponding phage or plasmid genomic sequence. Identifying the source of spacer sequences within the genomes of bacteria and archaea can provide insight into the individual microbe’s resistance and prior exposure to particular bacteriophages and/or plasmids. Detecting CRISPR sequences has typically relied on the identification of the 21- to 48-base pair directed DNA repeats which separate the spacers. While not originally designed for detecting CRISPR repeats, existing tools have been quite successful in identifying spacer sequences. The vast majority of these spacer sequences, however, do not BLAST to any known viral genome sequences. Herein we present a new software tool which utilizes known virus genomes and our knowledge of the structure of CRISPR loci. Rather than target identification of the repetitive elements, our approach looks specifically for spacer sequences. This tool was developed with the specific purpose of identifying CRISPRs in unassembled metagenomic next-generation sequencing reads. Using this tool we examined both publicly available genomic sequences as well as metagenomic sequence collections.
Single Nucleotide Polymorphisms (SNPs), the commonest types of genetic variation, have proven to be very important in the study of human health. One main challenge with studying SNPs in research is to determine the effect of a mutation on different biological systems. The enormous amount of data and data distribution among different resources make this task especially difficult. To assist researchers in this field we are continuing to develop MutDB which is a comprehensive, up-to-date, and user friendly web application providing annotations for non-synonymous SNPs on coding areas in the human genome. MutDB integrates data from dbSNP, Swiss-Prot, and Cosmic. In cases when the source of the mutation provides information only on the protein level or have missing information, we map the positions to the transcripts and protein of the gene as well as to the chromosome. We also apply intensive validation checks to assure reliability. We then offer a wide range of annotations like sources, somatic vs. germine, Uniprot classification, PubMed id, and structure annotations, and present the locations of the SNPs on the GRCh37 Genome Reference build. We use our tool MutPred to predict whether a change would have a deleterious effect as well as to predict the underlying cause for the disruption of function such as gain or loss of multiple post translational modifications, impact on protein stability, structure and other relevant attributes.

Originally designed to detect gene-gene and gene-environment interactions, Multifactor Dimensionality Reduction (MDR) is a popular data mining algorithm that uses discrete predictor variables to model a binary outcome. Given a model size $k$, the algorithm constructs all size $k$ combinations of predictors and reduces each combination to a MDR variable that is evaluated relative to disease status (case vs. control). The combinatorial nature of the algorithm results in running times which exponentially increase as $k$ increases from 1 to the number of variables/2. Thus, a parallel implementation of MDR for Graphics Processing Units (GPUs), video cards, was proposed. The original MDRgpu supported a maximum of 3-way interactions and was only compatible with CUDA capable NVIDIA GPUs. We address these limitations by enabling the modeling of sufficiently large $k$ and porting the algorithm to OpenCL. OpenCL is a C language extension which allows kernels to be executed on GPUs from multiple video
card vendors including NVIDIA and AMD. Thus, the MDRgpu performance benefits are magnified by enabling both large k-way interactions and the use of OpenCL compatible GPUs.

**P69: IDENTIFYING TREATMENT RELEVANT COLORECTAL CANCER SUBTYPES USING ITERATIVE NON-NEGATIVE MATRIX FACTORIZATION**

**Presenter:** ANDREAS SCHLICKER, Netherlands Cancer Institute  
**Authors:** Andreas Schlicker, Garry Beran, Christine M Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, Sara Davenport, Kerry Heathcote, Denis Alferz Castro, George Orphanides, Tim French, Lodewyk FA Wessels

Genetic and epigenetic features have been used to define colorectal cancer (CRC) subtypes and to take treatment decisions. To develop new targeted drugs, however, it is necessary to gain a better understanding of the molecular differences of CRC subtypes. We developed a new unsupervised approach for stratifying tumor samples using genome-wide mRNA expression data. Our method, iterative non-negative matrix factorization (iNMF), is based on the iterative application of non-negative matrix factorization based on randomly selected sets of genes. In a gene expression dataset consisting of 63 CRC tumors, we identified two dominant subtypes. These subtypes were highly concordant with the classes induced by an epithelial-mesenchymal-transition (EMT) gene expression signature. This finding is consistent with previous results. Further stratification of the tumor samples revealed five subtypes. These subtypes exhibit many differences, most notably differential activation of specific signaling pathways. Importantly, employing the derived subtype gene signatures, we stratified several independent, published datasets, suggesting that the signatures capture disease-relevant intrinsic features of CRC. Furthermore, application of the gene signatures to expression data obtained from cell lines revealed that the tumor subtypes were covered in all panels analyzed. Integrating pharmacological response data allowed us to identify several targeted compounds showing differential response across the subtypes. The CRC stratification obtained with our new method, iNMF, offers valuable insight into the differences between CRC subtypes at a functional level. Most importantly, it captures features of the disease that are highly relevant for the development of new targeted drugs in defined CRC patient sub-populations.

