

THURSDAY DECEMBER 10

* All conference scientific sessions will be held at the Silvertree Hotel

11:00 AM – 1:00 PM	REGISTRATION	<i>CABARET LOBBY</i>
1:00 PM – 1:45 PM	KEYNOTE 1	<i>CABARET ROOM</i>
Polypharmacology: Drug Discovery in the Era of Genomics and Proteomics <i>Philip E. Bourne, PhD Professor, Skaggs School of Pharmacy & Pharmaceutical Sciences, UCSD Associate Director, RCSB Protein Data Bank Editor in Chief, PLoS Computational Biology</i>		
1:45 PM – 2:45 PM	ORAL PRESENTATIONS 1–6	<i>CABARET ROOM</i>
2:45 PM – 3:00 PM	BREAK (15 MINUTES)	<i>CABARET LOBBY</i>
3:00 PM – 4:10 PM	ORAL PRESENTATIONS 7–13	<i>CABARET ROOM</i>
4:10 PM – 4:25 PM	BREAK (15 MINUTES)	<i>CABARET LOBBY</i>
4:25 PM – 5:10 PM	KEYNOTE 2	<i>CABARET ROOM</i>
Knowledge Acquisition for Knowledge Discovery in Cancer Metastasis <i>Anna Divoli, PhD Postdoctoral Scholar, Department of Medicine and Institute of Genomics and Systems Biology, University of Chicago</i>		
5:10 PM – 6:20 PM	ORAL PRESENTATIONS 14–20	<i>CABARET ROOM</i>
7:00 PM – 9:00 PM	BANQUET	<i>IL POGGIO RESTAURANT, SNOWMASS VILLAGE</i>

FRIDAY DECEMBER 11

* All conference scientific sessions will be held at the Silvertree Hotel

9:00 AM – 9:45 AM	KEYNOTE 3	<i>Cabaret Room</i>
Spanning Scales – The Combination of Mathematics and High Performance Computing impacts Computational Biology <i>Kirk E. Jordan, PhD, Emerging Solution Executive, Computational Science Center, IBM T.J. Watson Research Center</i>		
9:45 AM – 10:45 AM	ORAL PRESENTATIONS 21–26	<i>Cabaret Room</i>
10:45 AM – 11:00 AM	BREAK (15 MINUTES)	<i>Cabaret Lobby</i>
11:00 AM – 12:00 PM	ORAL PRESENTATIONS 27–32	<i>Cabaret Room</i>
12:00 PM – 4:00 PM	BREAK	

4:00 PM – 4:45 PM	KEYNOTE 4 (JOINT KEYNOTE)	<i>Cabaret Room</i>
Strategies for Elaborating Cognitive Requirements of Bioinformatics Tools <i>Ben Keller, PhD, Associate Professor, Computer Science Department, Eastern Michigan University and Barbara Mirel, Associate Research Scientist, School of Education, University of Michigan, NCIBI Core Director of Evaluation, Education and Training</i>		
4:45 PM – 5:45 PM	ORAL PRESENTATIONS 33–38	<i>CABARET ROOM</i>
5:45 PM – 8:00 PM	RECEPTION AND POSTER SESSION	<i>ELDORADO ROOM (3RD FLOOR)</i>

SATURDAY DECEMBER 12

* All conference scientific sessions will be held at the Silvertree Hotel

9:00 AM – 9:45 AM	KEYNOTE 5	<i>CABARET ROOM</i>
ChEMBL – Large-scale Open Access Data for Drug Discovery <i>John P. Overington, PhD CChem., European Bioinformatics Institute (EMBL-EBI)</i>		
9:45 AM – 10:35 AM	ORAL PRESENTATIONS 39–43	<i>CABARET ROOM</i>
10:35 AM – 12:00 PM	POSTER SESSION	<i>ELDORADO ROOM (3RD FLOOR)</i>
12:00 PM – 4:00 PM	BREAK	
4:00 PM – 5:00 PM	ORAL PRESENTATIONS 44–49	<i>ELDORADO ROOM (3RD FLOOR)</i>
5:00 PM – 5:45 PM	KEYNOTE 6	<i>ELDORADO ROOM (3RD FLOOR)</i>
Informatics Challenges for Pharmacogenetic <i>Russ B. Altman, MD, PhD, Professor of Bioengineering Genetics, Medicine (& Computer Science, by courtesy), Chair, Bioengineering Director, Biomedical Informatics Training Program, Stanford University</i>		
5:45 PM	ROCKY '09 CLOSING COMMENTS	<i>ELDORADO ROOM (3RD FLOOR)</i>

THURSDAY, DECEMBER 10

* All conference scientific sessions will be held at the Silvertree Hotel

11:00 AM – 1:00 PM REGISTRATION CABARET LOBBY

1:00 PM – 1:45 PM KEYNOTE 1 CABARET ROOM

Polypharmacology: Drug Discovery in the Era of Genomics and Proteomics

Philip E. Bourne, PhD Professor, Skaggs School of Pharmacy & Pharmaceutical Sciences, UCSD Associate Director, RCSB Protein Data Bank Editor in Chief, PLoS Computational Biology

1:45 PM – 2:45 PM ORAL PRESENTATIONS 1–6 CABARET ROOM

OP 1: Helping Biologists Understand their Data:
An Update on the Hanalyzer System

Presenter: William A Baumgartner Jr, University of Colorado, Denver
Authors: William A Baumgartner Jr, Hannah Tipney, Lawrence Hunter

OP 2: Thermodynamics-inspired ncRNA Search

Presenter: Jennifer Smith, Boise State University
Authors: Jennifer Smith, Pamila Ward

OP 3: Algorithm to Improve Gene Consistency Across Bacterial Genomes

Presenter: Judith D. Cohn, Los Alamos National Laboratory
Authors: Judith D. Cohn, Michael E. Wall, John Dunbar

OP 4: A Genome-wide Analysis of Poised Promoters in Bacteria

Presenter and Author: Marko Djordjevic, Arkansas State University & The Arkansas Biosciences Institute

OP 5: Structure-Based Prediction of DNA Binding Sites for Families of Transcription Factors

Presenter and Author: Julia Ponomarenko, University of California, San Diego

OP 6: Genome-wide Discovery of Human Heart Enhancers

Presenter: Ivan Ovcharenko, NIH
Authors: Leelavati Narlikar, Noboru Sakabe, Alexander Blanski, Fabio Arimura, Marcelo Nobrega, Ivan Ovcharenko

2:45 PM – 3:00 PM BREAK (15 MINUTES) CABARET LOBBY

DETAILED AGENDA

AGENDA

3:00 PM – 4:10 PM

ORAL PRESENTATIONS 7–13

CABARET ROOM

OP 7: Development of Methods for Integrating Diverse Sources of Genome-Scale Data

Presenter: Daniel Dvorkin, University of Colorado, Denver

Authors: Daniel Dvorkin, Katerina Kechris

OP 8: Visualizing Genomic Sequences in 2D

Presenter and Author: Josiah Seaman, Colorado State University

OP 9: Automated Inference of Molecular Mechanisms of Disease from Amino Acid Substitutions

Presenter: Biao Li, Indiana University

Authors: Biao Li, Vidhya G. Krishnan, Matthew E. Mort, Fuxiao Xin, Kishore K. Kamati, David N. Cooper, Sean D. Mooney, Predrag Radivojac

OP 10: A User Study of Attribute Visualization Tools and Their Role in Understanding Biological Networks

Presenter: Hande Kucuk, Eastern Michigan University

Authors: Hande Kucuk, Benjamin J. Keller, Terry Weymouth, Barbara Mirel

OP 11: NeuroIE: Extracting Neuroimaging Study Results from the Literature

Presenter: Yong Gao, Mass General Hospital/Harvard Medical School

Authors: Yong Gao, Dave Kennedy

OP 12: Analysis of a Local Huntingtin Protein Interaction Network

Presenter: Corey Powell, Buck Institute for Age Research

Authors: Corey Powell, Robert Hughes, Cendrine Tourette, Russell Bell, Sean Mooney

OP 13: In Silico Functional Profiling of Human Disease-Associated and Polymorphic Amino Acid Substitutions

Presenter: Vidhya G. Krishnan, Buck Institute for Age Research

Authors: Matthew Mort, Uday S. Evani, Vidhya G. Krishnan, Kishore K. Kamati Peter H. Baenziger, Anghuman Bagchi, Brandon Peters, Rakesh Sathyesh, Biao Li, Yanan Sun, Bin Xue, Nigam Shah, Maricel Kann, David N. Cooper, Predrag Radivojac, Sean D. Mooney

4:10 PM – 4:25 PM

BREAK (15 MINUTES)

CABARET LOBBY

4:25 PM – 5:10 PM KEYNOTE 2 CABARET ROOM

Knowledge Acquisition for Knowledge Discovery in Cancer Metastasis
Anna Divoli, PhD Postdoctoral Scholar, Department of Medicine and Institute of Genomics and Systems Biology, The University of Chicago

5:10 PM – 6:20 PM ORAL PRESENTATIONS 14–20 CABARET ROOM

OP 14: Accelerating Candidate Gene Discovery through Ontological Indexing of Large Scale Data Repositories

Presenter: Simon Twigger, Medical College of Wisconsin
Authors: Simon Twigger, Joey Geiger, Jennifer Smit

OP 15: Knowledge Network Approach: Pathways and Drugs

Presenter: Nikolai Daraselia, Ariadne
Authors: Nikolai Daraselia, Ekaterina Kotelnikova, Anton Yuryev

OP 16: Can We Accurately Determine the Fittest Genes in Nature

Presenter: Ramy K. Aziz, San Diego State University
Authors: Ramy K. Aziz, Mya Breitbart, Robert Edwards

OP 17: Assessing Models of Protein Interaction Network Evolution

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

OP 18: Gene and Genome Trees Conflict at Many Levels

Presenter: Leanne S. Haggerty, NUI Maynooth
Authors: Leanne S. Haggerty, Fergal J. Martin, David A. Fitzpatrick, James O. McInerney

OP 19: Protein-protein Interactions are Driven by Functional Evolution

Presenter: Yiqiang Zhao, Buck Institute for Age Research
Authors: Yiqiang Zhao, Sean Mooney

OP 20: Evolutionary Study and Prediction of Protein-protein Interactions in Chromatin Modification Complexes

Presenter: Xuejian Xiong, Hospital for Sick Children
Authors: Xuejian Xiong, Tuan On, Shuye Pu, Andrei Turinsky, Yunchen Gong, Andrew Emili, Zhaolei Zhang, Jack Greenblatt, Shoshana J. Wodak, John Parkinson

7:00 PM–9:00 PM BANQUET IL POGGIO RESTAURANT, SNOWMASS VILLAGE

FRIDAY DECEMBER 11

* All conference scientific sessions will be held at the Silvertree Hotel

AGENDA

9:00 AM – 9:45 AM KEYNOTE 3 CABARET ROOM

Spanning Scales – The Combination of Mathematics and High Performance Computing impacts Computational Biology

Kirk E. Jordan, PhD, Emerging Solution Executive, Computational Science Center, IBM T.J. Watson Research Center

9:45 AM – 10:45 AM ORAL PRESENTATIONS 21–26 CABARET ROOM

OP 21: An Individual Based Modelling Approach to Studying the Evolution of Mate Choice Strategy

Presenter and Author: Robert Williamson, Rose-Hulman Institute of Technology

OP 22: MtHaplogroups: A Curated Web Resource for Mitochondrial Variation

Presenter: Michael V. Osier, Rochester Institute of Technology

Authors: Kyle Dewey, Eric Stevens, Dina L. Newman, Michael V. Osier

OP 23: Posterior-Predictive Detection Of Molecular Co-Evolution Using Phylogenetically-Integrated Mutual Information

Presenter: A.P. Jason de Koning, University of Colorado, Denver

Authors: A.P. Jason de Koning, Todd A. Castoe, Hyun-min Kim, David D. Pollock

OP 24: Bioinformatics Characterization Of The Plasmodium Glutathione S-Transferase

Presenter: Emilee Colón, University of Puerto Rico

Authors: Emilee Colón, Adelfa Serrano, Hugh Nicholas Jr, Troy Wymore, Alexander Ropelewski, Ricardo González Méndez

OP 25: Leveraging Existing Biological Knowledge in Genome Wide Association Studies

Presenter: Ronald Schuyler, University of Colorado, Denver

Authors: Ronald Schuyler, Lawrence Hunter, Deborah Glueck

OP 26: Evaluating Microbial Diversity and Adaptation: New Opportunities for Insight in a Data Rich World

Presenter: Catherine Lozupone, University of Colorado, Boulder

Authors: Catherine Lozupone, Rob Knight

10:45 AM – 11:00 AM BREAK (15 MINUTES) CABARET LOBBY

11:00 AM – 12:00 PM ORAL PRESENTATIONS 27–32 CABARET ROOM

OP 27: Wow! That’s Really Interesting!

Presenter: Hannah Tipney, University of Colorado, Denver

Authors: Hannah Tipney, Lawrence Hunter

OP 28: Global Optimization in Nonlinear Models: Optimization and Evolution in Metabolic Networks

Presenter: Albert Sorribas, CMB – IRBLLEIDA – University of Lleida

Authors: Albert Sorribas, Carlos Pozo, Gonzalo Guillén-Gosálbez, Laureano Jimenez, Rui Alves

OP 29: Virus Discovery by Deep Sequencing and Assembly of Virus-derived Small Silencing RNAs

Presenter: Qingfa Wu, UC Riverside

Authors: Qingfa Wu, Yingjun Luo, Rui Lu, Nelson Lau, Eric C Lai, Wan-Xiang Li, Shou-Wei Ding

OP 30: Prediction of Protein Protein Interaction Sites and Their Impact on Genetic Disease

Presenter: Angshuman Bagchi, Buck Institute for Age Research

Authors: Angshuman Bagchi, Eunseog Youn, Matthew E. Mort, David N. Cooper, Sean D. Mooney

OP 31: Optimal Nearest Shrunken Centroids Method for Highdimensional Data Classification

Presenter: Tiejun Tong, University of Colorado, Boulder

Authors: Tiejun Tong, Herbert Pang

OP 32: Retrobiosynthesis: Searching Backwards, Looking Forwards

Presenter: Dan McShan, University of Colorado, Denver

Authors: Dan McShan

12:00 PM – 4:00 PM BREAK CABARET LOBBY

4:00 PM – 4:45 PM KEYNOTE 4 (JOINT KEYNOTE) CABARET ROOM

Strategies for Elaborating Cognitive Requirements of Bioinformatics Tools

Ben Keller, PhD, Associate Professor, Computer Science Department, Eastern Michigan University and Barbara Mirel, Associate Research Scientist, School of Education, University of Michigan, NCIBI Core Director of Evaluation, Education and Training

4:45 PM – 5:45 PM

ORAL PRESENTATIONS 33–38

CABARET ROOM

OP 33: Experiments with Biological Concept Recognition Tools

Presenter: Karin Verspoor, University of Colorado, Denver

Authors: Karin Verspoor, Kevin B. Cohen, Helen Johnson, Christophe Roeder, Cesar Mejia, William Baumgartner Jr, Larry Hunter

OP 34: Test Suite Design for Biomedical Ontology Concept Recognition Systems

Presenter: K. Bretonnel Cohen, University of Colorado, Denver

Authors: K. Bretonnel Cohen, Karin Verspoor, Christophe Roeder, William A. Baumgartner Jr, Lawrence Hunter

OP 35: Predicting Protein Linkages in Bacteria: Which Method is Best Depends on Task

Presenter: Anis Karimpour-Fard, University of Colorado, Denver

Authors: Anis Karimpour-Fard, Sonia M. Leach, Ryan T. Gill, Lawrence E Hunter

OP 36: A Metagenomic Study of a Microbial Mat in a Hawaiian Lava Cave

Presenter: Gayle K. Philip, NASA Astrobiology Institute

Authors: Gayle K. Philip, Mark V. Brown, Stuart P. Donachie

OP 37: Mobile Metagenomics: Annotating DNA Sequences on a Cell Phone

Presenter: Josh Hoffman, San Diego State University

Authors: Josh Hoffman, Daniel Cuevas, Robert Edwards

OP 38: The Colorado Richly Annotated Full-Text (CRAFT) Corpus: A Resource for BioNLP Research

Presenter: Michael Bada, University of Colorado, Denver

Authors: Michael Bada, Miriam Eckert, Kristin Garcia, Donald Evans, Dmitry Sitnikov, William A. Baumgartner Jr, Philip V. Ogren, Arrick Lanfranchi, Amanda Howard, William Corvey, Nianwen Xue, Kevin B. Cohen, Karin Verspoor, Judith A. Blake, Martha Palmer

5:45 PM – 8:00 PM

RECEPTION AND POSTER SESSION

ELDORADO ROOM
(3RD FLOOR)

SATURDAY DECEMBER 12

* All conference scientific sessions will be held at the Silvertree Hotel

9:00 AM – 9:45 AM **KEYNOTE 5** CABARET ROOM

ChEMBL – Large-scale Open Access Data for Drug Discovery

John P. Overington, PhD CChem., European Bioinformatics Institute (EMBL-EBI)

9:45 AM – 10:35 AM **ORAL PRESENTATIONS 39–43** CABARET ROOM

OP 39: Subgraphs of Protein-protein Interaction Networks with High Connectivity

Presenter: Suzanne Gallagher, University of Colorado, Boulder

Authors: Suzanne Gallagher, Debra Goldberg

OP 40: TreeHugger: A New Test for Enrichment of Gene Ontology Terms

Presenter: Daniel Jupiter, Texas A&M Health Science Center

Authors: Daniel Jupiter, Jessica Sahutoglu, Vincent VanBuren

OP 41: Pattern-Based Extraction of Argumentation from the Scientific Literature

Presenter and Author: Elizabeth White, University of Colorado, Boulder

OP 42: Early Host Response During Influenza Infections

Presenter and Author: Christian V. Forst, University of Texas

OP 43: Boolean Network Models of Human Aging

Presenter: Michael Verdicchio, Arizona State University

Authors: Michael Verdicchio, Seungchan Kim

10:35 AM – 12:00 PM **POSTER SESSION** ELDORADO ROOM (3RD FLOOR)

12:00 PM – 4:00 PM **BREAK**

4:00 PM – 5:00 PM **ORAL PRESENTATIONS 44–49** ELDORADO ROOM (3RD FLOOR)

OP 44: Positive Selection and Functional Shift: Human Heme Peroxidase Case Study

Presenter: Mary J. O’Connell, Dublin City University

Authors: Noeleen B. Loughran, Brendan O’Connor, Ciaran O’Fagain, William M. Nauseef, Mary J. O’Connell

OP 45: Delving into the Bacteriome:Protein-protein Interactions in E. coli

Presenter: John Parkinson, Hospital for Sick Children

Authors: Jose Peregrin-Alvarez, Xuejian Xiong, Chong Su, John Parkinson

OP 46: Simplified Clustering with Dirichlet Process and Other Process Mixtures

Presenter: Matthew Shotwell, Medical University of South Carolina

Authors: Matthew Shotwell, M.S., Elizabeth Slate, Ph.D.

OP 47: Real Time Metagenomics

Presenter: Robert Edwards, San Diego State University

Authors: Robert Edwards, Robert Olson, Terry Disz, Rick Stevens, Ross Overbeek

OP 48: Structure Discovery in PPI Networks Using Pattern-Based Network Decomposition

Presenter: Ying Liu, University of Texas, Dallas

Authors: Ying Liu, Chengcheng Shen, Phil Bachman

OP 49: CANCELLED

5:00 PM – 5:45 PM KEYNOTE 6
Informatics Challenges for Pharmacogenetic

ELDORADO ROOM (3RD FLOOR)

Russ B. Altman, MD, PhD, Professor of Bioengineering Genetics, Medicine (& Computer Science, by courtesy), Chair, Bioengineering Director, Biomedical Informatics Training Program, Stanford University

5:45 PM

ROCKY '09 CLOSING COMMENTS

*ELDORADO ROOM
(3RD FLOOR)*

RUSS B. ALTMAN, MD, PHD

Professor of Bioengineering, Genetics, Medicine (& Computer Science, by courtesy), Chair, Bioengineering Director, Biomedical Informatics Training Program, Stanford University

**Informatics Challenges for Pharmacogenetics**

ABSTRACT: Pharmacogenomics is the study of how human genetic variation impacts drug response phenotypes. We are building the PharmGKB (www.pharmgkb.org/) to catalog all knowledge of gene- drug relationships to support discovery and application of pharmacogenomics. From this effort arise important informatics challenges. I will discuss our work in text mining to extract relationships between drugs and gene variants, our use of this information to create tools for predicting gene-drug interactions, and our efforts building tools to assist in genetic association studies particularly focusing on drug response.

PHILIP E. BOURNE, PHD

Professor, Skaggs School of Pharmacy & Pharmaceutical Sciences, UCSD Associate Director RCSB Protein Data Bank, Editor in Chief PLoS Computational Biology

**Polypharmacology: Drug Discovery in the Era of Genomics and Proteomics**

ABSTRACT: The notion of one drug binding to one receptor to treat one disease becomes questionable as we understand more about the human genome and proteome. Rather we need to consider a collective effect associated with binding to multiple receptors which exist in a variety of different pathways by bringing to bear computational and systems biology. I will illustrate this with a bioinformatics approach^[1] that is akin to reverse engineering the drug discovery process. Rather than take a large library of ligand molecules and screen them against a known protein receptor, we take a known drug-receptor complex and search the human proteome for other proteins with similar binding sites. These off-targets are then mapped to pathways and systems and may explain a side effect of a drug or point to a possible repositioning of that drug to treat a different condition. Biological outcomes to date include repositioning Parkinson's disease drugs to treat TB^[2] and to explain why Torcetrapib failed after 15 years of development and \$850M was spent^[3].

REFERENCES

- [1] L. Xie and P.E. Bourne 2008 Detecting Evolutionary Linkages Across Fold and Functional Space with Sequence Order Independent Profile-profile Alignments. PNAS, 105(14) 5441-5446.
- [2] S.L. Kinnings, N. Buchmeier, N. Liu, P.J. Tonge L. Xie and P.E. Bourne 2009 Discovery of Novel Drug Leads to Treat Multi-drug and Extensively Drug Resistant Tuberculosis by Repositioning Safe Pharmaceuticals: A Chemical Genomics Approach with Subsequent Biological Validation. PLoS Comp. Biol. 5(7) e1000423.

[3] L. Xie, J. Li, L. Xie, and P.E.Bourne 2009 Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network To Explain the Side Effects of CETP Inhibitors, PLoS Comp. Biol. 5(5) e1000387.

ANNA DIVOLI, PHD

Postdoctoral Scholar, Department of Medicine and Institute of Genomics and Systems Biology, The University of Chicago

Knowledge Acquisition for Knowledge Discovery in Cancer Metastasis



ABSTRACT: The clinical importance of understanding cancer metastasis and the complex nature of the process have made it an extremely important research subject. We employ Knowledge Acquisition techniques to organize and represent the existing knowledge as it appears in the heads of experts and unveil any compelling trends and controversies. We interviewed 28 experts on several aspects of metastasis and asked them to share with us their subjective opinions. We were interested in their understanding of metastasis and their possible explanations for not yet scientifically answered research questions and unexplained clinical manifestations, as well as their views on the future of the metastasis research field. Detailed analysis of the interview data reveals areas of agreement but also of disagreement, several known theories' devotees and a few challengers, along with a number of interesting viewpoints and the inevitable introduction of some new questions. Besides the biologically interesting perspectives, we examine the language that experts use while communicating to us their views.

KIRK E. JORDAN, PHD

Emerging Solution Executive, Computational Science Center, IBM T.J. Watson Research Center

Spanning Scales – The Combination of Mathematics and High Performance Computing impacts Computational Biology



ABSTRACT: Computation is playing an ever increasing and vital role in the biological and healthcare sciences. In many instances, scientists are developing mathematical models and using high performance computing to carry out analysis and simulations that provide insight into biological systems. The complexity of these models often demands increasing compute power and sophisticated mathematics for the solution. In collaboration, biological scientists are using thousands of processors to look at their problems in new ways, leading to science breakthroughs. In this talk, I will briefly describe some of the trends we see in high performance computing and some of the challenges looming on the horizon. I then describe a solution in collaboration with colleagues used in investigating

blood perfusion in the brain as an example of a new approach for computational biology. In conclusion, I will point out how this approach is an example of coupling high performance computing and mathematics to tackle multi-scale biological science problems.