**P70: A NEW PROBABILISTIC MODEL IN PREDICTIVE MICROBIOLOGY (NPMPM)**

**Presenter:** NADINE SCHOENE, Goethe University Frankfurt  
**Authors:** Nadine Schoene, Alexander Bockmayr, Bernd Appel, Annemarie Kaesbohrer

Predictive microbiology is a basic component of microbial risk assessment. It can help to prevent foodborne disease outbreaks by detecting probable contamination sites and defining monitoring points. The goal of existing models
in predictive microbiology is to gain understanding of population kinetics. The
global error is minimized by fitting the parameters with pooled experimental
data. But for a risk assessment it is essential to predict the bacterial count at a
single time point (at the end of the process chain) as precise as possible, i.e., the
local error has to be minimized. The New Probabilistic Model in Predictive
Microbiology (NPMPM) is based on a new approach for including variability
and uncertainty into modeling of microbial growth. It was developed for risk
assessment of bacterial contamination in the food supply chain. The NPMPM
was implemented as an R package. Model assumptions are kept simple and
include exponential growth as basic form of population kinetics and log-
normal distributions for bacterial counts. Internal validation with simulated
data that fulfilled all model assumptions showed that the NPMPM is able to
reproduce the data it was built with. For certain conditions this was shown
theoretically, too.

P71: SIMPSONS PARADOX AND CORRELATION CANCELLATION IN GENOME-WIDE ASSOCIATION STUDY DATA
Presenter: RONALD SCHUYLER, University of Colorado Denver
Authors: Ronald Schuyler, Larry Hunter

Genome-wide association studies (GWAS) generate allele frequency data for
hundreds of thousands of markers in case and control population samples.
Evaluating the joint distributions of multiple markers using logistic or log-linear
methods reveals highly significant associations missed by standard single-
marker association tests. By testing all pairs of markers for deviation from
a null model in which neither marker contributes to disease susceptibility,
we observe a specific pattern of association that is not evident in the typical
single-marker analysis. While all significant pairs are in close physical proximity
on the genome, they display a range of linkage disequilibrium from 0.0 to 0.6
(Pearson’s correlation from -0.4 to 0.8). Examining the direction of effect at
each marker and the sign of the correlation of each pair provides a possible
explanation of why the effects at these highly significant loci are not observed
in a standard analysis. For all significant pairs in which the bivariate regression
coefficients are of opposite sign, the markers are positively correlated. For
significant pairs in which the bivariate regression coefficients are of the same
sign we observe negative correlation between the markers. This pattern, and
the lack of significance in single marker tests, may be explained by the effect
of correlation cancellation, corresponding to a variation of what is commonly
known as Simpson’s paradox. Further work is necessary to determine the true
causal structure underlying these association patterns.
Chromatin is organized in the nucleus into chromosome territories. The arrangement and movement of these territories affect gene expression and play a crucial role in cell differentiation. Recent discoveries have been rapidly expanding our understanding of this structure. However, the exact structure and physical mechanisms that cause it to form are still poorly understood. The newly developed Hi-C method has yielded the first whole genome map of human chromosome territories using wet lab experiments. This study utilizes Hi-C data to investigate the connection between sequence content and chromosome territories. I test the hypothesis that similar sequences physically cluster together. Chromosome sequence is broken into 1Mbp regions. For each region, a vector containing the number of occurrences of every possible 5bp oligomer is created. Sequence similarity between two regions is then calculated as the correlation between two vectors. The best predictors are identified using a hill climbing algorithm which includes/excludes oligomer counts from the similarity calculation. The results show a median correlation between sequence similarity and 3D position of 0.52 over 40,000 data points. This shows that similar sequences tend to cluster together creating . Using this method, human chromosome territories can be predicted computationally using the genome sequence. Furthermore, this study identifies the 70 oligomers most predictive of chromosome territories which are enriched for simple repeats and tRNA. The ability to predict chromosome territories is an important tool in understanding the mechanisms behind chromosome territories and cell differentiation.

Functional profiling of pharmacogenetic non-synonymous SNPs Little is known about the nature of pharmacogenetic (PGx) variants as compared to disease-causing genetic variants and neutral genetic polymorphisms. Similarly to disease-causing variants, the pharmacogenetic SNP may become invaluable for personal genetic applications. We are employing bioinformatic methods, to annotate the protein level consequences of pharmacodynamic (PD) and pharmacokinetic (PK) variants in the PharmGKB® database to with the goal of predicting candidate PGx variants. We are working towards a pharmacogenetic fingerprint of features that describe protein variants. Our research involves two separate strategies to separate PGx from neutral and PD variants from PK. The first involves gene and protein attributes determined from GO, KEGG, Reactome, HRPD and co-expression networks where known PGx entities are substantially
enriched relative to the human genome. These include such annotations as Glycosylation (PD, PK), Metal ion binding (PD, PK), Intracellular signaling (PD) and Oxidation-reduction (PK), using DAVID and GeneMANIA. The second strategy operates at the variant level wherein we employ machine-learning classifiers, such as Random Forest and Support Vector Machines, to known PGx variants, in order to evaluate features that will differentiate these from neutral SNPs or disease-causing mutations. Our dataset includes 143 known PGx variants and over 26,000 known and presumed neutral ones. The imbalance in this dataset results in a well-known challenge of having an excellent ROC-AUC along with relatively low precision. We will summarize our findings and describe how a PGx variant compares to the molecular attributes that describe disease-associated and neutral variants.

**P74: SEMI-SUPERVISED LEARNING IMPROVES GENE EXPRESSION-BASED PREDICTION OF CANCER RECURRENTNESS**

*Presenter: MINGGUANG SHI, Vanderbilt University*

*Authors: Mingguang Shi, Bing Zhang*

Motivation: Gene expression profiling has shown great potential in outcome prediction for different types of cancers. Nevertheless, small sample size remains a bottleneck in obtaining robust and accurate classifiers. Traditional supervised learning techniques can only work with labeled data. Consequently, a large number of microarray data that do not have sufficient follow-up information are disregarded. To fully leverage all of the precious data in public databases, we turned to a semi-supervised learning technique, low density separation (LDS). Results: Using a clinically important question of predicting recurrence risk in colorectal cancer patients, we demonstrated that 1) semi-supervised classification improved prediction accuracy as compared to the state of the art supervised method SVM; 2) performance gain increased with the number of unlabeled samples; 3) unlabeled data from different institutes could be employed after appropriate processing; and 4) the LDS method is robust with regard to the number of input features. To test the general applicability of this semi-supervised method, we further applied LDS on human breast cancer datasets and also observed superior performance. Our results demonstrated great potential of semi-supervised learning in gene expression-based outcome prediction for cancer patients.

**P75: COMPREHENSIVE ANALYSIS OF ASSOCIATIONS BETWEEN ALTERNATIVE SPLICING AND HISTONE MODIFICATIONS**

*Presenter: YUKI SHINDO, Keio University, Japan*

*Authors: Yuki Shindo, Tadasu Nozaki, Rintaro Saito, Masaru Tomita*

The regulation of alternative splicing (AS) was traditionally thought to be achieved by the enhancers and silencers located either in exons or introns. However, recent studies suggested that histone modifications affect the splicing...
machinery and can alter the exon inclusion pattern, although their extent of associations remains unclear. Here, we systematically analyzed the relationship between AS and histone modifications, especially focusing on the associations of the specific type of histone modification and exon inclusion/exclusion patterns. Using ChIP-Seq data, we first determined genomic positions of histones around intron-exon boundaries. We split the set of exons into “included” and “excluded” groups and calculated the frequency of the exons harboring modified histones within 200 bp upstream and downstream from the intron-exon boundaries in each group. The results showed that there were significant differences in the rate of modified histones between included and excluded groups in several types of modifications. For instance, H3K36me3 was enriched around the included exons, which was consistent with the previous studies. We also found histone modifications including H2AK5ac and H3K18ac were enriched around excluded exons, which might be the first example to suggest the role of these modifications in exon skipping. Moreover, we found that for some exons, changes in specific types of modifications of nearby histones among H1 and IMR90 cell lines are strongly associated with inclusion/exclusion patterns of these exons in each cell line. Taken together, our study provides further insights into the regulation of AS via histone modifications, which might reflect the cell type- and/or stage-specificity.

**P76: DEVELOPING A DECISION SUPPORT SYSTEM FOR STRATEGIC SELECTION OF DRUG TARGETS TO COMBAT TROPICAL DISEASES**

**Presenter:** FLORIANO SILVA-jr, FIOCruz  
**Authors:** Kele Belloze, Maria Cavalcanti, Sabrina Ferreira, Floriano Silva-Jr