JOINT KEYNOTE

BEN KELLER, PHD

Associate Professor, Eastern Michigan University,
Computer Science Department

and

BARBARA MIREL, D. ARTS

Associate Research Scientist, School of Education, University of Michigan,
NCIBI Core Director of Evaluation, Education and Training

Strategies for Elaborating Cognitive Requirements of Bioinformatics Tools

ABSTRACT: Most bioinformatics tools have been designed for isolated problems, and were not built with consideration to the broader context of their use. This fact creates a situation in which the tasks that scientists perform do not map well to the tools available to them. Our challenge as tool designers and developers is to understand how to better engineer these tools to support user cognition for larger tasks such as those in translational systems biology. This talk will discuss the two strategies we have followed to derive cognitive user requirements: first, by synthesizing best practices identified in the literature, and, second, by generalizing from Mirel's extensive fields studies of scientists conducting exploratory analysis with tools developed at the National Center for Integrative Biomedical Informatics. The first draws on discrete, independent observations to create general strategies applicable to broader sets of tools, while the second focuses on tools dealing with gene relationships and scientists' classification and comparison as a prelude to causative reasoning. Our goal is to reflect on how best to apply cognitive engineering to the development of translational bioinformatics and systems biology tools. This application of cognitive engineering includes developing user-centered rationales for requirements. Here we step back to look at and propose specific needs assessment processes that can be used to move incrementally from a model of scientific user cognition to high level user requirements, and ultimately to detailed uses cases and functional specifications for tools. (This work partially supported by NIH grant U54 DA021519.)



JOHN P. OVERINGTON, PHD CCHEM

Team Leader, Computational Chemical Biology, European Bioinformatics Institute (EMBL-EBI)

ChEMBL – Large-scale Open Access Data for Drug Discovery

ABSTRACT: The link between the biological and chemical worlds is of central importance in many fields, not least that of healthcare. For example, a major focus in systems biology research is the signalling networks and pathways describing the interactions and functions of large numbers of genes and proteins. Similarly, within healthcare-related chemistry research there is much interest in efficiently identifying drug-like compounds that specifically interact with these proteins/genes. However there has been relatively little research explicitly directed at understanding the linkages between these two historically distinct domains. Key to our work in this area has been the construction of a large and general structure activity relationship database, linking pharmacological activities of compounds through to their targets, and understanding how particular compounds recognise their cognate receptors. Application of rules derived from these databases leads to rapid, economic, and effective identification of quality target and lead combinations for subsequent pharmacological validation. These data have also been mapped to the currently drugged genome, launched drugs, and large numbers of clinical development candidates. These databases are now in the public domain with the specific aim of enabling new translational research. Current status and future challenges for these informatics resources will be discussed.

Oral Presentations

OP1: HELPING BIOLOGISTS UNDERSTAND THEIR DATA: AN UPDATE ON THE HANALYZER SYSTEM

Presenter: William A Baumgartner Jr, University of Colorado, Denver

Authors: William A Baumgartner Jr, Hannah Tipney, Lawrence Hunter

ABSTRACT: The advent of high-throughput genomic approaches has brought with it a new set of challenges faced by biologists, one the most common being what to do when given a large list of genes to analyze. The Hunter Lab has taken a data-integration approach to addressing this problem. We have previously reported on a system, the Hanalyzer, designed to help biologists leverage current gene-centric knowledge gathered from online resources and mined from the biomedical literature. The Hanalyzer outputs a network combining this gene-centric knowledge with experimental data provided by the biologist. We have demonstrated that by exploring this hybrid network biologists are able to decipher the underlying stories in their data, and perhaps more importantly, they can use the network to generate novel hypotheses that then drive further experimental studies. This talk will provide a brief introduction to the system and focus on recent advances made towards the goal of helping biologists explore their data both efficiently and effectively. Since its release, system development has continued on a number of fronts including knowledge network expansion, the construction of additional network navigational aids, and overall system enhancements ranging from handling data from different species to performance tuning. We will cover these improvements and conclude with a brief look at development issues that are on the horizon, specifically the challenges that remain to bring network construction to the biologist's desktop.

OP2: THERMODYNAMICS-INSPIRED NCRNA SEARCH

Presenter: Jennifer Smith, Boise State University

Authors: Jennifer Smith, Pamila Ward

ABSTRACT: The use of prior information in covariance model parameter estimation is crucial since many RNA families only have a very few known examples or all those that are known are in a sub-family of the actual family. Experimental thermodynamic measurements of RNA structures point to a number of regularities in molecular stability that are not captured in current covariance modeling practice. Among these are the dependence of stability on hairpin closing pair and loop end nucleotide identities and well as hairpin loop length. Preliminary evidence has been found that incorporation of these effects into priors and model structure may improve covariance model performance.

OP3: ALGORITHM TO IMPROVE GENE CONSISTENCY ACROSS BACTERIAL GENOMES

Presenter: Judith D. Cohn, Los Alamos National Laboratory

Authors: Judith D. Cohn, Michael E. Wall, John Dunbar

ABSTRACT: Identification of gene boundaries — the first step in genome annotation — provides the foundation for subsequent comparative genomics. Unfortunately, gene-finding algorithms are not always accurate. Even when gene-containing regions are correctly identified, gene prediction algorithms must select a single start site from a multitude of possible start sites. Errors in gene start sites not only alters the encoded peptide sequence but may affect the identification of orthologs. Further, the choice of start site affects the length of the intergenic region, which may impact a suite of other predictions, such as operon structure, regulatory motifs, and comparison of regulatory regions among genomes. In extreme cases, errors may lengthen intergenic regions to the extent that additional genes are predicted in the intervening space. Conversely, errors can remove intergenic content, sometimes resulting in spurious gene overlaps. Recently, we noted extensive inconsistencies in gene start sites among orthologous genes within the Burkholderia genus. In particular, we found many orthologous gene sets where predictions appeared consistent for all but one ortholog in a set, suggesting a possible gene-finding error. We posit that inconsistency in gene start sites among orthologs represents a gene-finding error in some cases and real biological variation in others. In this context, we present an algorithm to improve consistency across multiple genomes and characterize its performance for orthologs across the Burkholderia genus.

OP4: A GENOME-WIDE ANALYSIS OF POISED PROMOTERS IN BACTERIA

Presenter and Author: Marko Djordjevic, Arkansas State University and The Arkansas Bioscience Institute

ABSTRACT: As the first and usually rate-limiting step of transcription initiation, bacterial RNA polymerase binds to double stranded DNA (the closed complex formation) and subsequently opens the two strands of DNA (the open complex formation). Poised promoters in bacteria are sequences where RNAP binds with high binding affinity, but which do not have detectable levels of transcription initiation due to too slow transition from closed to open complex. Existence of a considerable number of poised promoters in genome has been often hypothesized, but poised promoters have not been systematically studied, since a large scale analysis of promoter kinetics is not experimentally feasible. To computationally address promoter poisoning on a genome-wide scale we use a recently developed biophysical model of transcription initiation [1]. We show that promoter poisoning is significantly reduced by i) Existence of -35 box interactions ii) Binding specificities of (physically independent) RNAP domains that interact with -10 box single-

stranded and double-stranded DNA. We show that the later (dominant) effect is not due to generic properties of protein-DNA interactions, and argue that RNAP is designed to reduce promoter poisoning in genome. However, despite this reduction, the number of poisoned promoters is still significant, and corresponds to ~30% of strongly bound sequences in bacteria [2]. This number roughly matches with lower bound of reported false positives in RNAP ChIP-chip experiments, which suggests that poisoned promoters are a major contributor to false positives in searches of bacterial promoters. [1] M Djordjevic and R Bundschuh, *Biophysical Journal* 94: 4233 (2008). [2] M Djordjevic, submitted (2009).

OP5: STRUCTURE-BASED PREDICTION OF DNA BINDING SITES FOR FAMILIES OF TRANSCRIPTION FACTORS

Presenter and Author: Julia Ponomarenko, University of California, San Diego

ABSTRACT: Gene transcription is regulated through binding of transcription factors (TF) to specific sites on DNA, known as TF binding sites (TFBS). Most algorithms for predicting TFBSs in genome sequences apply position-specific weight matrices (PWM) obtained from aligned TFBS sequences. That requires a significant amount of experimentally identified TFBSs. Alternative approaches include building PWMs either for families of TFs sharing similar DNA-binding domains or using 3D structures of TF-DNA complexes. Here a novel method for the prediction of TFBSs binding a family of TFs is presented. The method takes into account both 3D structures of TF-DNA complexes and sequences of TFBSs. Structures of DNA-binding protein domains are aligned, using the algorithm of structural alignment that favors matches between residues interacting with DNA. As a result the alignment of DNA bound to the proteins emerges. Thus aligned DNA sequences are used as a seed for aligning all TFBSs. A PWM is further defined, assuming that the probability of a base pair to be at a certain position of the site depends on its occurrence in the sequence alignment and the number of contacts with TF in the structural alignment. Applied to the sites binding the NF κ B family factors, the method has been shown to discriminate TFBS sites from non-sites significantly better than other tested sequence- and structure-based methods. Notwithstanding, the correlations between experimentally measured TF-DNA binding affinity values and binding scores predicted by the proposed method were comparable to those calculated using other methods. This work is funded by NIH grant R01GM085325.

OP6: GENOME-WIDE DISCOVERY OF HUMAN HEART ENHANCERS

Presenter: Ivan Ovcharenko, NIH

Authors: Leelavati Narlikar, Noboru Sakabe, Alexander Blanski, Fabio Arimura, Marcelo Nobrega, Ivan Ovcharenko

ABSTRACT: The various organogenic programs deployed during embryonic development rely on the precise expression of a multitude of genes in time and space. Identifying the cis-regulatory elements responsible for this tightly

orchestrated regulation of gene expression is an essential step in understanding the genetic pathways involved in development. We describe a strategy to systematically identify tissue-specific cis-regulatory elements that share combinations of sequence motifs. Using heart development as an experimental framework, we employed a combination of Gibbs sampling and linear regression to build a classifier that identifies heart enhancers based on the presence and/or absence of various sequence features, including known and novel TF binding specificities. In distinguishing heart enhancers from a large pool of random noncoding sequences, the performance of our classifier is vastly superior to four commonly used methods, with an accuracy reaching 92% in cross-validation. Furthermore, most of the binding specificities learned by our method resemble the specificities of TFs widely recognized as key players in heart development and differentiation, like SRF, MEF2, ETS1, SMAD, and GATA. Using our classifier as a predictor, a genome-wide scan identified over 40,000 novel human heart enhancers. Although the classifier used no gene expression information, these novel enhancers are strongly associated with genes expressed in the heart. Finally, in vivo tests of our predictions in mouse and zebra-fish achieved a validation rate of 60%, significantly higher than what is expected by chance. These results support the existence of underlying computationally amendable cis-regulatory codes dictating tissue-specific transcription in mammalian genomes.

OP7: DEVELOPMENT OF METHODS FOR INTEGRATING DIVERSE SOURCES OF GENOME-SCALE DATA

Presenter: Daniel Dvorkin, University of Colorado, Denver

Authors: Daniel Dvorkin, Katerina Kechris

ABSTRACT: Making use of multiple data sources across the genome is a major challenge in modern bioinformatics. Here we present some preliminary methods for combining signals from multi-species sequence conservation, transcription factor binding and gene expression data. Our approach is applied to data that cover most of the *D. melanogaster* genome and were curated to study embryo development. Methods developed allow identification of “hot spots” in the genome that are involved in the regulation of critical developmental processes and provide assessment of the significance of the results. The methods are general and designed to be applied to a wide variety of problems where there are data sources that span the genome but may be generated using diverse technologies, have varying genomic density or may follow different distributions.

OP8: VISUALIZING GENOMIC SEQUENCES IN 2D

Presenter and Author: Josiah Seaman, Colorado State University

ABSTRACT: It is increasingly evident that there are multiple and overlapping patterns within the genome, and that these patterns contain different types of information — regarding both genome function and genome history. In order to

discover additional genomic patterns which may have biological significance, novel strategies are required. To partially address this need, we introduce a new data visualization tool entitled Skittle. It first creates a 2-dimensional nucleotide display by assigning four colors to the four nucleotides, and then text-wraps to a user adjustable width. This nucleotide display is accompanied by a 'repeat map' which comprehensively displays all local repeating units, based upon analysis of all possible local alignments. Skittle includes a smooth-zooming interface which allows the user to analyze genomic patterns at any scale. Skittle is especially useful in identifying and analyzing tandem repeats, including repeats not normally detectable by other methods. However, Skittle is also more generally useful for analysis of any genomic data, allowing users to correlate published annotations and observable visual patterns, and allowing for sequence and construct quality control. Preliminary observations using Skittle reveal intriguing genomic patterns not otherwise obvious, including structured variations inside tandem repeats.

OP9: AUTOMATED INFERENCE OF MOLECULAR MECHANISMS OF DISEASE FROM AMINO ACID SUBSTITUTIONS

Presenter: Biao Li, Indiana University

Authors: Biao Li, Vidhya G. Krishnan, Matthew E. Mort, Fuxiao Xin, Kishore K. Kamati, David N. Cooper, Sean D. Mooney, Predrag Radivojac

ABSTRACT: Single nucleotide substitutions within protein coding regions are of particular importance owing to their potential to give rise to amino acid substitutions that affect protein structure and function which may ultimately lead to a disease state. Over the last decade, a number of computational methods have been developed to predict whether such amino acid substitutions result in an altered phenotype. Although these methods are useful in practice, and accurate for their intended purpose, they are not well suited to providing probabilistic estimates of the underlying disease mechanism. We have developed a new computational model, MutPred, that is based upon protein sequence, and which models changes of structural features and functional sites between wild-type and mutant sequences. These changes, expressed as probabilities of gain or loss of structure and function, can provide insight into the specific molecular mechanism responsible for the disease state. MutPred also builds on the established SIFT method but offers improved classification accuracy with respect to human disease mutations. Given conservative thresholds on the predicted disruption of molecular function, we propose that MutPred can generate accurate and reliable hypotheses on the molecular basis of disease for ~11% of known inherited disease-causing mutations. We also note that the proportion of changes of functionally relevant residues in the sets of cancer-associated somatic mutations is higher than for the inherited lesions in the Human Gene Mutation Database which are instead predicted to be characterized by disruptions of protein structure.

OP10: A USER STUDY OF ATTRIBUTE VISUALIZATION TOOLS AND THEIR ROLE IN UNDERSTANDING BIOLOGICAL NETWORKS

Presenter: Hande Kucuk, Eastern Michigan University

Authors: Hande Kucuk, Benjamin J. Keller, Terry Weymouth, Barbara Mirel

ABSTRACT: With the growth of databases of biological relationships, the visual analysis of molecular networks has become increasingly important. These networks can be large and complicated, making them difficult to comprehend. This research is directed at studying how different attribute visualization approaches affect users' ability to understand and analyze biological networks. We discuss a user study that we are performing of three Cytoscape tools that allow Gene Ontology annotations to be visualized and used to select proteins in networks retrieved from the MiMI (Michigan Molecular Interaction) database. These tools support varying levels of visualization, from value-based selection in the Cytoscape data panel, a list-based attribute browser, and an interactive chart-based attribute browser that shows the frequency of attributes in the network. Our goal is to observe whether the differences in these tools help the users better understand complex networks. The study, starting in late fall 2009, will ask users to find patterns or relationships among the molecules in a network constructed using other functional annotation, and capture user action and think-out-louds for analysis. We present the three strategies for attribute selection and visualization, the user study and preliminary results.

OP11: NEUROIE: EXTRACTING NEUROIMAGING STUDY RESULTS FROM THE LITERATURE

Presenter: Yong Gao, Mass General Hospital/Harvard Medical School

Authors: Yong Gao, Dave Kennedy

ABSTRACT: Experimental studies utilizing neuroimaging techniques, such as MRI and PET, have generated and reported increasing amount of results in the literature. These results typically show the correlation between certain aspects of the brain structure with behavioral observations or disease diagnosis for a group of demographically characterized subjects. Because these results mostly appear only in the full text body of the articles, they are not readily available for systematic retrieval and access via online databases of scientific literature such as PubMed. Our previous work on IBVD provides a searchable database of brain neuroanatomic volumetric observations that are manually curated from the journal articles. We are developing an information extraction system called NeuroIE, aiming to automate the literature curation process to keep pace with the growing number of publications. Initially, NeuroIE will focus on the extraction of the following information: demographic information of subjects, volumetric measurements, disease diagnosis, and the neuroimaging methods used in the experiment. NeuroIE demands several challenging text mining and NLP techniques. First, it is essential for NeuroIE to be able to process the full text of journal articles, as compared with most IE research focusing on abstracts. Second, NeuroIE must be capable of understanding

tabular information because much research results are reported as tables. Third, sentence and paragraph processing is critical when data is only reported descriptively as text.

OP12: ANALYSIS OF A LOCAL HUNTINGTIN PROTEIN INTERACTION NETWORK

Presenter: Corey Powell, Buck Institute for Age Research

Authors: Corey Powell, Robert Hughes, Cendrine Tourette, Russell Bell, Sean Mooney

ABSTRACT: Huntington's Disease is a neurodegenerative disorder caused by an abnormally long stretch of glutamines in the associated huntingtin protein. This study sheds light on possible functions for the huntingtin protein through analysis of a local protein-protein interaction network consisting of the huntingtin protein, proteins called primaries that have been found to interact with the huntingtin protein and secondary proteins that interact with the primary proteins. The first part of the analysis finds annotations that are overrepresented among the primary and secondary proteins. The second part of the analysis examines the network structure and finds functions and proteins that are more highly connected in the network than expected by chance. The third part of the analysis uses additional information, such as gene coexpression, to corroborate the results from the first two analyses.

OP13: IN SILICO FUNCTIONAL PROFILING OF HUMAN DISEASE-ASSOCIATED AND POLYMORPHIC AMINO ACID SUBSTITUTIONS

Presenter: Vidhya G. Krishnan, Buck Institute for Age Research

Authors: Matthew Mort, Uday S. Evani, Vidhya G. Krishnan, Kishore K. Kamati Peter H. Baenziger, Anghuman Bagchi, Brandon Peters, Rakesh Sathyesh, Biao Li, Yanan Sun, Bin Xue, Nigam Shah, Maricel Kann, David N. Cooper, Predrag Radivojac, Sean D. Mooney

ABSTRACT: We have co-opted a range of bioinformatic tools, designed to predict structural and functional sites in protein sequences, to the task of ascertaining whether intrinsic biases exist in terms of the distribution of different types of human amino acid substitutions (AAS) with respect to their structural, functional and pathological features. We applied these tools to compiled datasets of human disease-associated AAS in the contexts of inherited monogenic disease, complex disease, functional polymorphisms with no known disease association, somatic mutations in cancer, and neutral polymorphic AAS. The analysis revealed marked similarities in terms of the distribution of structural and functional sites between monogenic disease mutations and functional polymorphisms, with a bias toward those variants that impact protein function via structural disruption ($P=3.8 \times 10^{-24}$). Putative causative variants in both complex disease and cancer were significantly over-represented in intrinsically disordered regions ($P=8.83 \times 10^{-56}$) whilst cancer-associated mutations were enriched at certain molecular recognition sites ($P=1.6 \times 10^{-3}$). We postulate that missense mutations in complex disease and cancer are more likely than monogenic disease to impact on protein function

directly through disruption of functional sites (e.g. protein interaction) rather than indirectly via structural disruption. Further analysis of subtypes of inherited disease (e.g. cardiovascular disease) served to identify several disease entities that differed significantly in terms of the distribution of specific causative molecular changes. For example, blood coagulation disorders were found to exhibit a 19-fold depletion in AAS at O-linked glycosylation sites. In overall terms, however, the disruption of a specific molecular function does not constitute a disease-specific phenomenon.

OP14: ACCELERATING CANDIDATE GENE DISCOVERY THROUGH ONTOLOGICAL INDEXING OF LARGE SCALE DATA REPOSITORIES

Presenter: Simon Twigger, Medical College of Wisconsin

Authors: Simon Twigger, Joey Geiger, Jennifer Smith

ABSTRACT: Are any of these genes associated with my disease or phenotype? Is this candidate gene expressed in my tissue of interest? These are examples of common questions asked virtually every day by scientists attempting to identify genes contributing to human disease. Model Organism Databases such as the Rat Genome Database (RGD) curate published data related to these questions but there is much more information available than can be manually curated. Much of this information is being deposited into large scale data repositories but extracting useable information and knowledge from this stored data is a challenging problem. We are tackling this by annotating data in repositories such as NCBI's Gene Expression Omnibus (GEO) with biomedical ontologies using the National Center of Biomedical Ontology's web services. We are using an iterative process to automatically annotate the GEO records with ontology terms, followed by a manual curation/review using GMiner, a custom ruby on rails web application. Following review the data can then be explored on GMiner, allowing researchers to quickly find GEO datasets expressed in particular tissues and explore other attributes of those datasets. In addition we are creating an extensive annotation dataset linking genes to the tissues in which they have been expressed and making this available in RDF format to allow us to explore the benefits of integration with other semantic web resources. I will present the results of this annotation initiative and our plans to use these and other ontology annotations to accelerate candidate gene discovery.

OP15: KNOWLEDGE NETWORK APPROACH: PATHWAYS AND DRUGS

Presenter: Nikolai Daraselia, Ariadne

Authors: Nikolai Daraselia, Ekaterina Kotelnikova, Anton Yuryev

ABSTRACT: Providing a rich context for experimental data promises to offer new insights into mechanisms of molecular regulation. Employing a resource of millions of findings gleaned from a broad corpus of biomedical literature in which to evaluate genome-wide experimental data can highlight specific molecules otherwise easily missed. The challenge to use this large amount of data in decision-making can be met using appropriate hypothesis testing. Using the proprietary

high-content linguistics tool MedScan we compiled a database of knowledge networks associated with different diseases and small molecule effects by extracting the information from scientific literature. Different approaches towards reconstructing mechanistic models from the resulting knowledgebase and from microarray data will be described. By analyzing disease-specific gene expression data we were able to identify a potentially novel therapeutic target in glioblastoma. Further, by systematically mining the database for knowledge on existing drugs/drug candidates garnered from published findings, a new application for a known agent to inhibit glioblastoma pathway was suggested.

OP16: CAN WE ACCURATELY DETERMINE THE FITTEST GENES IN NATURE?

Presenter: Ramy K. Aziz, San Diego State University

Authors: Ramy K. Aziz, Mya Breitbart, Robert Edwards

ABSTRACT: Genes, like organisms, struggle for existence, and the fittest genes are those that persist and widely disseminate in nature. The unbiased determination of the fittest genes requires access to sequence data from a wide range of phylogenetic taxa and biogeographical ecosystems, a task that, until very recently, was not possible because genomic sequence data were biased towards organisms of human interest and did not fairly represent the tree of life. However, this goal has finally become achievable thanks to the emergence of the field of metagenomics, which allows unbiased sequencing of all living forms in a given ecosystem. Here, we introduce simple methods for the determination and normalization of gene abundance and ubiquity in publicly available genomic and metagenomic data sets. We analyzed 10 million protein-encoding genes in 2,137 sequenced bacterial, archaeal, eukaryotic, and viral genomes, and 187 metagenomes. Because metagenomic sequences widely vary in their read lengths and sequencing depth, we resorted to gene length normalization, which allowed more accurate determination of gene abundance in metagenomes. We present a list of the most abundant and the most ubiquitous genes with known biological function, and we discuss the essentiality of certain genes as opposed to the ecosystem-dependent pervasiveness of others. Surprisingly, one gene family was singled out as the most prevalent in both genomes and metagenomes.