As a knowledge-driven process, drug discovery and development should have been greatly empowered by the vastness of information available on public databases storing huge amounts of genomic, biochemical and pharmacological data, as well as in the biomedical literature. Nevertheless, it has been a formidable problem to extract useful data and integrate them to obtain answers to objective questions, such as “what is the best drug target in a pathogen proteome?” Hence, the objective of this work is to develop a decision support system for strategic selection of the most promising drug targets to combat tropical diseases. First, we will design a dimensional model for the construction of a Data Warehouse (DW) to allow for the efficient integration of diverse sources of structured and non-structured data. Then, an ETL (extraction, transformation and loading) process will be specified to feed the DW. ProtozoaDB, will be used as a staging area for storing new data on gene essentiality and protein druggability of five protozoa species responsible for important tropical diseases, such as malaria, leishmaniasis and Chagas disease. Because of the lack of experimental data on gene essentiality for the protozoa species, we will use homology to reuse data from model-organisms. Similarly, druggability will be inferred from homologous proteins targeted by chemicals listed in DrugBank, PubChem or BindingDB. Other approach will be information
retrieval from textual databases, using semantic annotation with multiple ontologies. Finally, analytical queries will be performed and genes of parasites with the highest success potential for drug development will be identified.

P77: CONCATABOMINATIONS: A CHARACTER RECODING STRATEGY IMPLEMENTED IN THE SAFE TAXONOMIC REDUCTION APPROACH FOR IDENTIFYING UNSTABLE TAXA

Presenter: KAREN SIU-TING SALVATIERRA, National University of Ireland Maynooth
Authors: Karen Siu-Ting, Christopher Creevey, Mark Wilkinson, Davide Pisani

Missing data is one of the sources of uncertainty in the resolution of nodes in phylogenetic trees. Under parsimony, the strategy of excluding the poorly known taxa has been commonly used to minimize the number of most parsimonious trees and maximise resolution. In this sense, Wilkinson’s (1995) Safe Taxonomic Reduction is useful to determine the potential instability of taxa a priori, and to determine which taxa are safe to exclude from an analysis. In the present work, we introduce a new step into this strategy called “Concatabomination”. This is based on the recoding of pairs of taxa that are found to be “potential taxonomic equivalents” but that still do not have enough data to place them unequivocally in a tree. By recoding potential equivalents we create a new “concatabominated” taxon that is replaced in the original matrix. A compatibility based approach is then used to determine whether two “potential equivalents” are true equivalents and hence if one can be safely deleted. This new strategy was implemented in a pipeline and tested using simulated matrices of non-conflicting characters and then in gap-rich paleontological datasets. The concatabominations were able to reduce the number of unstable taxa and as a result improved resolution in trees. We predict this approach can be useful in supertree analyses, when dealing with gap-rich phylogenomic datasets and to study effective overlap in multigene datasets.

P78: INVESTIGATION INTO CONSTRUCTION OF MHC LIGAND PREDICTORS AND IMPLICATIONS OF PROTOCOLS USED TO EVALUATE THEM

Presenter and Author: WERNER SMIDT, University of Pretoria

Prediction of MHC ligands in the field of Computational Immunology is pivotal in vaccine design strategies concerning CTL epitopes. Prediction tools have been developed over the last few decades with variable accuracies. Independent performance analysis of available MHC ligand predictors are often inconsistent between studies. Usually, cross-validation is performed to evaluate constructed tools. It will be shown here that random cross-validation is inconsistent with resampling. Using BLOSUM distances between peptides and standard clustering procedures in an MHC IC50 value dataset, a novel method has been developed to construct a cross-validation set. Constructing an Artificial Neural Network based MHC ligand predictor, it will be shown how removing certain training
features can improve prediction. Feature removal also improves predictor performance and can be important when making MHC ligand predictions of large peptide sets predictions for multiple HLA allotypes. Creating performance evaluation protocols that tend towards being independent of dataset is crucial in MHC ligand predictor construction, especially when datasets with vastly different peptides are used during construction of these networks.

**P79: MINING FOR CLASS-SPECIFIC MOTIFS IN PROTEIN SEQUENCE CLASSIFICATION**

*Presenter: SATISH SRINIVASAN, University of Nebraska Medical Center*

*Authors: Satish M. Srinivasan, Suleyman Vural, Chittibabu Guda*

Protein sequence data continue to become available at an exponential rate. Annotation of functional and structural attributes of these data lags far behind, with only a small fraction of the data understood and labeled by experimental methods. Automated classification methods that are based on supervised learning can significantly increase the overall accuracy of a method in many domains. While classification is important, identification of the sequence elements that can precisely discriminate between classes is a more interesting scientific question. Previously, we have developed a Bayesian supervised model for classifying protein sequence data using an n-gram approach. Discriminative n-grams are short peptide sequences that are highly frequent in one class but are either minimally present or absent in other classes. In this study, we present a scoring function for identifying discriminative n-grams that are highly specific to a class. We mapped these discriminative n-grams back to the protein sequences to obtain contiguous n-grams that represent short class-specific motifs in protein sequences. Since discriminative n-grams are striking features in a particular domain, classification of protein sequence data based on these short peptide sequences can help in functional and structural annotation of new protein sequences. To validate the performance of our scoring function, we are validating our enriched set of short peptide sequences against the functionally important motifs obtained from the ELM (Eukaryotic Linear Motif) and PROSITE databases. We anticipate to successfully apply this method to many classification problems such as predicting protein subcellular localization and identifying family-specific motifs in protein families.

**P80: RELATIONSHIPS BETWEEN POSITIONING OF SYNONYMOUS CODON AND GENE TRANSLATION EFFICIENCY IN BACTERIA**

*Presenter: SATOSHI TAMAKI, Keio University*

*Authors: Satoshi Tamaki, Masaru Tomita*

Amino acid in the protein peptide is coded by codons, consist of three DNA bases. While Amino acid is carried by tRNA recognizing each codons, most of amino acids has multiple tRNA partners. The frequency of occurrence of synonymous codons are not equal, and this codon usage bias is known to
influence gene transcription, translation, and expression levels. From the recent studies targeting Saccharomyces cerevisiae and other eukaryotes, it was shown that the closest same amino acid coded inside the coding sequence favors codons recognized by the same tRNA. The codon co-occurrence showed high correlation with the gene translation speed, followed by the tRNA recycling model during translation. In this study, we have confirmed the codon co-occurrence among bacterial genomes. From the exhaustive analysis using 696 bacterial genomes, high-codon correlation was observed in most of the bacterial genome. Spices, which had both entries in UniProt Knowledge Base complete proteome sets and tRNADB-CE database, was selected for this analysis. The tendency of codon co-occurrence was observed not only in frequently used codons, but also in rarely used codons. By comparing the GCSI (GC Skew Index) and mean TPI (tRNA Pairing Index) of the genome, codon-correlation was significantly observed in species with low GC skew spectrum intensity (GCSI < 0.1). Our result indicates the possibility of tRNA recycling during translation in bacteria, and will give new insight to the relationships between gene expression and it’s coding sequence.

P81: THE RAT GENOME CURATION: RGD AUTOMATED DATA INTEGRATION PIPELINES MAXIMIZE COVERAGE

Presenter: MAREK TUTAJ, Medical College of Wisconsin
Authors: Marek Tutaj, Mary Shimoyama, Elizabeth Worthey, Jennifer Smith, Howard Jacob

In model organism curation, relying solely on manual curation of the literature is impractical and would result in functional information for only a small portion of the genome. RGD uses a combination of targeted literature curation and a network of automated pipelines to provide comprehensive functional coverage of the rat genome. Quality control processes employed in the pipelines help identify conflicts, omissions and questionable relationships among data originating at other sources and data already in RGD. Conflict reports are sent to curators to resolve data issues. Through these pipelines, RGD 1) integrates and matches genes and gene models from multiple sources with links to sequence data, 2) identifies and loads orthologs and creates ortholog relationships, 3) identifies genes requiring nomenclature review and provides provisional nomenclature, 4) adds and updates genomic positions for rat, human and mouse genes, 5) updates multiple ontologies and identifies obsoleted terms and annotations, 6) provides experimentally determined human and mouse ortholog Gene Ontology annotations for rat genes, 7) provides identifiers and links for major protein information. RGD’s pipelines use a variety of data loading strategies, including drop-and-reload and incremental updates. Multiple quality control processes run in parallel on multicore machines allow dramatic reduction of total running time of the pipelines while providing the desired level of quality checking and high throughput of the data processed. This sophisticated data pipeline network allows RGD to provide comprehensive genome-wide functional and structural information with weekly updates.
P82: USING TRANSPOSABLE ELEMENTS TO UNDERSTAND GENOMIC EVOLUTION

Presenter: VIJETHA VEMULAPALLI, University of Colorado Denver
Authors: Vijetha Vemulapalli, David D Pollock

Transposable elements (TEs) form a large fraction of eukaryotic genomes and play a major role in genome evolution. TEs have been shown to have a significant impact on evolution of regulatory mechanisms, large-scale genomic rearrangements and disease. In order to understand the role of TEs in overall genomic evolution, it is extremely important to first accurately estimate the timing of expansion of various families of TEs. Yet, current methods of estimating the age distribution of TEs do not account for factors that can significantly confound age estimates. We have developed a Bayesian method to estimate the substitution process and the age distribution of TEs accurately. Human Alu elements have been analyzed using this method to get a more detailed assessment of the evolutionary history of these elements in the genome. A large number of copies of Alu elements are present in the Human genome, presumably evolving neutrally and hence make an excellent model for detailed inference of neutral substitution processes and the effect of genomic context on the substitution process. We have explored the effect of local GC content on the substitution process using Alu elements in the Human genome.