OP17: ASSESSING MODELS OF PROTEIN INTERACTION NETWORK EVOLUTION

Presenter: Todd A. Gibson, University of Colorado, Denver

Authors: Todd A. Gibson, Debra S. Goldberg

ABSTRACT: Theoretical models of eukaryotic protein interaction network evolution featuring gene duplication and protein interaction dynamics have contributed to our understanding of proteome evolution. These models are vetted by comparing topological similarities of networks generated by the models with empirically-derived biological networks. However, these models introduce biases

which limits their ability to inform our understanding of proteome evolution. These biases include their construction from researcher-selected seed graphs, and their failure to consider the contribution of horizontal gene transfer, which played a prominent role in pre-eukaryotic evolution. We avoid these biases with a novel method of model assessment—running the model in reverse on the empirical network itself. We apply our method on a well-regarded model of protein interaction network evolution and assess it against the *Saccharomyces cerevisiae* protein interaction network. Our analysis of the model's shortcomings reveals that highly-connected self-interacting proteins are preferentially retained in yeast's evolution.

OP18: GENE AND GENOME TREES CONFLICT AT MANY LEVELS

Presenter: Leanne S. Haggerty, NUI Maynooth

Authors: Leanne S. Haggerty, Fergal J. Martin, David A. Fitzpatrick, James O. McInerney

ABSTRACT: Horizontal gene transfer (HGT) plays a significant role in microbial evolution. It can accelerate the adaptation of an organism, it can generate new metabolic pathways and it can completely remodel an organism's genome. We examine 27 closely related genomes from the YESS group of gamma proteobacteria and a variety of four-taxon datasets from a diverse range of prokaryotes in order to explore the kinds of effects HGT has had on these organisms.

OP19: PROTEIN-PROTEIN INTERACTIONS ARE DRIVEN BY FUNCTIONAL EVOLUTION

Presenter: Yiqiang Zhao, Buck Institute for Age Research

Authors: Yiqiang Zhao, Sean Mooney

ABSTRACT: Many fundamental biological processes involve protein-protein interactions. It is an interesting question that how and why proteins are becoming interacted and contributes to the genetic complexity of organisms. The preferential attachment model, when applied to protein-protein interaction, asserts that a protein is more likely to evolve to have many interaction partners if it has many interaction partners before the evolution. Consistent with the model we find that older genes tended to have more interactions than newer ones. Two findings, on the other hand, were not consistent with the preferential attachment model. One analysis divided human genes into 7 temporal groups. Most interactions, both new and old, involved gene partners generated not in the oldest temporal groups, but rather in the temporal group representing the evolution from vertebrates to warm-blooded animals. A second analysis found that there was not an increased number of interactions in 645 gene clusters with historic duplication events, which is inconsistent with the hypothesis if the preferential attachment governs the evolution of protein-protein interaction. Based on these two analyses, it does not appear that the preferential attachment model adequately explains the evolution of protein-protein interactions and we suggest that the interactions are selected to increase species-special complexity and achieve species-special functions in the evolution history instead of simply being driven by "rich get richer".

OP20: EVOLUTIONARY STUDY AND PREDICTION OF PROTEIN-PROTEIN INTERACTIONS IN CHROMATIN MODIFICATION COMPLEXES

Presenter: Xuejian Xiong, Hospital for Sick Children

Authors: Xuejian Xiong, Tuan On, Shuye Pu, Andrei Turinsky, Yunchen Gong, Andrew Emili, Zhaolei Zhang, Jack Greenblatt, Shoshana J. Wodak, John Parkinson

ABSTRACT: Chromatin modification (CM) comprises an array of broadly conserved biological processes that modify chromatin to control access to DNA. Due to its fundamental role in processes such as transcription, DNA replication and repair, many components of CM are thought to play an important role in various forms of cancer and cancer related developmental processes. While CM machinery has been extensively characterized in yeast, less is known about its operation in other eukaryotes. Here 111 eukaryotic genome sequences are used to systematically profile the conservation and evolution of CM by applying the InParanoid algorithm based on the comprehensive surveys of literature and database resources. Furthermore, we examined the distribution of the evolutionary distinct groups within the yeast protein-protein interaction network. Many established CM complexes were readily identifiable. Particularly striking is the mosaic nature of the conservation patterns of the components of each complex. In addition, we exploit the conserved interactions among five model species, i.e. yeast, worm, fly, mouse and human for CM complexes. In RSC/SWI/SNF complexes, the interaction of SWI3-SNF5 is conserved in all five models, while that of SWI3-SNF2 is conserved in four models except worm. Based on the evolutionary study of the complexes, we can predict that the triangle interactions among SWI3, SNF2, and SNF5 are conserved across model species. Together these findings provide insights into the evolution of CM and will help facilitate an improved annotation of CM across eukaryotes, and the prediction of the protein-protein interactions.

OP21: AN INDIVIDUAL BASED MODELLING APPROACH TO STUDYING THE EVOLUTION OF MATE CHOICE STRATEGY

Presenter and Author: Robert Williamson, Rose-Hulman Institute of Technology

ABSTRACT: Sexual selection is a key force driving morphological and behavioral evolution. It can lead to alternate life strategies within a species (for example some males fighting for mates and some sneaking copulations), or the development of exaggerated traits (like large size or elaborate ornaments). Key factors determining how sexual selection will affect a species' evolutionary trajectory are the mate choice strategies of female choice and male-male competition. Female choice encompasses the good gene, resource supply, and sensory exploitation strategies; while male-male competition consists of the direct confrontation, sperm competition, and infanticide strategies. Because each of these strategies affects the life histories of species that exhibit them, it is important to understand what factors may influence the development of each strategy. I developed an individual based model (IBM) system using the Swarm java library to investigate how various

biotic and abiotic factors influence the evolution of different mate choice strategies, and which factors can cause shifts between the strategies. The factors included features like resource availability, gestation length, sex ratio, and infanticide rate. I will discuss the design and implementation of the system, including architecture and built in assumptions. I will also address the success of the IBM method to this type of scientific inquiry.

OP22: MTHAPLOGROUPS: A CURATED WEB RESOURCE FOR MITOCHONDRIAL VARIATION

Presenter: Michael V. Osier, Rochester Institute of Technology

Authors: Kyle Dewey, Eric Stevens, Dina L. Newman, Michael V. Osier

ABSTRACT: Within the mitochondrial genome, there exist a substantial number of variations. Based upon these variations, maternal ancestry of populations and individuals can be identified to varying degrees of refinement. The relatively high rate of mutation in the mitochondrial genome means that many variations are recent, often within the last 10,000 years. Haplogroups are therefore significant to the studies of population genetics and human migrations on an evolutionarily short timeframe. Additionally, some genetic disorders, including Parkinson's disease and Alzheimer's disease, are statistically correlated to specific haplogroups. From this knowledge, it should be possible to determine which specific mitochondrial variations are associated with such disorders. However, there are many variations associated with each major haplogroup. Additionally, only a minimal number of the variations that differentiate haplogroups are well established. There exist many other, less well known, variations which are specific to individual subhaplogroups, subsets of haplogroups. These less well known variations are more likely to be truly causative of diseases and disorders. To help lasso the relevant data for medical studies of mitochondrial variation, we constructed mtHaplogroups: a Web-accessible interface which allows users to mine the variation data relevant to their haplogroups of interest. As of this submission, mtHaplogroups contains 4029 entries, including mutations to 2185 unique basepairs of the mitochondrial genome. Associated changes to coding regions for each mutation are noted. mtHaplogroups can be found at "<http://momtong.rit.edu/cgi-bin/haplogroups/haplogroups.cgi>".

OP23: POSTERIOR-PREDICTIVE DETECTION OF MOLECULAR CO-EVOLUTION USING PHYLOGENETICALLY-INTEGRATED MUTUAL INFORMATION

Presenter: A.P. Jason de Koning, University of Colorado, Denver

Authors: A.P. Jason de Koning, Todd A. Castoe, Hyun-min Kim, David D. Pollock

ABSTRACT: The statistical detection of co-evolving amino acids has long been an important tool for understanding the determinants of protein structure, and for making predictions about the effects of mutations. Unfortunately, one of the most popular approaches for detecting molecular co-evolution is to calculate the mutual

information (MI) between columns of a multiple sequence alignment, using a frequency-based approximation that in effect ignores the process and consequences of evolution — including the phylogenetic non-independence of sequences, differences in evolutionary rates across sites, and non-uniform propensities for amino-acid replacement. Despite this, ‘tree ignorant’ methods appear to often be preferred over more well-justified phylogenetic approaches due to their simplicity and speed. We show that: 1) such non-evolutionary MI statistics can be easily misled by a variety of natural phenomena, making their interpretation difficult at best; and 2) it is much better to calculate MI by integrating the times spent in each state (and pairs of states) over the branches of a phylogeny, under a reasonable model of sequence evolution. We describe how this can be done very efficiently by exploiting our recently developed data augmentation scheme for the rapid MCMC analysis of large phylogenomic datasets (partial sampling of substitution histories’; de Koning, Gu, and Pollock, 2009). The resulting method can take only minutes on a desktop computer to reliably detect co-evolving residues in large datasets, has built-in significance testing, and suffers from none of the shortcomings of standard MI approaches. Tree ignorant MI approaches are subject to substantial biases, and should be abandoned.

OP24: BIOINFORMATICS CHARACTERIZATION OF THE PLASMODIUM GLUTATHIONE S-TRANSFERASE

Presenter: Emilee Colón, University of Puerto Rico

Authors: Emilee Colón, Adelfa Serrano, Hugh Nicholas Jr, Troy Wymore, Alexander Ropelewski, Ricardo González Méndez

ABSTRACT: Malaria is a global health problem caused by *Plasmodium* parasites. Glutathione S-transferase (GST) is involved in the conjugation of glutathione to drugs and toxic compounds. It is postulated that GST plays an important role in the development of drug resistance. The three-dimensional (3D) structure of *Plasmodium falciparum* GST (PfGST) has been solved. Previous work indicates that the PfGST cannot be assigned to any of the known GST classes. We performed sequence analyses and structural modeling of GSTs from *Plasmodium*, and structural alignments to known structures of the GST from other organisms, in order to classify PfGST into a GST family. Sequence alignments using ClustalW, motif analysis using MEME, and phylogenetic analysis using MEGA4, of PfGST, *Plasmodium vivax* GST (PvGST), *Plasmodium knowlesi* GST (PkGST) and *Plasmodium yoelii* GST (PyGST) and 38 other GST sequences were done. The alignments, motifs and phylogenies show a close relationship to the alpha and sigma class of GSTs. Models of the tertiary structure of the *P. vivax*, *P. knowlesi* and *P. yoelii* GST monomers were obtained using the Protein Homology/analogY Recognition Engine (PHYRE) server. The three-dimensional structures of GST enzymes from various classes (alpha, sigma, mu and pi) were analyzed by structural alignment with the PfGST 3D structure (1Q4J) using the MultiSeq feature in the

VMD program. The comprehensive comparison of PfGST with known GST structures reveals high structural similarity that allows PfGST to be classified into unique clade within the sigma class GSTs. These data may open new avenues for the development of novel antimalarials.

OP25: LEVERAGING EXISTING BIOLOGICAL KNOWLEDGE IN GENOME WIDE ASSOCIATION STUDIES

Presenter: Ronald Schuyler, University of Colorado, Denver

Authors: Ronald Schuyler, Lawrence Hunter, Deborah Glueck

ABSTRACT: Gnome-wide association studies (GWAS) have had success in detecting disease-associated genetic variants for several diseases by comparing the frequencies of each individually tested polymorphism between two populations. However, for many other diseases studied, only a fraction of the expected heritability has been explained. As it is expected that susceptibility to many common diseases may be influenced by the contribution of multiple distinct variations, we are developing an approach to consider multiple loci simultaneously. Our approach uses protein-protein interactions, biochemical pathways, gene regulation, shared Gene Ontology annotation terms and inferred relationships automatically extracted from the literature to combine association results from multiple loci that are unlinked genetically, but associated in our diverse knowledge-base. We test specifically for interacting loci, but avoid unnecessarily increasing the multiple testing burden that would be incurred by testing all possible loci pairs by limiting our search to biologically relevant gene pairs. By leveraging existing knowledge, we increase the power of GWAS to overcome the confounding effects of genetic heterogeneity and epistatic interactions between genes.

OP26: EVALUATING MICROBIAL DIVERSITY AND ADAPTATION: NEW OPPORTUNITIES FOR INSIGHT IN A DATA RICH WORLD

Presenter: Catherine Lozupone, University of Colorado, Boulder

Authors: Catherine Lozupone, Rob Knight

ABSTRACT: Advances in sequencing technology have produced exciting opportunities to understand microbial diversity and adaptation, but also challenges to interpret datasets generated on a whole new scale. UniFrac is a tool that uses phylogenetic information to compare sequences from many microbial assemblages simultaneously, and can determine the main factors, such as temperature, pH, or disease state, that explain variation between microbial assemblages. UniFrac has been used in a wide diversity of studies of both biomedical and ecological importance. By applying it to the analysis almost 100,000 sequences compiled from 181 studies of diverse microbial assemblages deposited in GenBank, I have illustrated that the bacteria that inhabit the vertebrate gut are particularly distinct from free-living communities, and that the distribution of bacterial diversity in free-living assemblages is largely governed by salinity and substrate type (i.e.

whether the samples were from soil/sediment or water). I will illustrate computational enhancements that will allow for the application of UniFrac to studies containing hundreds of samples and millions of sequences, which next-generation sequencing technology has made possible. Next, I will talk about the extension of UniFrac to whole genome comparisons, allowing us to test whether horizontal gene transfer, parallel gene loss and duplications cause the repertoires of functional genes to converge in organisms that inhabit the same environment. As an example, I will illustrate how the repertoires of carbohydrate active enzymes have converged in human gut microbes compared to their non-gut relatives.

OP27: WOW! THAT'S REALLY INTERESTING!

Presenter: Hannah Tipney, University of Colorado, Denver

Authors: Hannah Tipney, Lawrence Hunter

ABSTRACT: The advent of high-throughput genomic approaches to analyze biological systems has been a watershed period for biomedical research, facilitating breakthroughs and generating hypotheses that have advanced our understanding of the human condition. However, the large datasets generated by these methods have proved to be incredibly challenging for scientists to interpret, not only because of the sheer number of genes or proteins under investigation, but also due to the exceedingly rich and complicated knowledge which can be tied to each of those biological entities. When analyzing these large datasets, huge amounts of time can be spent exploring relationships, knowledge and experimental data, much of which has little bearing on what the scientist actually finds interesting or useful. It would therefore be incredibly useful to be able to determine those aspects of the data- and knowledge-space that are of interest to the scientist. This would enable the scientist to quickly and easily identify those aspects demanding their concentration, while paying less attention to those of lesser interest. However, what one scientist deems to be interesting may be in stark contrast to another, even when considering the same dataset, and so the identification of ‘interestingness’ needs to be determined and driven by the user, and be flexible enough to adjust to different users specifications. Here we will discuss the different ways a user can consciously and unconsciously assert their preferences, interests, likes and dislikes, and how this can help focus the analysis of multidimensional biomedical data.

OP28: GLOBAL OPTIMIZATION IN NONLINEAR MODELS: OPTIMIZATION AND EVOLUTION IN METABOLIC NETWORKS

Presenter: Albert Sorribas, CMB – IRBLLLEIDA – University of Lleida

Authors: Albert Sorribas, Carlos Pozo, Gonzalo Guillén-Gosálbez, Laureano Jimenez, Rui Alves

ABSTRACT: Global optimization methods for nonlinear models are required both in biotechnological applications and in studying evolution and adaptive responses in metabolic networks. In this presentation, we discuss a method that can be

applied to power-law models. These are canonical models that greatly facilitates modeling and analysis of metabolic networks. These methods are an alternative to stoichiometric models to go one step forward in analyzing systems responses to changes in gene expression. Optimization strategies and its application to finding feasibility regions related to specific physiological constraints are discussed in the case of the adaptive response of yeast to heat shock. We also briefly discuss the utility of recasting to extend this method to any nonlinear model.

OP29: VIRUS DISCOVERY BY DEEP SEQUENCING AND ASSEMBLY OF VIRUS-DERIVED SMALL SILENCING RNAS

Presenter: Qingfa Wu, UC Riverside

Authors: Qingfa Wu, Yingjun Luo, Rui Lu, Nelson Lau, Eric C Lai, Wan-Xiang Li, Shou-Wei Ding

ABSTRACT: In response to infection, invertebrates process replicating viral RNA genomes into small interfering RNAs (siRNAs) of discrete sizes to guide virus clearance by RNA interference. Here we report that viral siRNAs sequenced from fruitfly, mosquito and nematode cells were all overlapping in sequence, suggesting a possibility to use them for viral genome assembly and virus discovery. To test this idea, we examined contigs assembled from published small RNA libraries and discovered five new viruses from cultured *Drosophila* cells and adult mosquitoes, including three with a positive-strand RNA genome and two with a dsRNA genome. Notably, four of the identified viruses exhibited only low sequence similarities to known viruses so that none could be assigned into an existing virus genus. We detected virus-derived PIWI-interacting RNAs (piRNAs) in *Drosophila* ovary somatic sheet cells and demonstrated that the longer viral piRNAs could also assemble into viral genomes in absence of siRNAs. Thus, this study provides a powerful culture-independent approach for virus discovery in invertebrates by assembling viral genomes directly from host immune response products without prior virus enrichment or amplification. We propose that invertebrate viruses discovered by this approach may include new human and vertebrate viral pathogens that are transmitted by arthropod vectors.

OP30: PREDICTION OF PROTEIN PROTEIN INTERACTION SITES AND THEIR IMPACT ON GENETIC DISEASE

Presenter: Angshuman Bagchi, Buck Institute for Age Research

Authors: Angshuman Bagchi, Eunseog Youn, Matthew E. Mort, David N. Cooper, Sean D. Mooney

ABSTRACT: Protein-protein interactions play pivotal roles in many biological processes, for example, hormone-receptor binding and so on. There is increasing evidence that disease-causing mutations lead to disruption of protein interactions. In the present work, we used machine learning tools to discriminate between interface and non-interface residues of proteins using structure and sequence

information. We generated features from protein sequence and structure using a non-redundant set of protein hetero-complexes from the protein data bank. The training dataset comprised of (A) interface residues and non-interface surface residues for structure based prediction and (B) interface residues and non-interface surface and core residues for sequence based prediction. The datasets were used to build classification algorithms using random forest (RF) and support vector machine (SVM) coupled with 10 fold cross-validation for evaluation. Overall, RF outperformed SVM in most cases. The best performing sequence-based classification tool achieved an accuracy of 73% and when protein structure was included the accuracy was 75%. The predictors were then used to analyze mutation data including somatic mutations in cancer from tumor resequencing projects, the Human Gene Mutation Database (HGMD), and common human polymorphisms. The results showed an enrichment of protein interaction sites in the disease datasets compared to the neutral set (Seattle SNPs). Overall our results indicate that disease mutations are enriched in disruption of PPI interfaces and these interfaces can be predicted using bioinformatic approaches.

OP31: OPTIMAL NEAREST SHRUNKEN CENTROIDS METHOD FOR HIGH-DIMENSIONAL DATA CLASSIFICATION

Presenter: Tiejun Tong, University of Colorado, Boulder

Authors: Tiejun Tong, Herbert Pang

ABSTRACT: Class prediction with the nearest shrunken centroids (NSC) method has been shown to be very successful in many high-dimensional classification problems. Nevertheless, it has two major limitations: (i) the performance of the NSC method is not satisfactory when the sample size is relatively small owing to the large variation in feature selection; and (ii) the NSC method is ad hoc as the tuning parameter is chosen by cross-validation. In this paper, we propose a new algorithm that chooses the tuning parameter by minimizing some certain risk function. Simulation studies indicate that the proposed algorithm performs remarkably well compared to the originally NSC method by cross-validation when the sample size is small. Theoretical aspects of the tuning parameter estimation are also investigated. Finally, we conduct a real data study to evaluate the proposed findings.

OP32: RETROBIOSYNTHESIS: SEARCHING BACKWARDS, LOOKING FORWARDS

Presenter and Author: Dan McShan, University of Colorado, Denver

ABSTRACT: By now, the need for biosynthetic pathways as 'green' alternatives to the synthesis of industrial organic chemicals is well recognized. Inspired by the Nobel Prize winning work of E.J. Corey in chemical retrosynthesis, we will explore a retro-A* approach to finding biosynthetic pathways. For several years now, the presenter has advocated that the metabolic search algorithm developed in the

PathMiner A* metabolic search algorithm is provably complete, optimal as well as optimally efficient. In this research, we attempted to falsify these hypotheses by testing against a retro-A* algorithm which searches from the desired product backwards to the starting compound. These were experimentally tested by exhaustively searching from a common carbon source (alpha d-glucose, ADG) to every compound in the KEGG database twice over — once from ADG to everything, and then from everything to ADG. The results were quite surprising. While hypotheses of PathMiner A* algorithm's completeness do hold, the claim of optimal efficiency is seemingly falsified as retro-A* is significantly more efficient. More interesting still, the optimality of PathMiner A* is called into question as it appears that retro-A* is able to find shorter pathways in 24 instances. However, the mystery is unravelled as it turns out that retro-A* approach is algorithmically flawed in how it handles directionality of enzymatic reactions. In conclusion, we believe that PathMiner A* remains complete, optimal and optimally efficient for forward search algorithms, but that retroA* will ultimately prove to be substantially faster at finding the same pathways.

OP33: EXPERIMENTS WITH BIOLOGICAL CONCEPT RECOGNITION TOOLS

Presenter: Karin Verspoor, University of Colorado, Denver

Authors: Karin Verspoor, Kevin B. Cohen, Helen Johnson, Christophe Roeder, Cesar Mejia, William Baumgartner Jr, Larry Hunter

ABSTRACT: Information extraction from the biomedical literature requires recognition of core biological concepts, ranging from named entities such as proteins and diseases to more abstract concepts such as the biological processes and molecular functions captured in the Gene Ontology. To date the bulk of effort for concept recognition has focused on named entities, although several dictionary-based lookup tools are available for concept recognition of arbitrary ontology concepts. In this work, we report on the results of experiments aimed at gauging the effectiveness of these tools for recognition of concepts from the Gene Ontology and the Cell Type Ontology, exploring various parameter combinations of the tools. We take advantage of a manually annotated corpus of molecular biology journal publications, the CRAFT (Colorado Rich Annotation of Full Text) corpus, as well as a specifically designed test suite, as the gold standards for evaluation. We identify specific classes of variation in the natural language expression of these ontology concepts that pose difficulties for the dictionary-based tools and propose an alternative strategy for concept recognition based on the OpenDMAP system that is more robust to these variations in expression.

OP34: TEST SUITE DESIGN FOR BIOMEDICAL ONTOLOGY CONCEPT RECOGNITION SYSTEMS

Presenter: K. Bretonnel Cohen, University of Colorado, Denver

Authors: K. Bretonnel Cohen, Karin Verspoor, Christophe Roeder, William A. Baumgartner Jr, Lawrence Hunter

ABSTRACT: This talk describes an approach to evaluation of the workings of ontology concept recognition systems through use of a structured test suite and presents a publicly available test suite for this purpose. A structured test suite is built by enumerating the factors that might affect system performance and then systematically isolating, varying, and combining them in a data set in which inputs are paired with gold standard outputs. The goal is to build a test suite to which arbitrary ontology concept recognition systems can be applied. More broadly, we also seek to investigate what general principles, if any, might inform the construction of such test suites.

OP35: PREDICTING PROTEIN LINKAGES IN BACTERIA: WHICH METHOD IS BEST DEPENDS ON TASK

Presenter: Anis Karimpour-Fard, University of Colorado, Denver

Authors: Anis Karimpour-Fard, Sonia M. Leach, Ryan T. Gill, Lawrence E Hunter

ABSTRACT: Applications of computational methods for predicting protein functional linkages are increasing. In recent years, several bacteria-specific methods for predicting linkages have been developed. The four major genomic context methods are: Gene cluster, Gene neighbor, Rosetta Stone, and Phylogenetic profiles. These methods have been shown to be powerful tools and we provide guidelines for when each method is appropriate by exploring different features of each method and potential improvements offered by their combination. Using *Escherichia coli* K12 and *Bacillus subtilis*, linkage predictions made by each of these methods were evaluated against three benchmarks: functional categories defined by COG and KEGG, known pathways listed in EcoCyc, and known operons listed in RegulonDB. Each evaluated method had strengths and weaknesses, with no one method dominating all aspects of predictive ability studied. A common problem for computational methods is the generation of a large number of false positives that might be caused by an incomplete source of validation. By comparing two versions of a database, we demonstrated the dramatic differences on reported results. We used several benchmarks on which we have shown the comparative effectiveness of each prediction method, as well as provided guidelines as to which method is most appropriate for a given prediction task.

OP36: A METAGENOMIC STUDY OF A MICROBIAL MAT IN A HAWAIIAN LAVA CAVE

Presenter: Gayle K. Philip, NASA Astrobiology Institute

Authors: Gayle K. Philip, Mark V. Brown, Stuart P. Donachie

ABSTRACT: The existence of a subsurface biosphere on Earth has focused attention on the possibility that life on Mars may have retreated to subsurface 'oases' when surface conditions became unfavourable. High-resolution images show many Martian volcanoes were built from countless individual flows that were emplaced through channels and lava tubes, signalling a style of volcanism analogous to Hawaiian eruptions. Such lava tubes may be one of the few places on that planet where saturation levels of moisture might be maintained by the diffusion of geothermally heated groundwater. Furthermore, cave systems would shield organisms from high UV levels, whilst still providing sufficient light in the 'twilight' zone near cave entrances to facilitate low-light adapted photosynthesis. I will report how such a unique combination of environmental factors supports a complex microbial biofilm growing on the basalt roof of a lava cave, dating from 1922, in the volcanically active Kilauea Crater in Hawai'i Volcanoes National Park (HAVO). Anecdotal evidence suggests the biofilm is unique to this cave out of more than 200 other caves within the caldera. Using a combination of 16S ribosomal tag pyrosequencing and metagenomics, we have identified a highly novel community composition and complex metabolic interactions occurring within the biofilm.

OP37: MOBILE METAGENOMICS: ANNOTATING DNA SEQUENCES ON A CELL PHONE

Presenter: Josh Hoffman, San Diego State University

Authors: Josh Hoffman, Daniel Cuevas, Robert Edwards

ABSTRACT: Mobile Metagenomics is a metagenome annotation application for the Android mobile platform. The application allows for annotation to be performed in real-time. Results are finished on the order of a few minutes rather than hours or days. As partial results are completed, they are displayed to the user in real time. Mobile Metagenomics use Web services to send data to the seed (<http://www.theseed.org>) for processing. A typical Android interface provides a look and feel in line with user expectations. Fasta-formatted DNA sequence files are uploaded from the SD card on the phone using the OpenIntents open source file browser and a raw FORM-DATA bit stream. The server returns JSON-formatted data for the real-time annotation downloads. Each portion of the results is downloaded, processed, and displayed to the user in real time. Users are able to browse the partial results of their annotation as they wait for the full download to complete. All operations are fully interruptible; users can take calls, browse the web, and send/receive SMS messages during annotation. Mobile Metagenomics will continue to perform downloads in the background while the user utilizes their phone in any number of other ways. When results are complete, they can be saved

to the phone and loaded instantly later. The application implements the Android standard “Share” feature, which allows for users to email their data to other phone or computer users.

OP38: THE COLORADO RICHLY ANNOTATED FULL-TEXT (CRAFT) CORPUS: A RESOURCE FOR BIONLP RESEARCH

Presenter: Michael Bada, University of Colorado, Denver

Authors: Michael Bada, Miriam Eckert, Kristin Garcia, Donald Evans, Dmitry Sitnikov, William A. Baumgartner Jr, Philip V. Ogren, Arrick Lanfranchi, Amanda Howard, William Corvey, Nianwen Xue, Kevin B. Cohen, Karin Verspoor, Judith A. Blake, Martha Palmer,

ABSTRACT: Biomedical natural-language-processing (bioNLP) research increasingly relies on well-annotated corpora as gold standards for training and evaluation; we are therefore creating the Colorado Richly Annotated Full-Text (CRAFT) Corpus as such a resource. An initial corpus of 97 biomedical journal articles, primarily focused on the laboratory mouse, are being annotated in full. The annotation of this corpus is broadly divided into semantic and syntactic tasks, each of which includes a variety of annotation subtasks. As for the former, we are using high-quality terminologies, primarily ontologies from the Open Biomedical Ontologies (OBO) Consortium, as the terms used for semantic annotation. Furthermore, we are using these terminologies in their entirety, as opposed to using subsets of terms; ours is the first such effort of which we know to undertake this. Thus far, we have preliminarily finished annotation with the OBO Cell Type Ontology, the OBO GO cellular-component subontology, the OBO Chemical Entities of Biological Interest ontology, and the NCBI Taxonomy. Additionally, we are currently annotating the corpus with the GO biological-process and molecular-function subontologies, the OBO Sequence Ontology, and the Entrez Gene database, and we further intend to relationally link these semantic annotations. Syntactically, annotation tasks that we are performing include tokenization, part-of-speech tagging, and treebanking, and we have also begun coreferential annotation of these articles. All of our annotations and the articles will be freely available, and we believe this will be a significant resource providing ample avenues for bioNLP research.

OP39: SUBGRAPHS OF PROTEIN-PROTEIN INTERACTION NETWORKS WITH HIGH CONNECTIVITY

Presenter: Suzanne Gallagher, University of Colorado, Boulder

Authors: Suzanne Gallagher, Debra Goldberg

ABSTRACT: The edge connectivity of a graph is the minimum number of edges that must be removed in order to disconnect the graph. Likewise, the vertex connectivity (sometimes just called connectivity) is the minimum number of vertices and all adjacent edges that must be removed in order to disconnect the graph. Despite the fact that both edge and vertex connectivity are important properties of a graph, they have been used only rarely in the study of protein-protein interaction networks. We develop an algorithm to search for the subgraph

with the highest vertex and edge connectivity and apply this algorithm to various protein-protein interaction networks. In networks representing protein complexes, we found that most protein complexes have subgraphs that are 3-vertex-connected. Applying this algorithm to the entire protein-protein interaction network gave us an 8-connected clique of 9 proteins, 8 of which have been theorized to be co-complexed, and a 16-connected set of 49 membrane proteins whose significance we are still attempting to determine.

OP40: TREEHUGGER: A NEW TEST FOR ENRICHMENT OF GENE ONTOLOGY TERMS

Presenter: Daniel Jupiter, Texas A&M Health Science Center

Authors: Daniel Jupiter, Jessica Sahutoglu, Vincent VanBuren

ABSTRACT: The Gene Ontology (GO) project provides a structured vocabulary of biological terms, used by biological researchers as a tool for standardization of references to biological entities. Genes may be annotated with GO terms to indicate their roles or localizations in the cell. GO has been used in conjunction with high-throughput experimental methods, such as microarrays. In this setting, the interest is to determine whether sets of genes identified by the high-throughput experiment are enriched for GO terms: Do certain terms annotate more genes in the identified set than one might expect. Current methods for determining whether sets of genes are GO-enriched have certain well-known shortcomings. Many methods do not take the hierarchical structure of the ontology into account in determining enrichment. We address this drawback by introducing a new statistical test (TreeHugger) based on a novel per-gene scoring scheme for GO terms. Given a set of genes and a specified subset of those genes, our method determines enrichment of GO terms in the subset, taking into account the structure of the ontology and ascribing a lower weight to those terms that do not themselves directly annotate the given genes. Tests on simulated and real data indicate that our method is a conservative test for enrichment. Testing TreeHugger on a biological example reveals that it also reduces the redundancy caused by giving high scores to indirect annotations as provided by standard enrichment tests.

OP41: PATTERN-BASED EXTRACTION OF ARGUMENTATION FROM THE SCIENTIFIC LITERATURE

Presenter and Author: Elizabeth White, University of Colorado, Boulder

ABSTRACT: As the number of publications in the biomedical field continues its exponential increase, techniques for automatically summarizing information from this body of literature have become more diverse. In addition, the targets of summarization have become more subtle: initial work focused on extracting the factual assertions from full-text papers, but more recently, interest has shifted to recovering speculations and agreements or disagreements with other research.

Scientific writing is rife with such argumentation, and the premises, evidence, conjectures, objections and rebuttals that writers use to persuade the reader represent a rich vein of expert knowledge for summarization. Agreement, disagreement, and conjecture are often expressed in highly scripted ways; likewise, the higher-order discourse structures that underpin multisentence arguments tend to assume particular forms into which claims and evidence can be nested. These features make these kinds of arguments readily recoverable by pattern-based search. Here, I present PARROT, which uses OpenDMap patterns in combination with a Protégé ontology. PARROT first matches simple argumentative claims using a set of concepts relevant to scientific discourse and then exploits discourse cues and inference to combine these claims recursively into higher-order argument trees. PARROT outperforms an SVM classifier system in identifying statements of support and conflict at the sentence level. Additionally, PARROT provides a graphical representation of the arguments it finds, which makes it an valuable tool for summarizing the reasoning behind scientists' conclusions and identifying areas of consensus and contention.

OP42: EARLY HOST RESPONSE DURING INFLUENZA INFECTIONS

Presenter and Author: Christian V. Forst, University of Texas, Dallas

ABSTRACT: The recent emergence of the influenza virus from different reservoirs has raised concern about future strains of high virulence emerging that could easily infect humans. At present, it is not well understood what factors are responsible for variations in the virulence among different strains of influenza. Analysis of differential gene expression in cells is used here to determine the genetic response of cells to H5N1, benign viral infection by RSV, and to stressors found in regulating homeostasis. In a novel combination of the Gene Ontology database with a Human Network of biochemical interactions we have used these gene expression profiles to identify significant sub-networks characterizing early host responses. Characteristics of H5N1 infection compared to RSV infection show several factors that may contribute to increased virulence. These include faster timescales within the cell as well as a more focused activation of immunity factors. Many of the genes that are found to be significantly expressed in H5N1 response relative to the control trials are not found to cluster significantly in the Gene Ontology. These genes are, however, often closely linked to the clustered genes through the Human Network. We have further used RNA interference experiments to identify biochemical networks relevant for host defense and survival versus virus replication and host cell death. The corresponding response networks complement the gene expression data and identify specific sub-processes important for each phenotype. Host defense involves cell cycle control, inhibition of apoptosis and cell respiration. Virus survival includes B-cell signaling, Golgi trafficking and cytoskeleton control.

OP43: BOOLEAN NETWORK MODELS OF HUMAN AGING

Presenter: Michael Verdicchio, Arizona State University

Authors: Michael Verdicchio, Seungchan Kim

ABSTRACT: The systems biology of human aging is a complex, quantitative process. Many theories regarding senescence involve the roles of cellular components, such as mitochondria and lysosomes, as well the transportation and accumulation of various entities within and without the cell. In recent years, work by John Furber has amalgamated the research of many prominent aging biologists into a large chart illustrating many of the leading theories on human aging. The chart is organized into cellular components and describes many intricate, quantitative processes, along with their input and output entities. The representation, however, is not formalized as it is designed to be read and interpreted by humans. Boolean networks, which were first introduced by Kauffman and more recently applied to systems biology by Shmulevich et al., are a model well-suited in application to aging studies. The ability to interpret attracting states and their basins of attraction in light of their biological meanings could greatly simplify our understanding of essential processes in human aging. We construct a Boolean network representation of part of Furber's chart and perform a new type of analysis on the systems biology of aging through the exploration of Boolean network attractors and their basins. Preliminary results have shown a division of attractor states into healthy and unhealthy sides and analysis of essential variables within the large basins of attraction have revealed key entities and processes responsible for leading to these healthy and unhealthy attractors. Our current expansion of the model and collaboration with biologists will facilitate further understanding.

OP44: POSITIVE SELECTION AND FUNCTIONAL SHIFT: HUMAN HEME PEROXIDASE CASE STUDY

Presenter: Mary J. O'Connell, Dublin City University

Authors: Noeleen B. Loughran, Brendan O'Connor, Ciaran O'Fagain, William M. Nauseef, Mary J. O'Connell

ABSTRACT: A long-standing question in molecular evolution is whether computationally predicted positively selected residues cause an advantageous functional shift. Last year we presented the results of our analysis of positive selection in the mammalian heme peroxidase enzymes. They are a diverse group of enzymes involved in a number of defense and hormone related processes, and they play a major role in asthma, Alzheimers disease and inflammatory vascular disease. We are interested in determining the amino acid residues responsible for their specificities. Based on the resolved phylogeny for these enzymes we pinpointed the amino acid positions that have most likely contributed to their diverse functions (published last year). Many of these residues are in close proximity to sites implicated in protein misfolding, loss of function or disease. Here we address the link between positive selection and advantageous functional shift using in vitro

methods of analyses. We have created mutants for those positions predicted to be under positive selection. We have performed in depth biochemical analyses on the mutants using animal cell culture, including biosynthesis and activity studies. Our results indicate that (i) these mutants are not synthesized and processed efficiently, and (ii) they therefore have negligible levels of activity. We present our results of these biochemical studies here.

OP45: DELVING INTO THE BACTERIOME: PROTEIN-PROTEIN INTERACTIONS IN E. COLI

Presenter: John Parkinson, Hospital for Sick Children

Authors: Jose Peregrin-Alvarez, Xuejian Xiong, Chong Su, John Parkinson

ABSTRACT: It is widely appreciated that genes and proteins do not operate in isolation, but form components of highly integrated biological processes. Identifying the connections between these components is therefore critical to understanding how these processes are organized and function. *E. coli* is the leading model bacterium, however despite its importance in biological and medical discovery, a lack of large scale high quality interaction data has largely precluded a global 'systems' view of its genes and protein products. Here we describe an integrative approach, utilizing a Bayesian framework, that combines existing experimental and computational datasets to derive a highly reliable network of protein interactions that encompasses almost 50% of the *E. coli* proteome. While similar approaches have been applied to more limited datasets, rigorous statistical comparisons reveal that our approach results in a significant and biologically meaningful gain in performance. Combining our dataset with a set of recently generated experimentally derived interactions, we systematically organize these data into discrete functional modules to reveal known protein complexes and biochemical pathways. Finally we propose a new model of bacterial network evolution based on the integration of foreign genes acquired through horizontal gene transfer mechanisms. Together these data provide a comprehensive overview of the modular organization of the *E. coli* proteome and yield unique insights into functional and evolutionary relationships in bacterial networks.

OP46: SIMPLIFIED CLUSTERING WITH DIRICHLET PROCESS AND OTHER PROCESS MIXTURES

Presenter: Matthew Shotwell, Medical University of South Carolina

Authors: Matthew Shotwell, M.S., Elizabeth Slate, Ph.D.

ABSTRACT: Longitudinal clustering is a valuable tool for bioinformatics applications, for example in clustering gene expression probes with similar profiles in a microarray time series. Inferences drawn from such clustering aid in elucidating novel gene functions. The Dirichlet process (DP) mixture model is well suited to longitudinal clustering when the number of clusters is unknown, as inference on the number of clusters is a natural consequence of the model.

Bayesian inference for DP mixture models is dominated by posterior sampling through Markov chain Monte Carlo (MCMC) methods. These methods are computationally expensive, require expertise, and yield results that are burdensome to interpret in the context of clustering. We present a Bayesian inference mechanism for the DP mixture model that is based on optimization of a posterior likelihood, rather than sampling via MCMC. The method introduces cluster indicators and conditions on their maximum a posteriori (MAP) estimate. The resulting ‘profile’ posterior distributions for the longitudinal models are generally more tractable than the joint posterior distribution. In addition to the DP mixture, we present profile inference in a Dirichlet motivated process mixture, generated by modifying the prior distribution for the cluster indicators. This alternative process mixture is shown to have superior properties in many clustering problems. Profile inference in these models is evaluated through a simulated data example and with experimental data from a yeast cell cycle time series. Software implementation is available through the R package `profdpm`.

OP47: REAL TIME METAGENOMICS

Presenter: Robert Edwards, San Diego State University

Authors: Robert Edwards, Robert Olson, Terry Disz, Rick Stevens, Ross Overbeek

ABSTRACT: Metagenomics, extracting DNA from the environment and sequencing en masse, is revolutionizing microbiology. From health related studies through microbial ecology, all studies are being enhanced by our ability to sequence and characterize bacteria without growing them in the laboratory. Advances in next-generation sequencing technology have moved the bottle neck from extract and sequencing the DNA, to bioinformatics analysis of the samples. Typically, processing samples using BLAST will takes days to weeks. We have developed new technology to analyze and annotate metagenomes in minutes rather than weeks. Our new approach has opened the door to unique tools and techniques for data mining, and to providing real-time, in the field, sequence analysis. The improved algorithm, Web services interfaces, and web pages for data analysis are presented.

OP48: STRUCTURE DISCOVERY IN PPI NETWORKS USING PATTERN-BASED NETWORK DECOMPOSITION

Presenter: Ying Liu, University of Texas, Dallas

Authors: Ying Liu, Chengcheng Shen, Phil Bachman

ABSTRACT: The large, complex networks of interactions between proteins provide a lens through which one can examine the structure and function of biological systems. Previous analyses of these continually growing networks have primarily followed either of two approaches: large-scale statistical analysis of holistic network properties, or small-scale analysis of local topological features. Meanwhile, investigation of meso-scale network structure (above that of individual functional modules, while maintaining the significance of individual proteins) has been

hindered by the computational complexity of structural search in networks. Examining protein-protein interaction (PPI) networks at the mesoscale may provide insights into the presence and form of relationships between individual protein complexes and functional modules. Results: In this article, we present an efficient algorithm for performing sub-graph isomorphism queries on a network and show its computational advantage over previous methods. We also present a novel application of this form of topological search which permits analysis of a network's structure at a scale between that of individual functional modules and that of network-wide properties. This analysis provides support for the presence of hierarchical modularity in the PPI network of *Saccharomyces cerevisiae*.

OP49: CANCELLED

Poster Presentations Schedule
ELDORADO ROOM (3RD FLOOR)
Thursday, December 10

3:00 PM–6:00 PM PRESENTERS SET UP POSTERS (Maximum size 4x4 ft.)

Friday, December 11

9:00 AM–12:00 PM PRESENTERS SET UP POSTERS (Maximum size 4x4 ft.)

12:00 PM–5:45 PM POSTER VIEWING (room open, no authors present)

5:45 PM–8:00 PM POSTER SESSION (author present)

Saturday, December 12

10:35 AM–12:00 PM POSTER SESSION (author present)

12:00 PM REMOVE POSTERS
 (all posters must be removed at 12:00 p.m.)

Poster Presentations

ELDORADO ROOM (3RD FLOOR)

(Listed in alphabetical order by presenter's last name)

GENOMEWIDE HAPLOTYPE ASSOCIATION ANALYSIS IN SUB-THRESHOLD REGIONS

Presenter: Ryan Abo, University of Utah

Authors: Ryan Abo, Nicola J Camp

NOVEL METHOD FOR MICRORNA TARGET PREDICTION USING A GENETIC ALGORITHM

Presenter: Andrea Acquaviva, Politecnico di Torino

Authors: Paula Helena Reyes Herrera, Andrea Acquaviva, Elisa Ficarra, Enrico Macii

MODELING AND DESIGN OF PEPTIDO-MIMETIC COMPOUNDS TO BLOCK NOGGIN/BMP, ACTIVIN/FOLLISTATIN AND CROSSVEINLESS 2/BMP INTERACTIONS TO PROMOTE OSTEOGENESIS

Presenter: Shaila Ahmed, Queens College, CUNY

Authors: Shaila Ahmed, Boojala Vijay B Reddy, Sreedhara Sangadala, Sanjay Kumar

ASSEMBLER FOR SOLID DATA: BY IMPROVING MEMORY MANAGEMENT OF VELVET ASSEMBLER

Presenter: Sajja Akhter, San Diego State University

Authors: Sajja Akhter, Robert Edwards

APPLICATION OF DATA INTEGRATION METHODS TO AID IN THE DEVELOPMENT OF SECOND GENERATION BIOFUELS

Presenter: David Astling, National Renewable Energy Laboratory

Authors: David Astling, Peter Graf, Kofi Adragani, Jinsuk Lee, Mark Davis

PHANTOME (PHAGE ANNOTATION TOOLS AND METHODS): A PLATFORM FOR PHAGE ANNOTATION AND COMPARATIVE GENOMICS

Presenter: Ramy K. Aziz, San Diego State University

Authors: Ramy K. Aziz, Bhakti Dwivedi, Joe Anderson, Bonnie Hurwitz, JP Massar, Matthew Sullivan, Jeff Elhai, Mya Breitbart, Robert Edwards

PREDICTION OF PROTEIN PROTEIN INTERACTION SITES AND THEIR IMPACT ON GENETIC DISEASE

Presenter: Angshuman Bagchi, Buck Institute for Age Research

Authors: Angshuman Bagchi, Eunseog Youn, Matthew E. Mort, David N. Cooper, Sean D. Mooney

DEFINING WEB-SERVICE BASED BIOLOGICAL ANALYSIS WORKFLOWS

Presenter: Janaka Balasooriya, Arizona State University

Authors: Janaka Balasooriya, Swanson Morgan, Graciela Gonzalez

IDENTIFICATION OF NOVEL EPITOPES OF THE EBOLA VIRUS FOR RATIONAL VACCINE DESIGN

Presenter: Sophia Banton, Florida Atlantic University

Authors: Sophia Banton, Zvi Roth PhD

ASSESSING THE PREDICTIVE POWER OF SITE-PREDICTION METHODS FOR IDENTIFYING POSITIVE SELECTION WITHIN EMPIRICAL DATASETS

Presenter: Matthew L. Bendall, Brigham Young University
 Authors: Matthew L. Bendall, Matthew Dyer, Keith A. Crandall

CHARACTERIZATION AND ASSEMBLY OF THE FIRST SNAKE GENOME USING MULTI-PLATFORM NEXT-GENERATION SEQUENCE DATA

Presenter: Todd A. Castoe, University of Colorado, Denver
 Authors: Todd Castoe, Matthew La Bella, A. P. Jason de Koning, Kathryn Hall, Wanjun Gu, Peter Uetz, David D. Pollock

VISUALIZING METAGENOMES

Presenter: Nicholas Celms, San Diego State University
 Authors: Nicholas Celms, Elizabeth Dinsdale, Robert Edwards

PHYLOGENETIC ANALYSIS OF PLANT SESQUITERPENE SYNTHASES

Presenter: Brian Y. Chen, National Resource for Biomedical Supercomputing
 Authors: Brian Y. Chen, Ashley Young, Hugh B. Nicholas Jr, Alexander J. Ropelewski, Troy Wymore

SYSTEMATIC DRUG TARGET DISCOVERY VIA CHEMICAL AND GENETIC INTERACTION PROFILES

Presenter: Hon Nian Chua, Harvard Medical School
 Authors: Hon Nian Chua, Murat Cokol, Yo Suzuki, Frederick P. Roth

ON THE ACCURACY OF AUTOMATED INFERENCE OF PROTEIN FUNCTION

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

ALGORITHM TO IMPROVE GENE CONSISTENCY ACROSS BACTERIAL GENOMES

Presenter: Judith D. Cohn, Los Alamos National Laboratory
 Authors: Judith D. Cohn, Michael E. Wall, John Dunbar

BIOINFORMATICS CHARACTERIZATION OF THE PLASMODIUM GLUTATHIONE S-TRANSFERASE

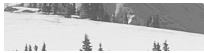
Presenter: Emilee Colón, University of Puerto Rico
 Authors: Emilee Colón, Adelfa Serrano, Hugh Nicholas Jr, Troy Wymore, Alexander Ropelewski, Ricardo González Méndez

KNOWLEDGE NETWORK APPROACH: PATHWAYS AND DRUGS

Presenter: Nikolai Daraselia, Ariadne
 Authors: Nikolai Daraselia, Ekaterina Kotelnikova and Anton Yuryev

MULTIVARIATE ANALYSIS OF METAGENOMES – AN UNDERGRADUATE REU STORY

Presenter: Elizabeth Dinsdale, San Diego State University
 Authors: Naneh Apkarian, Michelle Creek, Eric Guan, Mayra Hernandez, Kate Isaacs, Chris Peterson, Todd Regh, Robert Edwards, Barbara Bailey, Peter Salamon, Imre Tuba, Elizabeth Dinsdale



A COMPUTATIONAL TOOL FOR THE IDENTIFICATION OF SIGNATURE GENES AMONG PHAGES

Presenter: Bhakti Dwivedi, University of South Florida
Authors: Bhakti Dwivedi, Robert Edwards, Mya Breitbart

REAL TIME METAGENOMICS

Presenter: Robert Edwards, San Diego State University
Authors: Robert Edwards, Robert Olson, Terry Disz, Rick Stevens, Ross Overbeek

BIOINFORMATIC ANALYSIS OF SDS-INSOLUBLE PROTEIN AGGREGATES IN S. CEREVISIAE, C. ELEGANS, AND M. MUSCULUS.