P83: DETERMINING THE MINIMAL FUNCTIONAL ELEMENTS IN PROTEIN FAMILIES

Presenter: SULEYMAN VURAL, University of Nebraska
Authors: Suleyman Vural, Satish M. Srinivasan, Chittibabu Guda

Our aim in this project is to efficiently identify evolutionarily conserved core functional elements of all protein families in the public databases. We developed a new computational method that determines the specific and minimal functional elements for protein families using a Bayesian n-gram approach without having a need of multiple sequence alignments, which are the basis for building the models of most of the known protein families. Our alignment-free approach works with sequences that have little or no local sequence identity which is a desirable feature over the existing multiple sequence alignment based methods that work well only with sequences with at least a decent identity. The method efficiently determines the family specific discriminatory n-grams where n is considered to be from 4 to 8. Then family specific n-grams will be merged to obtain longer contiguous regions that represent the core functional elements of a given family. Our results indicate that a set of 50 families is sufficient for model building and extracting most of the core elements of a given family. Considering the extensive size of input databases we are currently parallelizing the code using MPI (Message Passing Interface) libraries. The project has a potential to identify the core and minimal functional motifs to better represent protein families.
P84: MACHINE APPROACHES TO RECOGNITION OF TRYpanosomal Variant Surface Glycoprotein Sequences

Presenter and Author: JON WILKES, Wellcome Trust Centre for Molecular Parasitology

Support vector machines (SVMs) attempt classification of two or more populations of elements by encoding properties of each element as multi-dimensional vectors and defining a hyper-plane(s) separating populations. An SVM’s efficacy in classifying samples depends upon appropriate numerical transformations. We have developed software to rapidly prototype methods to codify aspects of sequence architecture (SVMseq) and generate SVMs. African trypanosomes evade host immune responses by expression of variant surface glycoproteins (VSGs), shielding invariant components of the cell surface. Highly immunogenic, they randomly switch between forms enabling sub-populations expressing different VSGs to expand under immune selection. Extensive sub-telomeric arrays of genes and pseudogenes exist and undergo splicing to generate an unlimited range of novel genes, driving antigenic variation. The VSG repertoires of trypanosome populations will, of necessity, show little sequence similarity; novel techniques for identification and classification of putative VSG (pseudo)genes are required. We applied the software tools of SVMseq to create efficient SVM models for the detection of VSG sequence from canonical sequences derived from 9 of the 11 mega-chromosomes of Trypanosoma brucei Treu-927. Canonical sequences from the remaining 2 chromosomes were used to benchmark performance. We developed a 2 stage method involving the identification of VSG arrays, followed by identification of the N- and C-terminal domains of the component genes. ROC analysis of performance produced AUC values of >0.96. Application of the SVM model to unannotated genomic sequence identified arrays of novel genes.

P85: STOP AND DEFOG: WEB APPLICATIONS FOR A COMPREHENSIVE FUNCTIONAL GENE SET ANALYSIS

Presenter: TOBIAS WITTKOP, Buck Institute for Research on Aging
Authors: Tobias Wittkop, Emily Teravest, Ari E. Berman, Uday Evani, K. Mathew Fleisch, Corey Powell, Nigam Shah, Sean D. Mooney

One of the most common outcomes of high-throughput biological experiments is a list of genes or proteins of interest. In order to explain the observed changes of these specific genes and to create new hypotheses one needs to understand the functions and roles of the genes in the lists under the condition studied in the experiment. Here we present two novel applications that facilitate the extraction of functional content of lists of genes. With STOP we overcome the limitations of manually annotated ontologies (like GO) and use automatic annotation techniques using descriptive text as an annotation source. This
resource utilizes the National Center for Biomedical Ontology’s database, which includes over 200 biomedically related ontologies. These automatic annotations are then processed and used for enrichment analyses against submitted gene lists. In recent studies, we demonstrated how accurate these automatic annotations can be. STOP ranked 7th in a comparison of functional annotation tools when trying to predict novel annotations in the biological process category of GO. Our second application, DEFOG, eases the functional analysis of gene sets by hierarchically organizing them into functional related groups using data fusion of high-throughput experimental data. The underlying computational pipeline utilizes the state-of-the-art applications GeneMANIA, Transitivity Clustering, and Ontologizer for gene set specific network fusion, non-agglomerative hierarchical clustering, and GO term enrichment respectively. DEFOG allows for a novel visual analysis of gene sets that aids in the discovery of potentially important biological mechanisms and assists in the generation of new hypotheses from gene lists.