Presenter: Uday S Evani, Buck Institute
Authors: Uday S. Evani, Theodore W. Peters, Pedro Rodrigues, Gregg Czerwieniec, Sean D. Mooney, Gordon Lithgow, Bradford Gibson, Robert Hughes

PAIRWISE AND HIGHER-ORDER CORRELATIONS AMONG DRUG-RESISTANCE MUTATIONS IN HIV-1 SUBTYPE B PROTEASE

Presenter: Omar Haq, Rutgers University
Authors: Omar Haq, Ronald M Levy, Alexandre V Morozov, Michael Andrec

A USEFUL TOOL FOR CALCULATING BINDING-SITE RESIDUES ON PROTEINS FROM PDB STRUCTURES

Presenter: Jing Hu, Franklin & Marshall College
Authors: Jing Hu, Changhui Yan

A FAST APPROACH TO PROTEIN STRUCTURE ALIGNMENT BASED ON ONE-DIMENSIONAL ALPHABET CODE SEQUENCES

Presenter: Kenneth Hung, National Taiwan University
Authors: Kenneth Hung, Jui-Chih Wang, Kun-Nan Tsai, Cheng-Wei Chen, Chung-Ming Chen

DEALING WITH DATA DELUGE: DESIGNING AND IMPLEMENTING A DATABASE TO ENABLE SPECIALIZED STUDIES IN METAGENOMICS.

Presenter: Bonnie Hurwitz, University of Arizona
Authors: Bonnie Hurwitz, Matthew Sullivan, Robert Edwards, Adam Monier, Alexandra Z. Worden, Sudha Ram

VIRAL METAGENOMICS IN MARINE MICROBIAL SAMPLES

Presenter: Julio C. Ignacio, University of Arizona
Authors: Julio C. Ignacio, Alexandra Z. Worden, Matthew B. Sullivan

USING SYNTACTIC CONTEXT IN OPENDMAP PATTERNS

Presenter: Helen Johnson, University of Colorado, Denver
Authors: Helen L. Johnson, William Baumgartner Jr, Christophe Roeder, Karin Verspoor, Kevin Bretonnel Cohen, Larry Hunter

IN SILICO FUNCTIONAL PROFILING OF HUMAN DISEASE-ASSOCIATED AND POLYMORPHIC AMINO ACID SUBSTITUTIONS

Presenter: Vidhya G. Krishnan, Buck Institute for Age Research
Authors: Matthew Mort, Uday S. Evani, Vidhya G. Krishnan, Kishore K. Kamati Peter H. Baenziger, Anghuman Bagchi, Brandon Peters, Rakesh Sathyesh, Biao Li, Yanan Sun, Bin Xue, Nigam Shah, Maricel Kann, David N. Cooper, Predrag Radivojac, Sean D. Mooney

A USER STUDY OF ATTRIBUTE VISUALIZATION TOOLS AND THEIR ROLE IN UNDERSTANDING BIOLOGICAL NETWORKS

Presenter: Hande Kucuk, Eastern Michigan University
 Authors: Hande Kucuk, Benjamin J. Keller, Terry Weymouth, Barbara Mirel

AUTOMATED INFERENCE OF MOLECULAR MECHANISMS OF DISEASE FROM AMINO ACID SUBSTITUTIONS

Presenter: Biao Li, Indiana University
 Authors: Biao Li, Vidhya G. Krishnan, Matthew E. Mort, Fuxiao Xin, Kishore K. Kamati, David N. Cooper, Sean D. Mooney, Predrag Radivojac

LOSS OF POST-TRANSLATIONAL MODIFICATION SITES IN DISEASE

Presenter: Shuyan Li, Indiana University, Bloomington
 Authors: Shuyan Li, Lilia M. Iakoucheva, Sean D. Mooney, Predrag Radivojac

MACHINE LEARNING APPROACHES TO THE ASSESSMENT OF PEPTIDE-SPECTRUM MATCHES WITHOUT USING A DECOY DATABASE

Presenter: Yong Fuga Li, Indiana University, Bloomington
 Authors: Yong Fuga Li, Randy J. Arnold, Predrag Radivojac, Haixu Tang

STRUCTURE DISCOVERY IN PPI NETWORKS USING PATTERN-BASED NETWORK DECOMPOSITION

Presenter: Ying Liu, University of Texas, Dallas
 Authors: Ying Liu, Chengcheng Shen, Phil Bachman

PROMOTER PREDICTION IN HALOTHIOBACILLUS NEAPOLITANUS C2 BASED ON STRESS-INDUCED DNA DUPLEX DESTABILIZATION

Presenter: Aleksandra Markovets, Mississippi Valley State University
 Authors: Aleksandra Markovets, Charles Bland, Abigail Newsome

TOWARDS REALISTIC CODON MODELS: AMONG-SITE VARIABILITY AND DEPENDENCY OF SYNONYMOUS AND NON-SYNONYMOUS SUBSTITUTION RATES

Presenter: Itay Mayrose, University of British Columbia
 Authors: Itay Mayrose, Adi Doron-Faigenboim, Eran Bacharach, Tal Pupko

THE HUMAN GENE MUTATION DATABASE (HGMD) AND ITS EXPLOITATION IN THE ERA OF PERSONALIZED GENOMICS

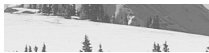
Presenter: Matthew Mort, Cardiff University
 Authors: P.D. Stenson, M. Mort, E. Ball, K. Howells, A. Phillips, N.S.T. Thomas, D.N. Cooper

META-ANALYTICAL TOOLS FOR DECIPHERING TRANSCRIPTIONAL NETWORKS IN A MODEL ANOXYGENIC PHOTOTROPH

Presenter: Oleg Moskvina, University of Wyoming
 Authors: Oleg Moskvina, Dmitry Bolotin, Pavel Ivanov, Mark Gomelsky

THE EVOLUTIONARY LANDSCAPE OF CHROMATIN MODIFICATION MACHINERY

Presenter: Tuan On, University of Toronto
 Authors: Tuan On, Xuejian Xiong, Shuye Pu, Andrei Turinsky, Yunchen Gong, Andrew Emili, Zhaolei Zhang, Jack Greenblatt, Shoshana J. Wodak, John Parkinson



BACTERIAL EVOLUTION: IMPLICATION FROM LIPID ABIOSYNTHESIS PATHWAY ENZYMES

Presenter: Stephen O. Opiyo, University of Nebraska, Lincoln

Authors: Stephen Opiyo, Rosevelt Pardy, Hideaki Moriyama, Etsuko Moriyama

CNGEN – A NEW TOOL FOR COPY-NUMBER GENOTYPES PARTITIONING

Presenter: Louis-Philippe Lemieux Perreault, Montreal University

Authors: Louis-Philippe Lemieux Perreault, Gregor Andelfinger, Géraldine Asselin, Marie-Pierre Dubé

QIIME (QUANTITATIVE INSIGHTS INTO MICROBIAL ECOLOGY) DYNAMIC RAREFACTION GRAPH

Presenter: Megan Pirrung, University of Colorado, Boulder

Authors: Megan Pirrung, Rob Knight

STRUCTURE-BASED PREDICTION OF DNA BINDING SITES FOR FAMILIES OF TRANSCRIPTION FACTORS

Presenter: Julia Ponomarenko, University of California, San Diego

Authors: Julia Ponomarenko

AUTOMATED IDENTIFICATION OF AMPLIFIABLE MICROSATELLITE LOCI AND PRIMER DESIGN FROM 454 HIGH-THROUGHPUT SEQUENCING READS

Presenter: Alex Poole, University of Colorado, Denver

Authors: Alex Poole, Todd Castoe, David Pollock

ANALYSIS OF A LOCAL HUNTINGTIN PROTEIN INTERACTION NETWORK

Presenter: Corey Powell, Buck Institute for Age Research

Authors: Corey Powell, Robert Hughes, Cendrine Tourette, Russell Bell, Sean Mooney

FLUX BALANCE ANALYSIS OF PLASMODIUM FALCIPARUM'S METABOLIC NETWORK

Presenter: Farhan Raja, University of Toronto

Authors: Farhan Raja, John Parkinson, James Wasmuth, Stacy Hung

GENEBOOK, A HIGH PRECISION MAMMALIAN PROTEIN THESAURUS

Presenter: Phoebe Roberts, Pfizer, Inc.

Authors: Robert Hernandez, Markella Skempri, Phoebe Roberts

INTEGRATING WEB SERVICES INTO BIOMEDICAL TEXT MINING.

Presenter: Christophe Roeder, University of Colorado, Denver

Authors: Christophe Roeder, William Baumgartner Jr, Larry Hunter

INTEGRATION OF DIYA OUTPUT WITH GMOD STANDARDS

Presenter: Inna Rytsareva, Mississippi Valley State University

Authors: Inna Rytsareva, Charles Bland, Abigail Newsome

PROTEIN SECONDARY STRUCTURE IS ROBUST UNDER ARTIFICIAL EVOLUTION WHILE PROTEIN DISORDER IS NOT

Presenter: Christian Schaefer, Columbia University, New York City

Authors: Christian Schaefer, Avner Schlessinger, Burkhard Rost

VISUALIZING GENOMIC SEQUENCES IN 2D

Presenter: Josiah Seaman, Colorado State University
Authors: Josiah Seaman

IMPROVING THE PREDICTION OF MICRORNA-TARGET GENES BY COMBINING TARGET PREDICTION ALGORITHMS

Presenter: Ashok Sharma, Medical College of Georgia
Authors: Ashok Sharma, Ryan Rimando, Richard McIndoe

SIMPLIFIED CLUSTERING WITH DIRICHLET PROCESS AND OTHER PROCESS MIXTURES

Presenter: Matthew Shotwell, Medical University of South Carolina
Authors: Matthew Shotwell, Elizabeth Slate

EVOLVING SPIKING NEURAL NETWORKS FOR THE PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES

Presenter: Heike Sichtig, UF Genetics Institute
Authors: Heike Sichtig, Alberto Riva

MCW PROTEOMICS ANALYSIS PLATFORM – A HYBRID CLOUD COMPUTING ARCHITECTURE FOR PROTEOMICS ANALYSIS

Presenter: Simon Twigger, Medical College of Wisconsin
Authors: Andrew Vallejos, Brian Halligan, Joey Geiger, Andrew Greene, Simon Twigger

ACCELERATING CANDIDATE GENE DISCOVERY THROUGH ONTOLOGICAL INDEXING OF LARGE SCALE DATA REPOSITORIES

Presenter: Simon Twigger, Medical College of Wisconsin
Authors: Simon Twigger, Joey Geiger, Jennifer Smith

DIGGING FOR GOLD – DATA ANNOTATION AND EXPLORATION WITH RATMINE

Presenter: Simon Twigger, Medical College of Wisconsin
Authors: Andrew Vallejos, Jennifer Smith, Richard Smith, Julie Sullivan, Gos Micklem, Simon Twigger

AN ALGORITHM TO DESCRIBE ALL OPTIMAL COMPARATIVE GENOME MAPS

Presenter: Zachary Vaughan, University of Colorado, Boulder
Authors: Zachary D. Vaughan, Debra S. Goldberg

LARGE-SCALE SEQUENCING OF T-CELL RECEPTOR REPERTOIRES IN DIABETIC NOD MICE

Presenter: Vijetha Vemulapalli, University of Colorado, Denver
Authors: Vijetha Vemulapalli, Todd A Castoe, Maki Nakayama, George Eisenbarth, David D Pollock

BOOLEAN NETWORK MODELS OF HUMAN AGING

Presenter: Michael Verdicchio, Arizona State University
Authors: Michael Verdicchio, Seungchan Kim

EVOLUTION OF A PLACENTA SPECIFIC REGULATORY NETWORK

Presenter: Thomas Walsh, Dublin City University

Authors: Thomas Walsh, Kieran Holohan, Anna O'Brien, Elinor Velasquez, Mary O'Connell

PATTERN-BASED EXTRACTION OF ARGUMENTATION FROM THE SCIENTIFIC LITERATURE

Presenter: Elizabeth White, University of Colorado, Boulder

Authors: Elizabeth White

AN INDIVIDUAL BASED MODELLING APPROACH TO STUDYING THE EVOLUTION OF MATE CHOICE STRATEGY

Presenter: Robert Williamson, Rose-Hulman Institute of Technology

Authors: Robert Williamson

EVALUATING GENE-DISEASE ASSOCIATION PREDICTIONS

Presenter: Laura Wojtulewicz, Arizona State University

Authors: Graciela H. Gonzalez, Fabian Spinnherin, Juan C. Uribe, Annie Skariah, Laura Wojtulewicz, Hailong Cui

FACTORS THAT CONTROL FUNCTIONALITY OF SESQUITERPENE SYNTHASES FROM PHYLOGENETIC AND BIOPHYSICAL SIMULATIONS

Presenter: Troy Wymore, National Resource for Biomedical Supercomputing

Authors: Troy Wymore, Brian Y. Chen, Hugh B. Nicholas Jr, Alexander J. Ropelewski, Charles L. Brooks III

STRUCTURE-BASED KERNELS FOR THE PREDICTION OF CATALYTIC RESIDUES AND THEIR INVOLVEMENT IN DISEASE

Presenter: Fuxiao Xin, Indiana University, Bloomington

Authors: Fuxiao Xin, Steven Myers, Yong Fuga Li, David N. Cooper, Sean D. Mooney, Predrag Radivojac

EVOLUTIONARY STUDY AND PREDICTION OF PROTEIN-PROTEIN INTERACTIONS IN CHROMATIN MODIFICATION COMPLEXES

Presenter: Xuejian Xiong, Hospital for Sick Children

Authors: Xuejian Xiong, Tuan On, Shuye Pu, Andrei Turinsky, Yunchen Gong, Andrew Emili, Zhaolei Zhang, Jack Greenblatt, Shoshana J. Wodak, John Parkinson

SIMPLEMHC: A NOVEL METHOD FOR IN-SILICO PREDICTION OF PEPTIDE BINDING TO MHC CLASS II MOLECULES

Presenter: Li Xue, Iowa State University

Authors: Li Xue, Arthur Fridman

PROTEIN-PROTEIN INTERACTIONS ARE DRIVEN BY FUNCTIONAL EVOLUTION

Presenter: Yiqiang Zhao, Buck Institute for Age Research

Authors: Yiqiang Zhao, Sean Mooney

GENOMEWIDE HAPLOTYPE ASSOCIATION ANALYSIS IN SUB-THRESHOLD REGIONS

Presenter: Ryan Abo, University of Utah
 Authors: Ryan Abo, Nicola J. Camp

ABSTRACT: Haplotype analysis is potentially a powerful method for establishing genotype-phenotype associations in genome-wide association (GWA) studies. In particular, the ability to identify new regions that single SNP analyses do not detect will be important. However, current haplotype analysis methods that scale to a genome-wide level are limited and have only been developed for independent samples. We present a two-stage strategy which explores haplotypes in loci that are sub-threshold in single SNP analyses. In the first step, single marker associations are tested to identify sub-threshold regions of interest and then regions are scored across the genome based on the number of SNPs in the region that achieved a significance between 10^{-2} and 10^{-4} . Haplotype analysis within sub-threshold regions was performed using hapConstructor, a software developed to efficiently data-mine haplotypes that implements a stepwise forward-backward procedure to search for haplotypes associated with disease. We used the Genetic Analysis Workshop 16 Problem Set 2 and analyzed data from the Framingham 500K SNP set for association with high-density lipoprotein (HDL) cholesterol levels. We were able to identify a two-locus haplotype on chromosome 5 (p -value = 8.0×10^{-6}), illustrating the possibility for associated haplotypes within sub-threshold GWAS regions.

NOVEL METHOD FOR MICRORNA TARGET PREDICTION USING A GENETIC ALGORITHM

Presenter: Andrea Acquaviva, Politecnico di Torino
 Authors: Paula Helena Reyes Herrera, Andrea Acquaviva, Elisa Ficarra, Enrico Macii

ABSTRACT: MicroRNAs are non-coding RNAs involved in gene regulation. Nowadays particular computational approaches for binding site prediction have gained importance in the field considering the quantity of available experimental data. The microRNA target prediction have been undertaken by target-oriented approaches. We present an innovative and flexible method for the target prediction, able to overcome limitations of traditional approaches, proposing a miRNA oriented approach. The method explores the space of possible binding sites using a Genetic Algorithm (GA) and drives the target search using microRNA derived features (structural properties, expression information, and functional correlation between miRNA and target genes) considered in the GA fitness function. The method was validated against experimental data and frequently used methods, obtaining improved specificity and sensitivity. Further target predictions were done for conserved microRNAs in different species showing the effectiveness of the proposed method for cross-genomic explorations. In addition, the proposed method lends itself to be refined in a target-specific way by adding target-dependent features inside the algorithm (GA) in order to give hints toward microRNA binding sites

characteristics for a given set of target genes correlated in the same biological process. It was possible to evaluate the effect of the inclusion of these new features in the predictions (in particular were used sequence motifs as an initial approach). The key idea is to be able to predict possible targets, independent on the amount of available data but at the same time be able to easily incorporate new

MODELING AND DESIGN OF PEPTIDO-MIMETIC COMPOUNDS TO BLOCK NOGGIN/BMP, ACTIVIN/FOLLISTATIN AND CROSSVEINLESS 2/BMP INTERACTIONS TO PROMOTE OSTEOGENESIS

Presenter: Shaile Ahmed, Queens College, CUNY

Authors: Shaile Ahmed, Boojala Vijay B Reddy, Sreedhara Sangadala, Sanjay Kumar

ABSTRACT: TGF- β superfamily members such as bone morphogenic proteins (BMPs), activins, inhibins execute distinct and intricate roles in numerous biological events such as cell growth, differentiation, embryogenesis etc. When they exert their biological activity, they are sternly regulated by extracellular antagonists such as noggin, follistatin, crossveinless 2 (CV2) etc. Blocking these antagonists' interactions with their target receptor proteins help to promote their respective biological responses when needed. This work is an attempt to use the current understanding of some of the receptors and ligands of TGF-beta superfamily members, namely BMPs, and activins and their antagonists such as noggin, CV2, and follistatin interactions through the analyses of their available complexes to develop modified peptide-mimetic compounds with an ultimate goal of promoting their respective biological responses to enhance osteogenesis. Designing peptidomimetic compounds is an alternative approach in developing potential drugs instead of using small molecular databases. In our study, we identified several, di-, tri- and tetra-peptides derived from the key binding regions of BMP-2 and activin that are in contact with their corresponding antagonists noggin, CV2 and follistatin based on the analyses on the BMP-2/noggin, BMP-2/CV2 and activin/follistatin complex structures. We modeled structures of these peptide sequences and performed some preliminary experiments using Glide and GOLD to dock these peptides onto respective binding sites on the corresponding antagonists' surfaces. We analyzed the scores from different docking procedures and compared the binding residues of noggin and CV2 to BMP-2 and follistatin to activin. This study is simply the initiative for composing modified peptidomimetics to accomplish the proposed goal.

ASSEMBLER FOR SOLID DATA: BY IMPROVING MEMORY MANAGEMENT OF VELVET ASSEMBLER

Presenter: Sajia Akhter, San Diego State University

Authors: Sajia Akhter, Robert Edwards

ABSTRACT: The SOLiD' System is a highly accurate next-generation sequencing technology that supports a wide range of applications. This platform delivers very short length high volume sequences which are significantly different from the

traditional sequencing data in both read length and volume. So it is challenging to adapt these data to many applications, including de novo assembly. There are couple of different assembler like VELVET, SHARCGS, EULER-SR, ALLPATHS 2 which can assemble solexa data very nicely. As both solid data and solexa data has same read length but the volume of the data in Solid technology is much higher than solexa, these assemblers might not perform well for Solid data. Here we are in a progress of the improvement of the velvet assembler so that it can assemble Solid data. The algorithms for velvet assembler work very fast but memory is a big issue for this assembler. There are four different steps in velvet assembler: sequencing, hashing, simplification of linear stretches and error removal. Most of the memory is required for sequencing and hashing. We are able to improve 54% memory per read in sequencing step and more than 50% memory in hashing step.

APPLICATION OF DATA INTEGRATION METHODS TO AID IN THE DEVELOPMENT OF SECOND GENERATION BIOFUELS

Presenter: David Astling, National Renewable Energy Laboratory

Authors: David Astling, Peter Graf, Kofi Adragani, Jinsuk Lee, Mark Davis

ABSTRACT: Biomass derived from plants such as poplar has been identified as one of the key feedstocks for the development of second generation biofuels. In order to convert woody biomass to fermentable sugars we must have a good understanding of the genetic and metabolic processes involved in cell wall biosynthesis. We have used high-throughput techniques such as microarrays and molecular beam mass spectrometry (MBMS) to elucidate metabolic pathways and signal transduction cascades involved in altering cell wall chemistry. To integrate and draw correlations amongst these large py-MBMS and microarray data sets, we are exploring three published multi-variate statistical methods; O2PLS, Principal Fitted Components and Elastic Net. i) The O2PLS method, a 2-way Orthogonal correction to Partial Least Squares (PLS) regression, is ideally suited for analyzing information across multiple platforms because it separates the systematic orthogonal variation from the joint covariation for both datasets. ii) The Principal Fitted Components (PFC) method shares some similarity to the O2PLS method by estimating the joint variation in both sets of data. In addition the PFC method can be used to obtain a sufficient reduction of the data, i.e. fewer number of variables that can be used to simplify further analysis. iii) And finally the Elastic Net is an alternative method for obtaining correlations between each data set. We are currently comparing and evaluating the results of each method with several test cases.

PHANTOME (PHAGE ANNOTATION TOOLS AND METHODS): A PLATFORM FOR PHAGE ANNOTATION AND COMPARATIVE GENOMICS

Presenter: Ramy K. Aziz, San Diego State University

Authors: Ramy K. Aziz, Bhakti Dwivedi, Joe Anderson, Bonnie Hurwitz, JP Massar, Matthew Sullivan, Jeff Elhai, Mya Breitbart, Robert Edwards

ABSTRACT: Phages are the most abundant biological entities on Earth. Despite the indisputable importance of phages in the biosphere and their pivotal role in molecular biology, their genomes are poorly annotated. Having a consistent and accurate phage gene nomenclature is critical to phage research, and knowing which genes are viral versus microbial will help all researchers struggling to understand microbial genomes and metagenomes. Here, we present PhAnToMe (PHage ANnotation TOols and MEthods), a new platform for phage genome annotations and comparative phage genomics. PhAnToMe relies on the SEED database to handle the nuances of phages and prophages, develop a controlled vocabulary and a consistent nomenclature for phage genes, and create a new tool for the identification of prophages. PhAnToMe will provide high quality annotations to over 1,000 existing phage and prophage genomes and dozens of existing phage metagenomes. Because most biologists have little experience in computer programming, PhAnToMe also implements BioBIKE (Biological Integrated Knowledge/programming Environment), which makes use of familiar graphical conventions to facilitate problem solving. The workflows devised through the project will be made available to users through this interface, and likewise, workflows that users devise may be readily packaged and made available to others in the research community. Researchers may analyze the currently available genomes or bring their own sequences, analyze and annotate them with the advantage of comparative analysis provided by the resource. The tools and high quality annotation developed in this project will serve as a solid basis for future efforts to comprehend phage and microbial genomes.

PREDICTION OF PROTEIN PROTEIN INTERACTION SITES AND THEIR IMPACT ON GENETIC DISEASE

Presenter: Angshuman Bagchi, Buck Institute for Age Research

Authors: Angshuman Bagchi, Eunseog Youn, Matthew E. Mort, David N. Cooper, Sean D. Mooney

ABSTRACT: Protein-protein interactions play pivotal roles in many biological processes, for example, hormone-receptor binding and so on. There is increasing evidence that disease-causing mutations lead to disruption of protein interactions. In the present work, we used machine learning tools to discriminate between interface and non-interface residues of proteins using structure and sequence information. We generated features from protein sequence and structure using a non-redundant set of protein hetero-complexes from the protein data bank. The training dataset comprised of (A) interface residues and non-interface surface

residues for structure based prediction and (B) interface residues and non-interface surface and core residues for sequence based prediction. The datasets were used to build classification algorithms using random forest (RF) and support vector machine (SVM) coupled with 10 fold cross-validation for evaluation. Overall, RF outperformed SVM in most cases. The best performing sequence-based classification tool achieved an accuracy of 73% and when protein structure was included the accuracy was 75%. The predictors were then used to analyze mutation data including somatic mutations in cancer from tumor resequencing projects, the Human Gene Mutation Database (HGMD), and common human polymorphisms. The results showed an enrichment of protein interaction sites in the disease datasets compared to the neutral set (Seattle SNPs). Overall our results indicate that disease mutations are enriched in disruption of PPI interfaces and these interfaces can be predicted using bioinformatic approaches.

DEFINING WEB-SERVICE BASED BIOLOGICAL ANALYSIS WORKFLOWS

Presenter: Janaka Balasooriya, Arizona State University

Authors: Janaka Balasooriya, Swanson Morgan, Graciela Gonzalez

ABSTRACT: The advent of high-throughput technologies for gene expression analysis has revolutionized the way scientific analysis is conducted. In order to handle analysis of complex and large datasets, medical scientists turn to biostatisticians and bioinformaticians for help in selecting the optimal array of tools and methods. It is often the case that several steps of processing are required in order to take raw data from experiments and turn it into something that can assist the scientists in scientific discovery and data distribution. On average, about 80% of the invested time goes into assembling the right data to prepare for analysis and gathering resources for data processing from different universities, institutes, and centers around the world. Bringing all of these resources together presents challenges due to system, network, and application-level discrepancies. Web services and cloud computing have emerged as a platform for hiding system and network heterogeneity issues. SOA architecture separates functions into distinct unites, or services, which then can be utilized over a network or the World Wide Web. An increasing number of biological data bases and applications are available as web services, but it is still difficult to stitch together the services that are needed for a specific analysis workflow. Ideally, one would be interested in a 'plug and play' platform that can facilitate the definition and execution of such workflows, hiding any such discrepancies. In this talk we introduce BondFlow, an environment for configuring and executing workflows on the fly over heterogeneous web objects, including web services.

IDENTIFICATION OF NOVEL EPITOPES OF THE EBOLA VIRUS FOR RATIONAL VACCINE DESIGN

Presenter: Sophia Banton, Florida Atlantic University
 Authors: Sophia Banton, Zvi Roth PhD

ABSTRACT: Despite the overall success of global vaccination efforts there is still a great need for new and improved vaccines against numerous human pathogens. Rational vaccine design (RVD) seeks to manipulate the immune system to “work harder” via utilization of the cellular components that are heavily involved in antigen recognition. Ideally, such vaccines rely on epitopes — antigenic determinants usually made of proteins — that are recognized by the cellular arm of the immune system. The first step is therefore epitope identification and selection. Epitope prediction servers were used to elucidate novel vaccine targets towards Ebola, a hemorrhagic virus and one of the deadliest infectious agents known. Whole antigens of the Ebola virus, the glycoprotein and nucleoprotein, were computationally analyzed for antigenic peptide sequences. Specific individual epitopes were uncovered for both B cells and T cells and a universal sequence (a 9mer) ‘WIPYFGPAA’ from the Ebola glycoprotein was a potential candidate for both B and T cell activation. Another sequence with universal potential ‘AIVNAQ’ was extracted from a lone 3D structure of the Ebola virus glycoprotein bound to a human antibody and was also reported by numerous servers. The methodology is validated by consistent overlap with confirmed Ebola antigens in the Immune Epitope Database and Analysis Resource (IEDB) server. Analysis of the epitopes indicates that they were indeed ‘rationally’ selected by their structural and molecular properties and may have the potential to stimulate the cells of the adaptive immune system. Thus, these epitopes may be valid for future RVD against Ebola.

ASSESSING THE PREDICTIVE POWER OF SITE-PREDICTION METHODS FOR IDENTIFYING POSITIVE SELECTION WITHIN EMPIRICAL DATASETS

Presenter: Matthew L. Bendall, Brigham Young University
 Authors: Matthew L. Bendall, Matthew Dyer, Keith A. Crandall

ABSTRACT: Several computational methods have been proposed for identifying amino acid sites which evolve under positive Darwinian selection. We assess the predictive power of six site-prediction methods: (1) Suzuki and Gojobori’s counting method (SG99); (2) single likelihood ancestor counting (SLAC); (3) fixed effects likelihood (FEL); (4) random effects likelihood (M8); (5) dual rate random effects likelihood (DUAL); and (6) selection on amino acid properties (SAAP). For our model system, we used sequences from 139 longitudinal HIV drug resistance studies; the convergent evolution of drug resistant mutations (DRMs) is evidence of positive selection. All six methods lack statistical power when applied to our data, failing to recover even one-third of the positively selected sites. SAAP has the highest true positive rate, correctly identifying ‘30% of the known sites ($\alpha = 0.01$). DUAL identified 17% of the sites (BF = 50); all other

methods detected $< 5\%$ of the known sites. All six methods identified selection on sites where the actual selective pressure is unknown, possibly indicating poor specificity of the methods. To estimate a false positive rate, we generated 2,250 simulation datasets designed to mimic the empirical HIV datasets. Receiver operating characteristic analysis suggests that DUAL has the most predictive power as the discrimination threshold is varied. We demonstrate that current methods for identifying positively selected sites can be improved by using the AIC to select random effects likelihood models which best explain the data.

CHARACTERIZATION AND ASSEMBLY OF THE FIRST SNAKE GENOME USING MULTI-PLATFORM NEXT-GENERATION SEQUENCE DATA

Presenter: Todd A. Castoe, University of Colorado, Denver

Authors: Todd Castoe, Matthew La Bella, A. P. Jason de Koning, Kathryn Hall, Wanjun Gu, Peter Uetz, David D. Pollock

ABSTRACT: Snakes have become important model organisms for a broadening diversity of research, however, there is extremely little known about the genomes of snakes. To remedy this situation, we have begun an ambitious program to sequence multiple snake genomes, starting with that of perhaps the most important model species, the Burmese Python. Here, we report results of a preliminary genome assembly for the Python based on the combination of 454 and Illumina sequencing. Additionally, based on this sampling, we are able to characterize the repetitive content of the Python genome, which provides important data for comparison to other tetrapod genomes available.

VISUALIZING METAGENOMES

Presenter: Nicholas Celms, San Diego State University

Authors: Nicholas Celms, Elizabeth Dinsdale, Robert Edwards

ABSTRACT: The Line Islands offer a rare opportunity to study human influence on microbial lifeforms, due in part to the location of the islands (they're in the middle of the Pacific Ocean) but also to the varying levels of population across the islands. By collecting samples from sites with variable human exposure, metagenomes were sequenced that can be utilized to understand the evolutionary adjustments humanity imposed upon the microbial community. By creating an application to graph the various metagenomic samples, the areas of interest in the microbial genomes that have occurred with speciation were identified. This tool offers a quick way to view, study, and evaluate metagenomes. This has the value of helping to focus further research more narrowly. Multiple metagenomes were sequenced and genomic comparisons were done using BLAST. The beta version of the application serves as a stepping-stone to further biological and computational research. Further development goals with the application include broadening accessibility and applicability. The same approach will answer questions about many other environments, and could benefit other research groups studying microbial metagenomes worldwide.

PHYLOGENETIC ANALYSIS OF PLANT SESQUITERPENE SYNTHASES

Presenter: Brian Y. Chen, National Resource for Biomedical Supercomputing
Authors: Brian Y. Chen, Ashley Young, Hugh B. Nicholas Jr, Alexander J. Ropelewski, Troy Wymore

ABSTRACT: Since there is a clear evolutionary relationship between the monoterpene, sesquiterpene and diterpene synthase sequences, phylogenetic analysis has typically been focused at the superfamily level. Yet, because 1) the different classes of Terpene Synthases (TSs) bind substrates that differ by at least five carbons and 2) the lack of characterized plant sequences and 3) the rapid evolution of plant species, robust structure-function relationships have not emerged from such bioinformatics analyses. In this presentation, we will describe our efforts to characterize the sesquiterpene synthases (STSs) that bind farnesyl diphosphate and then undergo a range of cyclization mechanisms to produce a variety of products. Our phylogenetic analyses of STSs reveal that sequences that carry out a common cyclization/mechanistic step in the catalytic cycle form distinct sub-groups within the tree. From these classifications, specific residues have been identified that uniquely characterize the sub-groups. We will demonstrate how these residues may impact the mechanism by visualizing where they are located in the 3-dimensional structure.

SYSTEMATIC DRUG TARGET DISCOVERY VIA CHEMICAL AND GENETIC INTERACTION PROFILES

Presenter: Hon Nian Chua, Harvard Medical School
Authors: Hon Nian Chua, Murat Cokol, Yo Suzuki, Frederick P. Roth

ABSTRACT: A variety of genome-scale studies have revealed the interplay between genes and drugs, and pointed to mechanisms of drug action. Haploinsufficiency profiling (HIP) can often reveal drug targets directly, while Homozygous Profiling (HOP) tends to reveal genes that compensate for loss of the drug target's activity. Genetic interaction profiles can be combined with HOP to reveal drug targets (HOP-GI). Here we describe the systematic combination of these approaches to reveal drug targets. We integrated multiple genome-scale chemical-genomic profiles data with genetic interaction profiles covering 75% of all genes in *S. cerevisiae*. Using a collection of gold standard drug/target relationships assembled from the literature, we found that HIP alone can detect 22.8% of known drug targets (48.1% of those examined by HIP). An integrated approach combining HIP, HOP and HOP-GI detects 61.3% of known drug targets examined by either HIP or HOP (70.8% of those examined by both methods). Our results provide new drug target predictions and support the continued application of large-scale studies of genetic and chemical-genomic profiles to systematically reveal mechanisms of drug action.

ON THE ACCURACY OF AUTOMATED INFERENCE OF PROTEIN FUNCTION

Presenter: Wyatt T. Clark, Indiana University

Authors: Wyatt T. Clark, Predrag Radivojac

ABSTRACT: With the gap between known protein sequences vs. the number of sequences experimentally annotated with molecular function reaching two orders of magnitude, the prospects for experimentally annotating all or most of the known proteins in near future are grim. Here, we study the quality of methods for protein automated functional annotation from its amino acid sequence and propose a new algorithm for functional inference. We find that the traditional functional transfer by similarity is surprisingly inaccurate and that even best current methods are in need of improvement. This implies that the consequences of annotating new proteins by electronic annotation may be serious and that the amount of trust in such methods by experimental researchers may be overly high. We then present a new computationally inexpensive and robust method for generating features for query sequences based on sequence alignments and predicted protein properties (e.g. secondary structure, transmembrane helices) and discuss its value against current state-of-the-art. The algorithms were evaluated using gene ontology categories molecular function and biological process. We discuss the value of joint multi-class learning algorithms vs. traditional one-against-all training.

ALGORITHM TO IMPROVE GENE CONSISTENCY ACROSS BACTERIAL GENOMES

Presenter: Judith D. Cohn, Los Alamos National Laboratory

Authors: Judith D. Cohn, Michael E. Wall, John Dunbar

ABSTRACT: Identification of gene boundaries — the first step in genome annotation — provides the foundation for subsequent comparative genomics. Unfortunately, gene-finding algorithms are not always accurate. Even when gene-containing regions are correctly identified, gene prediction algorithms must select a single start site from a multitude of possible start sites. Errors in gene start sites not only alters the encoded peptide sequence but may affect the identification of orthologs. Further, the choice of start site affects the length of the intergenic region, which may impact a suite of other predictions, such as operon structure, regulatory motifs, and comparison of regulatory regions among genomes. In extreme cases, errors may lengthen intergenic regions to the extent that additional genes are predicted in the intervening space. Conversely, errors can remove intergenic content, sometimes resulting in spurious gene overlaps. Recently, we noted extensive inconsistencies in gene start sites among orthologous genes within the *Burkholderia* genus. In particular, we found many orthologous gene sets where predictions appeared consistent for all but one ortholog in a set, suggesting a possible gene-finding error. We posit that inconsistency in gene start sites among orthologs represents a gene-finding error in some cases and real biological variation

in others. In this context, we present an algorithm to improve consistency across multiple genomes and characterize its performance for orthologs across the *Burkholderia* genus.

BIOINFORMATICS CHARACTERIZATION OF THE PLASMODIUM GLUTATHIONE S-TRANSFERASE

Presenter: Emilee Colón, University of Puerto Rico

Authors: Emilee Colón, Adelfa Serrano, Hugh Nicholas Jr, Troy Wymore, Alexander Ropelewski, Ricardo González Méndez

ABSTRACT: Malaria is a global health problem caused by *Plasmodium* parasites. Glutathione S-transferase (GST) is involved in the conjugation of glutathione to drugs and toxic compounds. It is postulated that GST plays an important role in the development of drug resistance. The three-dimensional (3D) structure of *Plasmodium falciparum* GST (PfGST) has been solved. Previous work indicates that the PfGST cannot be assigned to any of the known GST classes. We performed sequence analyses and structural modeling of GSTs from *Plasmodium*, and structural alignments to known structures of the GST from other organisms, in order to classify PfGST into a GST family. Sequence alignments using ClustalW, motif analysis using MEME, and phylogenetic analysis using MEGA4, of PfGST, *Plasmodium vivax* GST (PvGST), *Plasmodium knowlesi* GST (PkGST) and *Plasmodium yoelii* GST (PyGST) and 38 other GST sequences were done. The alignments, motifs and phylogenies show a close relationship to the alpha and sigma class of GSTs. Models of the tertiary structure of the *P. vivax*, *P. knowlesi* and *P. yoelii* GST monomers were obtained using the Protein Homology/analogy Recognition Engine (PHYRE) server. The three-dimensional structures of GST enzymes from various classes (alpha, sigma, mu and pi) were analyzed by structural alignment with the PfGST 3D structure (1Q4J) using the MultiSeq feature in the VMD program. The comprehensive comparison of PfGST with known GST structures reveals high structural similarity that allows PfGST to be classified into unique clade within the sigma class GSTs. These data may open new avenues for the development of novel antimalarials.

KNOWLEDGE NETWORK APPROACH: PATHWAYS AND DRUGS

Presenter: Nikolai Daraselia, Ariadne

Authors: Nikolai Daraselia, Ekaterina Kotelnikova, Anton Yuryev

ABSTRACT: Providing a rich context for experimental data promises to offer new insights into mechanisms of molecular regulation. Employing a resource of millions of findings gleaned from a broad corpus of biomedical literature in which to evaluate genome-wide experimental data can highlight specific molecules otherwise easily missed. The challenge to use this large amount of data in decision-making can be met using appropriate hypothesis testing. Using the proprietary high-content linguistics tool MedScan we compiled a database of knowledge networks associated with different diseases and small molecule effects by extracting

the information from scientific literature. Different approaches towards reconstructing mechanistic models from the resulting knowledgebase and from microarray data will be described. By analyzing disease-specific gene expression data we were able to identify a potentially novel therapeutic target in glioblastoma. Further, by systematically mining the database for knowledge on existing drugs/drug candidates garnered from published findings, a new application for a known agent to inhibit glioblastoma pathway was suggested.

MULTIVARIATE ANALYSIS OF METAGENOMES – AN UNDERGRADUATE REU STORY

Presenter: Elizabeth Dinsdale, San Diego State University

Authors: Naneh Apkarian, Michelle Creek, Eric Guan, Mayra Hernandez, Kate Isaacs, Chris Peterson, Todd Regh, Robert Edwards, Barbara Bailey, Peter Salamon, Imre Tuba, Elizabeth Dinsdale

ABSTRACT: Microbial activity shapes the health of individual organisms, entire ecosystems and the planet. Metagenomes, which are random samples of the microbial genomes within an environment, are constructed to explore variations in microbial activities. Bioinformatics analysis of the metagenome sequences provides a description of the metabolic processes that are important for the growth and survival of the microbes in any given environment. The number of metagenomes is increasing exponentially, making it challenging to analyze and present biological interpretations across all datasets. To address this problem seven Math REU summer students conducted a statistical comparison across 203 metagenomes, a dataset consisting of about 2 billion base pairs (bp) of DNA sequences. Their statistical analyses, which included both supervised and unsupervised techniques, will be described. The former required input from the researcher to obtain groupings, whereas in the later the grouping is provided by the statistical analyses. We demonstrated that the combination of determining group size, using the K-mean silhouette analysis, identifying important variables using the random forest variable plot, and clustering using a canonical discriminant analysis explained 91.2 % of the variance in the microbial metabolic functions across environments with an error rate of 12.5 %. Our results showed that the metabolic profiling provide by metagenomes are highly accurate in distinguishing the activity of microbial communities from different environments and identifying microbial communities within an environment that are perturbed.

A COMPUTATIONAL TOOL FOR THE IDENTIFICATION OF SIGNATURE GENES AMONG PHAGES

Presenter: Bhakti Dwivedi, University of South Florida

Authors: Bhakti Dwivedi, Robert Edwards, Mya Breitbart

ABSTRACT: Phages (viruses that infect bacteria) are the most abundant biological entities on the planet. Phages have been central to many molecular biology tools and discoveries, and serve important ecological functions, including structuring

microbial communities, driving evolution through horizontal gene transfer, and playing major roles in biogeochemical cycling. Unlike other cellular organisms, there is no single gene that is found in all phages. This makes it difficult to understand their evolution and diversity. However, groups of related phage genomes often contain conserved genes ('signature genes') that can be used to classify phages and further explore the diversity of restricted phage groups.

Here we describe the development of an automated bioinformatics analysis tool to identify signature genes conserved across completely sequenced phage genomes. The program includes three successive steps: 1) comparison of all proteins encoded in user-selected phage genomes using a Blastp similarity search, 2) generation of detailed graphical and text-based outputs of genes conserved amongst the phage genomes, and 3) generation of amino acid sequence alignments of the signature genes using ClustalW. Downstream applications of the data also include primer design for PCR amplification of phage genes from environmental samples. Overall, this bioinformatics tool will advance the field of phage biology by identifying signature genes that can be used as genetic markers to study phage biodiversity, phylogeny, and evolution.

REAL TIME METAGENOMICS

Presenter: Robert Edwards, San Diego State University

Authors: Robert Edwards, Robert Olson, Terry Disz, Rick Stevens, Ross Overbeek

ABSTRACT: Metagenomics, extracting DNA from the environment and sequencing en masse, is revolutionizing microbiology. From health related studies through microbial ecology, all studies are being enhanced by our ability to sequence and characterize bacteria without growing them in the laboratory. Advances in next-generation sequencing technology have moved the bottle neck from extract and sequencing the DNA, to bioinformatics analysis of the samples. Typically, processing samples using BLAST will takes days to weeks. We have developed new technology to analyze and annotate metagenomes in minutes rather than weeks. Our new approach has opened the door to unique tools and techniques for data mining, and to providing real-time, in the field, sequence analysis. The improved algorithm, Web services interfaces, and web pages for data analysis are presented.

BIOINFORMATIC ANALYSIS OF SDS-INSOLUBLE PROTEIN AGGREGATES IN S. CEREVISIAE, C. ELEGANS, AND M. MUSCULUS

Presenter: Uday S Evani, Buck Institute

Authors: Uday S. Evani, Theodore W. Peters, Pedro Rodrigues, Gregg Czerwieniec, Sean D Mooney, Gordon Lithgow, Bradford Gibson, Robert Hughes

ABSTRACT: Protein aggregation is known to be linked with both aging and neurodegenerative diseases like Alzheimer's disease, Parkinson's disease, Huntington's disease, and prion diseases. As there is a clear association between

protein aggregation and late onset disease states, we wanted to analyze the SDS-insoluble fraction of the proteome in *S. cerevisiae*, *C. elegans*, and cultured mouse cells. We compared the SDS-insoluble fraction in young and aged samples in the three models and found that proteins accumulate in an age dependent manner. Proteins found in the insoluble fraction of the aged cells were essentially absent in the young cells. We then compared proteins from the insoluble fraction in the three models and found significant overlap. This suggests that there are distinct classes of proteins conserved across these species which are particularly susceptible to aggregation. It is also becoming increasingly clear that protein aggregation is a complex process involving cascade of molecular reactions initiated by a subset of the insoluble proteins and this analysis will help elucidate this. Our analyses include looking for enrichment of specific molecular functions, biological processes, protein structural features, ubiquitination, aggregation propensities, protein half-life, and expression data associated with proteins in the insoluble fraction. In this presentation we will discuss results from these analyses.

PAIRWISE AND HIGHER-ORDER CORRELATIONS AMONG DRUG-RESISTANCE MUTATIONS IN HIV-1 SUBTYPE B PROTEASE

Presenter: Omar Haq, Rutgers University

Authors: Omar Haq, Ronald M Levy, Alexandre V Morozov, Michael Andreac

ABSTRACT: The reaction of HIV protease to inhibitor therapy is characterized by the emergence of complex mutational patterns which confer drug resistance. Here we develop a probabilistic approach based on connected information that allows us to study residue, pair level and higher-order correlations within the same framework. We apply our methodology to a database of approximately 13,000 sequences which have been annotated by the treatment history of the patients. We show that including pair interactions is essential for agreement with the mutational data, since neglect of these interactions results in order-of-magnitude errors in the probabilities of the simultaneous occurrence of many mutations. The magnitude of these pair correlations changes dramatically between sequences obtained from patients that were or were not exposed to drugs. Higher-order effects make a contribution of as much as 10% for residues taken three at a time, but increase to more than twice that for 10 to 15-residue groups. We find that higher-order interactions have a significant effect on the predicted frequencies of sequences with large numbers of mutations. While relatively rare, such sequences are more prevalent after multi-drug therapy. The relative importance of these higher-order interactions increases with the number of drugs the patient had been exposed to. Correlations are critical for the understanding of mutation patterns in HIV protease. Pair interactions have substantial qualitative effects, while higher-order interactions are individually smaller but may have a collective effect. Together they lead to correlations which could have an important impact on the dynamics of the evolution of cross-resistance.