P86: BUILDING AN INTERACTOME TO IDENTIFY SIGNALING COMPONENTS

Presenter: SARAH WYATT, Ohio University
Authors: Sarah Waytt, Kaiyu Shen

A genomic-level analysis was utilized to study the gravitropic signal transduction in Arabidopsis. First, the gravity persistent signal (GPS) treatment was used to isolate the mechanisms of gravitropic signal transduction from those of response mRNA was extracted across a time course during the GPS treatment and probed against an Agilent Arabidopsis 4X44k gene expression array. Instead of analyzing the data to find the “top statistically significant” genes, we used the raw data to construct a gravitropic signaling-specific interactome. First, an Arabidopsis gene interaction network was built from five databases: PAIR—Experimentally reported interaction, PAIR — High coverage interactions, TAIR, AtPIN, BioGRID. Next, a collection of expression profile data sets, representing potential background parameters, were obtained from TAIR and NASC databases to be used as filters. Pearson Correlation Coefficient (PCC) values were calculated for the GPS microarray data and the background parameters, and the original interactome was then reduced: gene interactions were kept only if there was a significant PCC value in the GPS data but not in the background parameter profiles. To further fine tune and train the interactome, text-mining of the literature was performed and functional annotations were collected and applied. These analyses generated an interactome that was specific for gravitropic signal transduction by adding/deleting interactions as well as assigning probabilities for the interactions using a Bayesian network approach. Thus, the important gravitropic signal transduction genes, along with the functional clusters and crucial pathways, could be selected for biological validation and study.
**P87: POST-TRANSLATIONAL MODIFICATIONS INDUCE SIGNIFICANT YET NOT EXTREME CHANGES TO PROTEIN STRUCTURE**

*Presenter: FUXIAO XIN, Indiana University Bloomington*
*Authors: Fuxiao Xin, Predrag Radivojac*

Anecdotal studies have shown that post-translational modifications (PTMs) can induce structural rearrangements of their target proteins, but the extent of the structural impact has not been systematically characterized. In this study, we take available structure data from Protein Data Bank (PDB) and use statistical methods to provide the general trend of the structural impact of PTMs. We primarily focus on two most abundant PTMs in PDB, glycosylation and phosphorylation, but show that acetylation and methylation, which can be found in a smaller number of structures, have similar tendencies. We exploited redundancy in PDB and, using root mean-square deviation (RMSD), compared structures of identical protein chains in their modified and unmodified forms. Our results provide evidence that PTMs induce significant protein conformational changes at both local and global level. Specifically, phosphorylation increases global RMSD by 52% and local RMSD by 137%, while glycosylation increases global RMSD by 40% and local RMSD by 35%. However, the proportion of extreme changes is small: at the global level, only 7% of glycosylated and 13% of phosphorylated proteins undergo changes that are greater than 2Å. Similarly at the local level, an increase of RMSD greater than 0.5Å upon glycosylation and phosphorylation occurs in only 8% and 20% of cases, respectively. Further comparison of RMSDs between modified structures with RMSDs between unmodified structures shows no significant difference for glycosylation, but suggests that phosphorylation stabilizes protein structure by reducing global RMSD by 50%.

**P88: CO- EVOLUTION OF SP1 AND DNA BINDING SITES**

*Presenter and Author: KEN YOKOYAMA, University of Colorado Denver*

Regulatory proteins bind specifically to hundreds of sites across the genome, and are therefore thought to rarely change binding specificity. Contradicting this, we show that transcription factor SP1 motif preferences evolved convergently in eutherian mammals and birds, causing coevolutionary changes throughout ~800 regulatory regions. Structural and phylogenetic evidence implicates a single causative amino acid replacement at the same SP1 position along both lineages. Furthermore, multiple convergent events subsequently occurred at the homologous position in co-regulatory paralogs SP3 and SP4, which bind to SP1 sites. Theoretical models suggest that regulatory system transformations, such as seen in SP1, are eminently plausible in diploids, and shifts in the regulatory element birth/death rates are well predicted. The unprecedented scale of co-evolution and convergence observed here strongly implicates positive selection.
P89: COexpression of Linked Genes Is Not Explained by Similar Chromatin Environments

Presenter: YIQIANG ZHAO, Buck institute
Authors: Yiqang Zhao, Sean Mooney

Neighboring genes in the genome tend to be coexpressed and it holds true when the similarity of transcription regulation is controlled for. Thus, it is hypothesized that this is because adjacent genes share similar chromatin environments. By using a distance-based method, we found clustering of genes with similar epigenetic states, as determined by histone modifications. This clustering can be extended to multiple tissues (co-modification). These clusters are usually larger than 10 genes, which is nearly three times larger than the observed clusters derived from co-expression or co-function, which are also measured by our method. Interestingly, by calculating the correlation coefficient between them, we found that co-modification, co-expression and co-function are independent. Our result suggests there is a component of the clustering of gene co-expression that cannot be explained by the current observations.

P90: What’s New at EcoGene

Presenter: JINDAN ZHOU, University of Miami Miller School of Medicine
Authors: Jindan Zhou, Kristi L. Jones, Kenneth E. Rudd

EcoGene is a database dedicated to annotation of the E. coli K-12 genome. The recent EcoGene 3.0 web interface is powered by the Drupal open source content management platform and is supported by a MySQL database. We describe several new features for EcoTopics, EcoSearch and EcoArray. EcoArray is a curated E. coli Gene Expression Database easily accessible using PubMed identifiers, experimental conditions or the regulated genes. The data sources for EcoArray are GEO records and publication supplemental materials. Geometadb is used to retrieve GEO metadata describing details we use for assembling and annotating expression ratios. PHP and R scripts were written to parse, extract, assemble, and reformat experimental data into MySQL tables for querying and analysis of expression profiles. Genesets are groups of genes and proteins, including regulons identified from experiments in EcoArray. Information describing E. coli is not limited to individual genes; this is accommodated by clustering many genesets within EcoTopics. These include high-throughput results datasets from shotgun proteomics, genetic screens, and molecular interaction studies. Genesets can also be derived from bioinformatics analyses, such as the identification of paralogous families. Gene Query includes an interactive Venn diagram as the graphic interface to access Boolean query results in table format. Additional EcoGene features include dynamically generated GeneMaps on every GenePage, with intergenic region details and customized restriction sites. MapSearch, a restriction map alignment utility, and PrintMap, a PDF map drawing program, have been renovated for EcoGene. Finally, PrimerPair enables the genome scale design of PCR primers for deletions and clones.
IBM’s Deep Computing organization is the high performance computing organization in IBM Systems and Technology Group. This group is responsible for the strategy, marketing and identification of areas that can benefit from IBM’s high end technology. The life sciences is such an area, and IBM is and will continue to bring valued solutions to life sciences.

IBM’s Research Division is a partner with IBM’s Deep Computing organization, developing the next generation of high performance computers. In addition, the Research Division has many groups investigating numerous application areas in collaboration with IBM’s customers and partners. This includes IBM’s Computational Biology Center with IBM’s new Computational Science Center.

NEC Corporation (TSE: 6701) is one of the world’s leading providers of Internet, broadband network and enterprise business solutions dedicated to meeting the specialized needs of its diverse and global base of customers.

NEC delivers tailored solutions in the key fields of computer, networking and electron devices, by integrating its technical strengths in IT and Networks, and by providing advanced semiconductor solutions through NEC Electronics Corporation. The NEC Group employs more than 140,000 people worldwide. For additional information, please visit the NEC home page at: www.nec.com.
SomaLogic, Inc.
2945 Wilderness Pl.
Boulder, CO 80301
http://www.somalogic.com
information@somalogic.com

SomaLogic, Inc. is a privately-held biomarker discovery and clinical proteomics company based in Boulder, Colorado. The company’s mission is to use its proprietary Slow-Offrate Modified Aptamer (“SOMAmer”) technology to develop enhanced protein analysis tools and reagents for the life sciences community, to facilitate target validation, and to develop and commercialize clinical diagnostic products that will improve the delivery of healthcare by offering timely and accurate diagnostic information to physicians and their patients. Further information about SomaLogic can be found at www.somalogic.com.

Silver Sponsor

Biodesix, Inc.
520 Zang Street, Suite 213
Broomfield, CO 80021

Heinrich Röder, D. Phil. hroder@biodesix.com
www.biodesix.com

Biodesix is a personalized medicine company focused on the development of diagnostic products that inform treatment decisions and improve patient care. The Company’s goal is to give physicians more information about the patient and their disease; understanding the clinically meaningful information contained within each patient’s molecular profile leads to better care and better outcomes. The Company’s unique approach is based on ProTS®, proprietary technology which exploits the power of mass spectrometry and enables the discovery of specific molecular profiles. Biodesix collaborates with clinical investigators to address critical clinical questions, and partners with biotechnology and pharmaceutical companies to develop diagnostics to select patients most likely to benefit from novel therapies. For more information about Biodesix, please visit www.Biodesix.com.

Book Donations for Student Prizes

Springer
Cambridge University Press
Rocky 2011 is an official conference of the International Society for Computational Biology.

Rocky 2011 is supported by the Computational Bioscience Program at the University of Colorado School of Medicine.

Conference Chair
Lawrence Hunter, PhD
University of Colorado Denver
School of Medicine