A USEFUL TOOL FOR CALCULATING BINDING-SITE RESIDUES ON PROTEINS FROM PDB STRUCTURES

Presenter: Jing Hu, Franklin & Marshall College

Authors: Jing Hu, Changhui Yan

ABSTRACT: Proteins perform various functions through interactions with other molecules, such as DNA, RNA, proteins, carbohydrates, and ligands. To study the mechanisms of these interactions, researchers often need to identify binding-site residues on proteins. A commonly adopted strategy is to find a complex structure from the Protein Data Bank (PDB) that consists of the protein of interest and its interacting partner(s) and calculate binding-site residues based on the complex structure. However, one serious flaw with this approach is that a protein may participate in multiple interactions, but the binding-site residues calculated based on one complex structure usually do not reveal all binding sites on a protein. Thus, it is necessary to find all PDB complexes that contain the protein of interest and combine all the binding-site information on the protein collected from them. This process is very painful and time-consuming when there is a large set of proteins to analyze. Especially, combining the binding-site information obtained from different PDB structures requires tedious work of aligning protein sequences. We have developed a tool for calculating binding-site residues on proteins (TCBRP). For an input protein, TCBRP can quickly find all binding-site residues by automatically integrating the binding-site information from all PDB structures that contain the protein of interest. Additionally, TCBRP presents the binding-site residues in different categories based on the molecule types they interact with, e.g., DNA, RNA, protein, carbohydrates, and ligands. TCBRP also allows users to choose the definition of binding-site residues. The tool is available at <http://yanbioinformatics.cs.usu.edu:8080/ppbindingsubmit>.

A FAST APPROACH TO PROTEIN STRUCTURE ALIGNMENT BASED ON ONE-DIMENSIONAL ALPHABET CODE SEQUENCES

Presenter: Kenneth Hung, National Taiwan University

Authors: Kenneth Hung, Jui-Chih Wang, Kun-Nan Tsai, Cheng-Wei Chen, Chung-Ming Chen

ABSTRACT: Protein structure alignment is one of the major steps in understanding the interplay of protein structure and evolution. It is also an essential process for finding homologous proteins based on identification of structural similarity and may further help the elucidation of protein function. To achieve a reliable and fast structural alignment, a new protein structure alignment approach was proposed based on the coded one-dimensional alphabet code sequences. The sequences were obtained from decomposing recurrent protein chains into four consecutive residues in an overlapping way. The initial alignment was carried out by aligning secondary structure elements in two proteins using combination of orientation independent and dependent score functions to account for global structural similarity. The protein chains were then encoded into one-dimensional alphabet code sequences

and dynamic programming was employed to achieve local structural alignment. Finally, superposition of two structures was applied to refine alignment results by calculating inter-molecular distance. The test data used in the study were protein structures, downloaded from SCOP with less than 40 % sequence identity (613 protein pairs randomly chosen from 161 families). We compared the proposed method with Combinatorial Extension (CE) and Secondary Structure Matching (SSM), which are two of the most utilized structural comparison tools. The computational time was reduced to 6m36s by the proposed method, in contrast to 37m38s of CE and 13m37s of SSM. The Match Index (MI) was employed as quality indicator to demonstrate that the coded one-dimensional sequence method achieved alignment quality comparable to SSM.

DEALING WITH DATA DELUGE: DESIGNING AND IMPLEMENTING A DATABASE TO ENABLE SPECIALIZED STUDIES IN METAGENOMICS

Presenter: Bonnie Hurwitz, University of Arizona

Authors: Bonnie Hurwitz, Matthew Sullivan, Robert Edwards, Adam Monier, Alexandra Z. Worden, Sudha Ram

ABSTRACT: Metagenomics has transformed health and environmental sciences through examination of DNA from wild microbes that previously evaded identification by culturing or microscopy. These complex datasets are used to reveal patterns in sequence data correlated to environmental metadata in hopes of explaining biological phenomena. Recently established metagenomics data standards (e.g., MIGS/MIMS) maximize their comparative utility, and new metagenomic databases complement sequence repositories like Genbank with environmental data. While these databases offer researchers pre-computed analyses and well-defined GUIs to analyze preexisting datasets, mechanisms for managing unpublished metagenomic datasets and seamlessly combining data from multiple sources remain underdeveloped. We introduce a design for a MIGS/MIMS compliant MySQL relational database system that equips researchers with the ability to describe metadata associated with a habitat, disease or sample treatment, in addition to documenting how raw data were processed, filtered and analyzed. We follow the well-known database design life cycle that includes conceptual design using entity relationship modeling, relational design and normalization, and completeness checks to assure common data objects are included. To ease de novo annotation efforts, we develop scripts to interface between the database and automated annotation systems such as RAST (rast.nmpdr.org). We demonstrate the capabilities of this database system by examining ocean viruses in an unpublished dataset of ~5 million sequence reads originating from twelve discrete ocean metagenomes sampled along a coastal to open ocean transect off Monterey Bay, California. This first step offers powerful new tools to translate metagenomic sequence datasets to interpret the biology of human, aquatic and earth systems.

VIRAL METAGENOMICS IN MARINE MICROBIAL SAMPLES

Presenter: Julio C. Ignacio, University of Arizona

Authors: Julio C. Ignacio, Alexandra Z. Worden, Matthew B. Sullivan

ABSTRACT: Marine viruses numerically dominate the oceans, and impact microbial communities through mortality, gene transfer, and modulating microbial metabolisms. Here we analyze viral sequences from size-fractionated marine microbial metagenomes originating from four sites: coastal surface, mesotrophic surface, open-ocean surface, and open-ocean deep-chlorophyll-maximum (DCM) waters. Overall, environmental genomic 'viral' fragments ranged from ~2% of the total sequence reads in the largest-size-fraction (3-20µm) from coastal and upwelling water to ~15% in the smallest-size-fraction (0.1-0.8µm) open-ocean DCM sample. The T4-like viruses, prominent in this dataset, significantly varied among sites with the lowest (~34%) signal from coastal and upwelling waters, and the biggest from open ocean (up to 85% at the DCM). Within the T4-like viruses, we wondered what fraction were cyanophages, and approached the problem using a dataset of 26 genomes and their hierarchical core gene sets (e.g., T4-like vs cyano-T4-like core genes) to estimate their abundance, relative sequence diversity and sequence characteristics in the metagenomes. These analyses suggest that cyanophages represent a larger fraction of the T4-like viruses in open ocean (>75%) than coastal or upwelling regions (<50%), consistent with abundance estimates of cyanobacterial host genes. Notably, however, even with >300K sequence reads per sample, single-copy widespread microbial and viral 'core' genes do not occur in the expected 1:1 ratio, suggesting undersampling in these complex communities. Finally, short-comings of our current analyses suggest the need for new reference genomes from a diversity of water types, as well as defining within- and between-population genomic variability.

USING SYNTACTIC CONTEXT IN OPENDMAP PATTERNS

Presenter: Helen Johnson, University of Colorado, Denver

Authors: Helen L. Johnson, William Baumgartner Jr, Christophe Roeder, Karin Verspoor, Kevin Bretonnel Cohen, Larry Hunter

ABSTRACT: Curated data recorded in biological databases is a critical resource for biological data analysis. This data, however, is vastly incomplete. The biomedical literature contains a lot of information that is not represented in databases. The OpenDMAP concept recognition system uses patterns to extract events from biomedical text. Patterns applied using OpenDMAP for extraction of various biologic interaction types, such as protein-protein interaction, phosphorylation, localization, etc., so far have largely relied on matching a continuous sequence of pattern elements including text literals, semantically typed categories, and shallow syntactic categories. Historically, the precision of OpenDMAP output has been high, but the recall has been low. However, OpenDMAP's pattern language allows

for the use of dependency parse information, the use of which may improve recall. We report on manual pattern creation practices and preliminary experiments using patterns that rely on syntactic dependency relationships to extract events expressed in complex sentential structures in biomedical text.

IN SILICO FUNCTIONAL PROFILING OF HUMAN DISEASE-ASSOCIATED AND POLYMORPHIC AMINO ACID SUBSTITUTIONS

Presenter: Vidhya G. Krishnan, Buck Institute for Age Research

Authors: Matthew Mort, Uday S. Evani, Vidhya G. Krishnan, Kishore K. Kamati Peter H. Baenziger, Anghuman Bagchi, Brandon Peters, Rakesh Sathyesh, Biao Li, Yanan Sun, Bin Xue, Nigam Shah, Maricel Kann, David N. Cooper, Predrag Radivojac, Sean D. Mooney

ABSTRACT: We have co-opted a range of bioinformatic tools, designed to predict structural and functional sites in protein sequences, to the task of ascertaining whether intrinsic biases exist in terms of the distribution of different types of human amino acid substitutions (AAS) with respect to their structural, functional and pathological features. We applied these tools to compiled datasets of human disease-associated AAS in the contexts of inherited monogenic disease, complex disease, functional polymorphisms with no known disease association, somatic mutations in cancer, and neutral polymorphic AAS. The analysis revealed marked similarities in terms of the distribution of structural and functional sites between monogenic disease mutations and functional polymorphisms, with a bias toward those variants that impact protein function via structural disruption ($P=3.8X10^{-24}$). Putative causative variants in both complex disease and cancer were significantly over-represented in intrinsically disordered regions ($P=8.83X10^{-56}$) whilst cancer-associated mutations were enriched at certain molecular recognition sites ($P=1.6X10^{-3}$). We postulate that missense mutations in complex disease and cancer are more likely than monogenic disease to impact on protein function directly through disruption of functional sites (e.g. protein interaction) rather than indirectly via structural disruption. Further analysis of subtypes of inherited disease (e.g. cardiovascular disease) served to identify several disease entities that differed significantly in terms of the distribution of specific causative molecular changes. For example, blood coagulation disorders were found to exhibit a 19-fold depletion in AAS at O-linked glycosylation sites. In overall terms, however, the disruption of a specific molecular function does not constitute a disease-specific phenomenon.

A USER STUDY OF ATTRIBUTE VISUALIZATION TOOLS AND THEIR ROLE IN UNDERSTANDING BIOLOGICAL NETWORKS

Presenter: Hande Kucuk, Eastern Michigan University

Authors: Hande Kucuk, Benjamin J. Keller, Terry Weymouth, Barbara Mirel

ABSTRACT: Peptide identification conducted by various peptide search algorithms is a primary step in shotgun proteomics data analysis that serves as the basis for further analyses such as protein identification and quantification. Given a set of fragmentation spectra from a shotgun proteomics experiment, peptide

identification can be reduced to the problem of discriminating the true peptide-spectrum matches (PSMs) from the false ones. Although the commonly used peptide search algorithms are quite efficient, it is shown that their discriminative power for true versus false PSMs is limited. Hence, post-processing methods like the percolator algorithm have been developed which re-rank the PSMs and significantly improve the performance of peptide identification. The percolator algorithm utilized the PSMs from a decoy protein database as the negative training set in the learning process, which may result in an over-optimistic estimation of false discovery rate (FDRs), also based on the search in the decoy database. To address this issue, we propose two learning approaches to assigning probabilistic scores to PSMs without using a decoy database, one based on a semi-supervised learning algorithm and the other based on a regression model. Further, contrasting all existing algorithm, our algorithm does not assume any uniform distribution for the scoring, but instead performs peptide and spectra specific evaluation on each PSM. The testing of our methods on a few shotgun proteomics experiments showed that they reported reliable estimation of probability, while achieved higher discrimination power than the percolator algorithm.

AUTOMATED INFERENCE OF MOLECULAR MECHANISMS OF DISEASE FROM AMINO ACID SUBSTITUTIONS

Presenter: Biao Li, Indiana University

Authors: Biao Li, Vidhya G. Krishnan, Matthew E. Mort, Fuxiao Xin, Kishore K. Kamati, David N. Cooper, Sean D. Mooney, Predrag Radivojac

ABSTRACT: Single nucleotide substitutions within protein coding regions are of particular importance owing to their potential to give rise to amino acid substitutions that affect protein structure and function which may ultimately lead to a disease state. Over the last decade, a number of computational methods have been developed to predict whether such amino acid substitutions result in an altered phenotype. Although these methods are useful in practice, and accurate for their intended purpose, they are not well suited to providing probabilistic estimates of the underlying disease mechanism. We have developed a new computational model, MutPred, that is based upon protein sequence, and which models changes of structural features and functional sites between wild-type and mutant sequences. These changes, expressed as probabilities of gain or loss of structure and function, can provide insight into the specific molecular mechanism responsible for the disease state. MutPred also builds on the established SIFT method but offers improved classification accuracy with respect to human disease mutations. Given conservative thresholds on the predicted disruption of molecular function, we propose that MutPred can generate accurate and reliable hypotheses on the molecular basis of disease for ~11% of known inherited disease-causing mutations. We also note that the proportion of changes of functionally relevant residues in the sets of cancer-associated somatic mutations is higher than for the inherited lesions

in the Human Gene Mutation Database which are instead predicted to be characterized by disruptions of protein structure.

LOSS OF POST-TRANSLATIONAL MODIFICATION SITES IN DISEASE

Presenter: Shuyan Li, Indiana University, Bloomington

Authors: Shuyan Li, Lilia M. Iakoucheva, Sean D. Mooney, Predrag Radivojac

ABSTRACT: Understanding and predicting molecular cause of disease is one of the major challenges for biology and medicine. One particular area of interest continues to be computational analyses of disease-associated amino acid substitutions. To this end, various studies have been performed to identify molecular functions disrupted by disease-causing mutations. Here, we investigate the influence of disease-associated mutations on post-translational modifications. In particular, we study the loss of modification target sites as a consequence of disease mutation. We find that about 5% of disease-associated mutations may affect known modification sites, either partially (4%) or fully (1%), compared to about 2% of putatively neutral polymorphisms. Most of the fifteen post-translational modification types analyzed were found to be disrupted at levels higher than expected by chance. Molecular functions and physiochemical properties at sites of disease mutation were also compared to those of neutral polymorphisms involved in the process of post-translational modification site disruption. Disease-associated mutations in the neighborhood of post-translationally modified sites were found to be enriched in mutations that change polarity, charge, and hydrophobicity of the wild-type amino acids. Overall, these results further suggest that disruption of modification sites is an important but not the major cause of human genetic disease.

MACHINE LEARNING APPROACHES TO THE ASSESSMENT OF PEPTIDE-SPECTRUM MATCHES WITHOUT USING A DECOY DATABASE

Presenter: Yong Fuga Li, Indiana University, Bloomington

Authors: Yong Fuga Li, Randy J. Arnold, Predrag Radivojac, Haixu Tang

Peptide identification conducted by various peptide search algorithms is a primary step in shotgun proteomics data analysis that serves as the basis for further analyses such as protein identification and quantification. Given a set of fragmentation spectra from a shotgun proteomics experiment, peptide identification can be reduced to the problem of discriminating the true peptide-spectrum matches (PSMs) from the false ones. Although the commonly used peptide search algorithms are quite efficient, it is shown that their discriminative power for true versus false PSMs is limited. Hence, post-processing methods like the percolator algorithm have been developed which re-rank the PSMs and significantly improve the performance of peptide identification. The percolator algorithm utilized the PSMs from a decoy protein database as the negative training set in the learning process, which may result in an over-optimistic estimation of false discovery rate

(FDRs), also based on the search in the decoy database. To address this issue, we propose two learning approaches to assigning probabilistic scores to PSMs without using a decoy database, one based on a semi-supervised learning algorithm and the other based on a regression model. Further, contrasting all existing algorithm, our algorithm does not assume any uniform distribution for the scoring, but instead performs peptide and spectra specific evaluation on each PSM. The testing of our methods on a few shotgun proteomics experiments showed that they reported reliable estimation of probability, while achieved higher discrimination power than the percolator algorithm.

STRUCTURE DISCOVERY IN PPI NETWORKS USING PATTERN-BASED NETWORK DECOMPOSITION

Presenter: Ying Liu, University of Texas, Dallas

Authors: Ying Liu, Chengcheng Shen, Phil Bachman

ABSTRACT: The large, complex networks of interactions between proteins provide a lens through which one can examine the structure and function of biological systems. Previous analyses of these continually growing networks have primarily followed either of two approaches: large-scale statistical analysis of holistic network properties, or small-scale analysis of local topological features. Meanwhile, investigation of meso-scale network structure (above that of individual functional modules, while maintaining the significance of individual proteins) has been hindered by the computational complexity of structural search in networks. Examining protein-protein interaction (PPI) networks at the mesoscale may provide insights into the presence and form of relationships between individual protein complexes and functional modules. Results: In this article, we present an efficient algorithm for performing sub-graph isomorphism queries on a network and show its computational advantage over previous methods. We also present a novel application of this form of topological search which permits analysis of a network's structure at a scale between that of individual functional modules and that of network-wide properties. This analysis provides support for the presence of hierarchical modularity in the PPI network of *Saccharomyces cerevisiae*.

PROMOTER PREDICTION IN HALOTHIOBACILLUS NEAPOLITANUS C2 BASED ON STRESS-INDUCED DNA DUPLEX DESTABILIZATION

Presenter: Aleksandra Markovets, Mississippi Valley State University

Authors: Aleksandra Markovets, Charles Bland, Abigail Newsome

ABSTRACT: In the post-genomic era when scientists can sequence the genomes of many organisms, one of the biggest challenges is the correct identification of promoter regions, which is essential for the understanding of gene regulation. Since wet-lab promoter prediction techniques are time consuming, in silico methods have been used to facilitate the process. Most of these traditional computational methods are based on motifs searching, which are insufficiently conserved to

predict at a high level. To compensate for this shortcoming, DNA structural properties, such as curvature, stacking energy and stress-induced DNA duplex destabilization (SIDD), have been used. Of particular concern for this study is the prediction of promoters in the proteobacteria *Halothiobacillus neapolitanus* c2. We have implemented a method that predicts promoter sequences in this organism by using the DNA SIDD of the genome. Recent studies have shown that SIDD is a distinctive structural attribute of promoter regions. The SIDD predicted promoter-containing sites of this research can be used as targets for experimental verification or further bioinformatics investigation.

TOWARDS REALISTIC CODON MODELS: AMONG-SITE VARIABILITY AND DEPENDENCY OF SYNONYMOUS AND NON-SYNONYMOUS SUBSTITUTION RATES

Presenter: Itay Mayrose, University of British Columbia

Authors: Itay Mayrose, Adi Doron-Faigenboim, Eran Bacharach, Tal Pupko

ABSTRACT: The evolutionary selection forces acting on a protein-coding gene are commonly inferred using codon evolutionary models by computing the rate of non-synonymous (amino-acid altering; K_a) to synonymous (silent; K_s) substitution rate. Current codon models usually assume that K_a can vary between sites of the gene due to selection forces that operate at the protein level, but that K_s is constant and represents the neutral substitution rate. Towards a more realistic description of sequence evolution, a model that accounts for among site-variation of both synonymous and non-synonymous substitution rates was developed. This model allows the inference of selection forces that operate at the DNA and mRNA levels and further increases the accuracy of K_a/K_s estimation.

THE HUMAN GENE MUTATION DATABASE (HGMD) AND ITS EXPLOITATION IN THE ERA OF PERSONALIZED GENOMICS

Presenter: Matthew Mort, Cardiff University

Authors: P.D. Stenson, M. Mort, E. Ball, K. Howells, A. Phillips, N.S.T. Thomas, D.N. Cooper

ABSTRACT: The Human Gene Mutation Database (HGMD) is a comprehensive core collection of germline mutations in nuclear genes that underlie or are associated with human inherited disease. HGMD includes rare variants causative of monogenic disease, putatively functional polymorphisms associated with complex disease and polymorphism of functional significance with as yet no reported disease-association. By October 2009, the database contained over 93,000 different lesions detected in 3,453 different genes, with new entries currently accumulating at a rate exceeding 9,000 per annum. HGMD has a broad utility for researchers, physicians, clinicians and genetic counselors as well as for companies specializing in biopharmaceuticals, bioinformatics and personalized genomics. Its application to the post-genomic era is therefore diverse but of increasing importance is its application and exploitation in the interpretation and analysis of personal genomics data.

META-ANALYTICAL TOOLS FOR DECIPHERING TRANSCRIPTIONAL NETWORKS IN A MODEL ANOXYGENIC PHOTOTROPH

Presenter: Oleg Moskvina, University of Wyoming

Authors: Oleg Moskvina, Dmitry Bolotin, Pavel Ivanov, Mark Gomelsky

ABSTRACT: We designed a web-based meta-analytical tool to help decipher transcriptional regulatory networks in the metabolically versatile anoxygenic phototrophic bacterium *Rhodobacter sphaeroides*. Expression data from 100 GeneChips corresponding to various growth conditions and regulatory mutants were processed using Robust Multi-Array Analysis. For every pair of genes, Pearson correlation coefficients for transcriptional patterns were calculated. Gene-centric networks were visualized using the approach introduced in StarNet (PLoS ONE. 2008 3(3):e1717) as a basis for further development. To simplify transcriptional network interpretation, we added the option to display genes in the context of predicted operons and superoperons. Regulatory regions (-400 to +20 nt) of genes belonging to each deduced regulatory network were analyzed for putative binding sites of bacterial transcription factors (TFs). 188 position-specific weight matrices from the ProDoric database were used. For every gene-TF combination (797,496 total), the match scores were calculated for a range of Prior Probability values. We designed an expert system, which selects statistically significant TF hits in a genelist using the 'Score vs. Prior Probability' 2-dimensional matrix. The selection is based on overrepresentation of the binding site hits (genes in network versus whole genome) estimated from binomial distribution. Evaluation of the performance of this algorithm using lists of genes responded to regulatory mutations revealed relevant selection of putative TFs. The database of global transcriptional profile correlations, tools for network visualization and optimized search for TF binding sites are accessible via web interface at <http://rhodobase.org>. The core of the described meta-analytical system can be expanded to include other bacteria.

THE EVOLUTIONARY LANDSCAPE OF CHROMATIN MODIFICATION MACHINERY

Presenter: Tuan On, University of Toronto

Authors: Tuan On, Xuejian Xiong, Shuye Pu, Andrei Turinsky, Yunchen Gong, Andrew Emili, Zhaolei Zhang, Jack Greenblatt, Shoshana J. Wodak, John Parkinson

ABSTRACT: Chromatin modification (CM) comprises an array of broadly conserved biological processes that modify chromatin to control access to DNA. Due to its fundamental role in processes such as transcription, DNA replication and repair, many components of CM are thought to play an important role in various forms of cancer and cancer related developmental processes. While CM machinery has been extensively characterized in yeast, less is known about its operation in other eukaryotes. Here we exploit over 100 eukaryotic genome sequences to systematically profile the conservation and evolution of CM. Comprehensive surveys of literature and database resources revealed 312 CM

components in yeast, 75 in worm, 118 in fly and 171 in human. Applying the InParanoid algorithm, we are able to significantly expand the numbers of known CM components on the basis of orthology relationships. These include an additional 309 human genes that merit experimental investigation. Surprisingly, while we identify many highly conserved 'core' components, many components are found to be restricted to specific lineages. In addition, we note the presence of many lineage specific gene family expansions, particularly among the vertebrates, together with an apparent loss and/or significant divergence of many components in parasitic organisms. Phylogenetic tree construction of two CM complexes reveal that CM complexes are comprised of a mosaic of components with varying evolutionary histories. Together these findings provide insights into the evolution of CM and will help facilitate an improved annotation of CM across eukaryotes.

BACTERIAL EVOLUTION: IMPLICATION FROM LIPID ABIOSYNTHESIS PATHWAY ENZYMES

Presenter: Stephen O. Opiyo, University of Nebraska, Lincoln

Authors: Stephen Opiyo, Rosevelt Pardy, Hideaki Moriyama, Etsuko Moriyama

ABSTRACT: Lipid-A, a complex glycolipid, is the highly immunoreactive endotoxic center of lipopolysaccharide (LPS). It anchors the LPS into the outer membrane of most gram-negative bacteria. Lipid-A can be recognized by animal cells, triggers some defense-related responses, and causes gram-negative sepsis. The lipid-A biosynthesis pathway consists of nine enzymatic steps. Using BLASTP, TBLASTN, and a more sensitive profile hidden Markov model, we searched for lipid-A enzymes across 62 bacteria genomes. We found that not all gram-negative bacteria have the nine canonical lipid-A biosynthesis enzymes found in *Escherichia coli* K12, some bacteria have no lipid-A enzymes and others have only LpxA, LpxC, LpxD, and LpxB. lpxH gene appeared to have arisen from a duplication of lpxH2 gene after beta/gammaproteobacteria were diverged from other proteobacteria, lpxM gene is also a duplicate of lpxL gene found only in the gammaproteobacteria. This study clearly showed that the currently known nine-enzyme pathway for lipid-A biosynthesis, which has been mainly studied in *E. coli* K12 and related bacteria, should not be considered as representative nor ancestral to all bacteria.

CNGEN – A NEW TOOL FOR COPY-NUMBER GENOTYPES PARTITIONING

Presenter: Louis-Philippe Lemieux Perreault, Montreal University

Authors: Louis-Philippe Lemieux Perreault, Gregor Andelfinger, Géraldine Asselin, Marie-Pierre Dubé

ABSTRACT: Copy number polymorphisms (CNPs) and variations (CNVs) may be at least as important as single nucleotide polymorphisms (SNPs) in assessing human genetic variability, since conservative estimates have shown that they might affect more than 10% of the human genome. Integrated genotypes, derived by the genotyping of SNPs, CNPs and CNVs using the Birdsuite software on Affymetrix's

Genome-Wide Human SNP array, are now being used for association studies of complex traits. The use of those genotypes in linkage analysis with multi-generational family data is limited by the requirement of chromosome-specific copy number assignment, or partitioning. We have developed new software which, once applied to familial trios or extended pedigrees, produces partitioned copy number genotypes with distinct parental alleles. Those multi-allelic partitioned copy number polymorphisms have the potential to offer a new and powerful tool for linkage analysis. CNGen has been validated using simulations on complex pedigree structures. The simulation steps have shown that CNGen will not result in an excess of false calls in the presence of genotyping error or de novo mutations, supporting its robustness. The new method has been applied successfully to a real dataset of 300 genotyped samples from 42 pedigrees segregating congenital left ventricular outflow tract obstruction. CNGen partitioned 55% of Birdsuite's results and identified 3,500 (0.44%) Mendelian errors in the process, a rate within expectations for multi-allelic markers. CNGen is a flexible, open source and platform independent Python program.

QIIME (QUANTITATIVE INSIGHTS INTO MICROBIAL ECOLOGY) DYNAMIC RAREFACTION GRAPH

Presenter: Megan Pirrung, University of Colorado, Boulder

Authors: Megan Pirrung, Rob Knight

ABSTRACT: Current high-throughput sequencing technology can return half a million sequences in a single run, and this number is increasing dramatically every few months as technology improves. These datasets are both large and highly multivariate, complicating data analysis. Knight Lab of Boulder Colorado has developed a pipeline for analyzing such data, called QIIME, Quantitative Insights Into Microbial Ecology (manuscript in preparation). The QIIME pipeline includes several new visualizations that assist in analyzing data, and can be run easily on a user's laptop. One such visualization set is the Dynamic Rarefaction Graph, which utilizes javascript web technologies to dynamically graph rarefaction data produced by the QIIME pipeline. Many different javascript graphical libraries were tested to determine the most effective package for producing dynamic graphs. The current software takes user supplied parameters such as a mapping file containing relevant metadata, and rarefaction data, and then produces a webpage containing a graph that the user may interact with. Since the software is dynamic, series averages and error bars are calculated in real time instead of statically, as was the case in previous software packages. To facilitate analysis of the data, the user can also dynamically recolor the graphs according to user-defined metadata. The web-based functionality of the graphs means the software is platform agnostic. These rarefaction graphs are an integral part of the QIIME pipeline and high-throughput sequence data analysis, leading to discoveries about microbial ecology and its intricate role in different diseases.

STRUCTURE-BASED PREDICTION OF DNA BINDING SITES FOR FAMILIES OF TRANSCRIPTION FACTORS

Presenter and Author: Julia Ponomarenko, University of California, San Diego

ABSTRACT: Gene transcription is regulated through binding of transcription factors (TF) to specific sites on DNA, known as TF binding sites (TFBS). Most algorithms for predicting TFBSs in genome sequences apply position-specific weight matrices (PWM) obtained from aligned TFBS sequences. That requires a significant amount of experimentally identified TFBSs. Alternative approaches include building PWMs either for families of TFs sharing similar DNA-binding domains or using 3D structures of TF-DNA complexes. Here a novel method for the prediction of TFBSs binding a family of TFs is presented. The method takes into account both 3D structures of TF-DNA complexes and sequences of TFBSs. Structures of DNA-binding protein domains are aligned, using the algorithm of structural alignment that favors matches between residues interacting with DNA. As a result the alignment of DNA bound to the proteins emerges. Thus aligned DNA sequences are used as a seed for aligning all TFBSs. A PWM is further defined, assuming that the probability of a base pair to be at a certain position of the site depends on its occurrence in the sequence alignment and the number of contacts with TF in the structural alignment. Applied to the sites binding the NFkB family factors, the method has been shown to discriminate TFBS sites from non-sites significantly better than other tested sequence- and structure-based methods. Notwithstanding, the correlations between experimentally measured TF-DNA binding affinity values and binding scores predicted by the proposed method were comparable to those calculated using other methods. This work is funded by NIH grant R01GM085325.

AUTOMATED IDENTIFICATION OF AMPLIFIABLE MICROSATELLITE LOCI AND PRIMER DESIGN FROM 454 HIGH-THROUGHPUT SEQUENCING READS

Presenter: Alex Poole, University of Colorado, Denver

Authors: Alex Poole, Todd Castoe, David Pollock

ABSTRACT: Optimal integration of next-generation sequencing into mainstream research requires re-evaluation of how problems can be reasonably overcome and what questions can be asked. One potential application is the rapid acquisition of genomic information to identify microsatellite loci for evolutionary, population genetic and chromosome linkage mapping research on non-model and not previously sequenced organisms. Here, we report on results using high-throughput sequencing to obtain a large number of microsatellite loci from a venomous snake. Our approach to identify microsatellite loci and corresponding primer pairs was rapid, user-friendly, and provides primer analysis tailored for unsequenced genomes

ANALYSIS OF A LOCAL HUNTINGTIN PROTEIN INTERACTION NETWORK

Presenter: Corey Powell, Buck Institute for Age Research

Authors: Corey Powell, Robert Hughes, Cendrine Tourette, Russell Bell, Sean Mooney

ABSTRACT: Huntington's Disease is a neurodegenerative disorder caused by an abnormally long stretch of glutamines in the associated huntingtin protein. This study sheds light on possible functions for the huntingtin protein through analysis of a local protein-protein interaction network consisting of the huntingtin protein, proteins called primaries that have been found to interact with the huntingtin protein and secondary proteins that interact with the primary proteins. The first part of the analysis finds annotations that are overrepresented among the primary and secondary proteins. The second part of the analysis examines the network structure and finds functions and proteins that are more highly connected in the network than expected by chance. The third part of the analysis uses additional information, such as gene coexpression, to corroborate the results from the first two analyses.

FLUX BALANCE ANALYSIS OF PLASMODIUM FALCIPARUM'S METABOLIC NETWORK

Presenter: Farhan Raja, University of Toronto

Authors: Farhan Raja, John Parkinson, James Wasmuth, Stacy Hung

ABSTRACT: *Plasmodium falciparum* is the causative agent of human malaria, responsible for upwards of three million deaths each year. Due to the recent emergence of drug-resistance, there is an urgent need for the development of novel anti-malarial therapeutics. As many drugs function by inhibiting metabolic pathways, a thorough understanding of *P. falciparum* metabolism is key for effective drug discovery. Flux balance analysis (FBA) of metabolic networks has been shown, in many organisms, to be a useful tool in the search and analysis of metabolic drug targets. We have curated a metabolic model of *P. falciparum*: 317 genes, 547 reactions, 490 metabolites, and five subcellular localizations. Taking production of cellular biomass as the objective, our FBA simulations identified optimal reaction flux distributions in response to a variety of metabolic network perturbations. Through the systematic inhibition of nutrient transporters, we predict that ten amino acids cannot be synthesized by the parasite and must be obtained from its host. Additionally, we have looked at other nutrients whose essentiality is debated. For example, we show that restricting p-aminobenzoic acid (PABA) disrupts parasite growth rate, but is not lethal. Importantly, drug inhibition simulations suggest that many of the currently suggested drug targets function via toxicity effects rather than direct inhibition of biomass production. Further simulations are aimed towards classifying global effects in response to drug

inhibition and searching for potentially lethal drug combinations. To strengthen the model we are introducing expression and proteomic datasets, thus uncovering the metabolic subtleties of this complex and important parasite.

GENEBOOK, A HIGH PRECISION MAMMALIAN PROTEIN THESAURUS

Presenter: Phoebe Roberts, Pfizer, Inc.

Authors: Robert Hernandez, Markella Skempri, Phoebe Roberts

ABSTRACT: Pfizer has an extensive literature retrieval system for populating internal databases that support target identification and characterization. Central to this system is normalization of gene mentions in the literature, enabling integration of extracted facts from the literature with other internal and external data sources. GeneBook is Pfizer's manually curated gene/protein dictionary with 265,874 terms (173,623 of which are synonyms) culled from public gene name sources and updated weekly using automated and manual approaches. In anticipation of augmenting our retrieval process with disambiguation tools, we recently measured GeneBook's baseline performance using dictionary-lookup methods. Benchmarking against the GENIA corpus revealed that GeneBook is a highly precise protein thesaurus, with precision of 98% and recall ~60%. To approximate gene normalization performance against all of Medline, dictionary lookup via Oracle Text using GeneBook was compared to an external gene normalization solution with a published f-measure of ~82% against a semi-random set of 80,000 Medline records enriched for publications from the last 15 years. The gene mentions retrieved by the two methods, recorded as the gene identifier (GeneID):Medline record identifier (PMID) pair, were pooled. Analysis of 100 unique GeneID:PMID pairs found by both methods showed 93% precision. Similar analysis of results retrieved only by GeneBook showed 50% precision, suggesting an overall estimated precision of 71%. Together, these data show that GeneBook is a highly precise gene thesaurus, and, as noted in related work, that test corpora like GENIA overestimate precision rates relative to all of Medline.

INTEGRATING WEB SERVICES INTO BIOMEDICAL TEXT MINING

Presenter: Christophe Roeder, University of Colorado, Denver

Authors: Christophe Roeder, William Baumgartner Jr, Larry Hunter

ABSTRACT: Ease of installation, platform independence and interoperability have been key factors in the adoption of some of the most widely used biomedical text mining tools. Coinciding with the maturation of the Semantic Web has been the increasing availability of biomedical text mining web services, which inherently need no installation and are platform-independent. In terms of interoperability, the Apache Unstructured Information Management Architecture (UIMA) standard has taken integration to new heights. We have adapted a number of text mining tools into UIMA including the National Center for Biomedical Ontology's (NCBO) Open Biomedical Annotator (OBA). The OBA service recognizes biomedical

ontology concepts in text and returns annotations indicating the location of all matched concepts. Leveraging the integration efforts of the entire NCBO, this service hides from the user the installation, mapping, and maintenance efforts for >100 ontologies, a process that would be daunting for single users. In this talk we will describe our efforts to integrate web services in general, and the OBA in particular, into our text mining infrastructure using UIMA. We will discuss the benefits of service-oriented architectures from the user's perspective, as well as the drawbacks. Our discussion will conclude with an overview of the tools that have been integrated into the UIMA framework and made available publicly by the Hunter Lab.

INTEGRATION OF DIYA OUTPUT WITH GMOD STANDARDS

Presenter: Inna Rytsareva, Mississippi Valley State University

Authors: Inna Rytsareva, Charles Bland, Abigail Newsome

ABSTRACT: Genome annotation is the process of embellishing raw DNA sequences with predictions of features such as genes and transcription factor binding sites. These assignments are necessary to identify important gene functions and to enable comparative analysis. Unfortunately the path from DNA sequence to annotated genome is hampered by the lack of standards in various output formats within the annotation pipeline. GMOD is the Generic Model Organism Database project, a collection of open source software tools for creating and managing genome-scale biological databases. There are two general GMOD standards: GFF and Chado. GFF is a compact format for describing sequence and sequence annotations. Chado is an ontology-based modular schema for representing genome-associated biological information. DIYA (Do-It-Yourself Annotator) has been approved as an official GMOD project. It is a modular and configurable open source pipeline framework, written in Perl, used for the rapid annotation of microbial genome sequences. The software is currently used to take nucleotide sequence contigs as input, either in the form of complete genomes or the result of shotgun sequencing, and produce an annotated sequence. In the present study GFF3 and Chado compliance for DIYA output has been completed and hence facilitated immediate connectivity with GMOD databases and tools. The proposed approach has allowed for the generation of GFF files from DIYA output as it exists already and the identification of areas that need modification in order to increase compatibility with the BioPerl library and Chado databases.

PROTEIN SECONDARY STRUCTURE IS ROBUST UNDER ARTIFICIAL EVOLUTION WHILE PROTEIN DISORDER IS NOT

Presenter: Christian Schaefer, Columbia University, New York City

Authors: Christian Schaefer, Avner Schlessinger, Burkhard Rost

ABSTRACT: Single amino acid mutations often impact protein function and structure. Here, we study a particular aspect of robustness under mutations,

namely regular secondary structure (helices and sheets) and intrinsically disordered regions (often in non-regular secondary structure). Is the formation of regular secondary structure an intrinsic feature of amino acid sequences? Similarly, is disorder an intrinsic sequence feature? To tackle these questions, we in silico mutated native protein sequences gradually into random sequence-like ensembles and monitored the change in predicted secondary structure (PROFsec) and disorder (IUPred, MD, VSL2). By first establishing that by our coarse-grained measures for change, predictions and observations were similar we ruled out that our results were biased by prediction mistakes. Strings of secondary structure (three states) and disorder (two states) began to differ from the native at a rate roughly linearly proportional to the change in sequence. Surprisingly, neither the secondary structure content nor the length distribution of helices and strands changed substantially. Regions with long disorder (>30 consecutive residues) behaved very differently: they rapidly disappeared during in silico mutation. Our findings suggest that the formation of regular secondary structure is an intrinsic feature of random amino acid sequences, while the formation of long disordered regions is not an intrinsic feature of proteins with disordered regions. Put differently: helices and strands are easy to maintain by evolution, whereas disordered regions are difficult to maintain. Mutations that are neutral with respect to disorder are therefore extremely unlikely.

VISUALIZING GENOMIC SEQUENCES IN 2D

Presenter and Author: Josiah Seaman, Colorado State University

ABSTRACT: It is increasingly evident that there are multiple and overlapping patterns within the genome, and that these patterns contain different types of information — regarding both genome function and genome history. In order to discover additional genomic patterns which may have biological significance, novel strategies are required. To partially address this need, we introduce a new data visualization tool entitled Skittle. It first creates a 2-dimensional nucleotide display by assigning four colors to the four nucleotides, and then text-wraps to a user adjustable width. This nucleotide display is accompanied by a ‘repeat map’ which comprehensively displays all local repeating units, based upon analysis of all possible local alignments. Skittle includes a smooth-zooming interface which allows the user to analyze genomic patterns at any scale. Skittle is especially useful in identifying and analyzing tandem repeats, including repeats not normally detectable by other methods. However, Skittle is also more generally useful for analysis of any genomic data, allowing users to correlate published annotations and observable visual patterns, and allowing for sequence and construct quality control. Preliminary observations using Skittle reveal intriguing genomic patterns not otherwise obvious, including structured variations inside tandem repeats.

**IMPROVING THE PREDICTION OF MICRORNA-TARGET GENES BY
COMBINING TARGET PREDICTION ALGORITHMS**

Presenter: Ashok Sharma, Medical College of Georgia
Authors: Ashok Sharma, Ryan Rimando, Richard McIndoe

MicroRNAs are about 22 nucleotide-long, non-coding, endogenous RNA molecules which can regulate gene expression by the translational repression or cleavage of mRNA targets. It is now well established that miRNAs play key roles in many pathways and processes in both physiological and pathological conditions. It is estimated that up to 30% of mammalian genes are regulated by miRNAs, but relatively few have been experimentally validated. The accurate prediction and validation of genes targeted by miRNAs is a major hurdle for miRNA research. Several computational approaches have been developed to predict the miRNA target genes. However, the degree of overlap between predicted targets using these algorithms is very poor. As a result, there is a great need for a method to provide a measure of reliable predictions of miRNA target genes. We have developed a method and resource to assess reliable miRNA-target predictions by combining five available algorithms (miRanda, MirTarget, PicTar, PITA, TargetScan). Most of these algorithms devised to predict the miRNA targets use parameters such as sequence matches, structural information, free energy of the interaction, and conservation across species. By utilizing a wide variety of miRNA-target prediction algorithms, we can capture as many parameters as possible. We evaluated each algorithm against a set of experimentally validated targets and developed a method to calculate the normalized combined weighted score for each miRNA-target gene pair. We also developed a user interface with which one could query for each miRNA or target gene to find its interaction partners and their respective target quality scores.

SIMPLIFIED CLUSTERING WITH DIRICHLET PROCESS AND OTHER PROCESS MIXTURES

Presenter: Matthew Shotwell, Medical University of South Carolina
Authors: Matthew Shotwell, M.S., Elizabeth Slate, Ph.D.

ABSTRACT: Longitudinal clustering is a valuable tool for bioinformatics applications, for example in clustering gene expression probes with similar profiles in a microarray time series. Inferences drawn from such clustering aid in elucidating novel gene functions. The Dirichlet process (DP) mixture model is well suited to longitudinal clustering when the number of clusters is unknown, as inference on the number of clusters is a natural consequence of the model. Bayesian inference for DP mixture models is dominated by posterior sampling through Markov chain Monte Carlo (MCMC) methods. These methods are computationally expensive, require expertise, and yield results that are burdensome to interpret in the context of clustering. We present a Bayesian inference

mechanism for the DP mixture model that is based on optimization of a posterior likelihood, rather than sampling via MCMC. The method introduces cluster indicators and conditions on their maximum a posteriori (MAP) estimate. The resulting ‘profile’ posterior distributions for the longitudinal models are generally more tractable than the joint posterior distribution. In addition to the DP mixture, we present profile inference in a Dirichlet motivated process mixture, generated by modifying the prior distribution for the cluster indicators. This alternative process mixture is shown to have superior properties in many clustering problems. Profile inference in these models is evaluated through a simulated data example and with experimental data from a yeast cell cycle time series. Software implementation is available through the R package profdpm.

EVOLVING SPIKING NEURAL NETWORKS FOR THE PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES

Presenter: Heike Sichtig, UF Genetics Institute

Authors: Heike Sichtig, Alberto Riva

ABSTRACT: Our understanding of complex biological adaptive systems, from the cellular to the molecular level, can be used to develop valuable computational tools for interdisciplinary research in bioinformatics and biomedical engineering. We propose the use of spiking neural networks, able to realistically model the neurological system, to address challenging problems in computational biology, and of artificial evolutionary processes, such as genetic algorithms, to tune the network parameters. These tools can be extremely useful for interdisciplinary research because of their generality and their applicability to any complex system of interest. We will present work in progress using artificial spiking neurons applied to the well-known problem of predicting transcription factor binding sites (TFBSs) in DNA sequences. The system is trained using real TFBS data from the TRANSFAC database, and uses a top-down modeling approach to simulate biological information processing based on neurological coding. The goal of our work is to reduce the number of false positives in the predicted TFBSs, through a more accurate modeling of the information contained in the alignments in the training data. We will present an evaluation of our system’s performance for the detection of TFBSs and compare it to alternative methods.

MCW PROTEOMICS ANALYSIS PLATFORM – A HYBRID CLOUD COMPUTING ARCHITECTURE FOR PROTEOMICS ANALYSIS

Presenter: Simon Twigger, Medical College of Wisconsin

Authors: Andrew Vallejos, Brian Halligan, Joey Geiger, Andrew Greene, Simon Twigger

ABSTRACT: Mass spec-based proteomics rests on the availability of computing resources with which to analyze the resulting spectra and make protein identifications. Over recent years we have built an internal proteomics analysis workflow incorporating Sequest and Mascot, two of the major commercial analysis

packages. These are run on local clusters, with the size of the cluster limited both by availability of hardware and cost of the per-node licenses. A shift to incorporate two open source algorithms, OMSSA and !XTandem, has allowed us to explore additional architectures, including incorporating Amazon's cloud computing platforms. We have previously released ViPDAC, a stand alone cloud computing solution for analysis using OMSSA and X!Tandem available as a public amazon machine image. Building on this we have integrated our local cluster with the cloud architecture to create a hybrid model, enabling researchers to run analyses across all four algorithms on both local and cloud hardware, all from a single interface. Incorporation of the cloud has allowed us to provide expandable analysis capacity at very low cost and has enabled us to consider different analysis approaches that would have been unattainable with the financial and infrastructure constraints associated with local hardware and proprietary software. We will present an overview of our system and of our experiences using the Amazon cloud in a production environment.

ACCELERATING CANDIDATE GENE DISCOVERY THROUGH ONTOLOGICAL INDEXING OF LARGE SCALE DATA REPOSITORIES

Presenter: Simon Twigger, Medical College of Wisconsin

Authors: Simon Twigger, Joey Geiger, Jennifer Smith

ABSTRACT: Are any of these genes associated with my disease or phenotype? Is this candidate gene expressed in my tissue of interest? These are examples of common questions asked virtually every day by scientists attempting to identify genes contributing to human disease. Model Organism Databases such as the Rat Genome Database (RGD) curate published data related to these questions but there is much more information available than can be manually curated. Much of this information is being deposited into large scale data repositories but extracting useable information and knowledge from this stored data is a challenging problem. We are tackling this by annotating data in repositories such as NCBI's Gene Expression Omnibus (GEO) with biomedical ontologies using the National Center of Biomedical Ontology's web services. We are using an iterative process to automatically annotate the GEO records with ontology terms, followed by a manual curation/review using GMiner, a custom ruby on rails web application. Following review the data can then be explored on GMiner, allowing researchers to quickly find GEO datasets expressed in particular tissues and explore other attributes of those datasets. In addition we are creating an extensive annotation dataset linking genes to the tissues in which they have been expressed and making this available in RDF format to allow us to explore the benefits of integration with other semantic web resources. I will present the results of this annotation initiative and our plans to use these and other ontology annotations to accelerate candidate gene discovery.

DIGGING FOR GOLD – DATA ANNOTATION AND EXPLORATION WITH RATMINE

Presenter: Simon Twigger, Medical College of Wisconsin

Authors: Andrew Vallejos, Jennifer Smith, Richard Smith, Julie Sullivan, Gos Micklem, Simon Twigger

ABSTRACT: RatMine (<http://ratmine.mcw.edu>) provides the ability to create, evaluate and manipulate a wide variety of information about rat genes and proteins through a user-friendly web interface. One of the challenges facing researchers today is not only the growing availability of data and data sources but also the need to manipulate the data found in these resources. Databases such as the Rat Genome Database curate and store a large amount of information across a wide variety of biological areas. Ratmine complements RGD by providing a way to further explore and manipulate the data found in RGD in conjunction with data from other bioinformatic resources. Built as a data warehouse using the Intermine data mining platform, Ratmine combines the data from RGD with SNP data from Ensembl, pathway data from KEGG, protein data from Uniprot within a single environment. At its basic level Ratmine allows researchers to quickly evaluate a list of rat genes by providing enrichment analyses for GO, disease, phenotype and pathway annotations and listing known pathways and genome locations. Lists of genes or proteins can be compared and contrasted, combined, split, saved for later analysis and even made public for others to use. More complex queries can be created through the use of the query ‘template’ builder and saved for future use and can also be shared with the wider community. Overall, the use of the Intermine infrastructure has enabled us to create a powerful, yet user-friendly, data mining platform with comparatively little pain.

AN ALGORITHM TO DESCRIBE ALL OPTIMAL COMPARATIVE GENOME MAPS

Presenter: Zachary Vaughan, University of Colorado, Boulder

Authors: Zachary D. Vaughan, Debra S. Goldberg

ABSTRACT: Comparative genome maps, which identify related chromosomal regions in two species, are a powerful tool for leveraging what we know about one organism to gain insights about related organisms. They are useful in many ways, for instance to combine related genetic information about different organisms, for inferring phylogenetic relationships, and for gaining valuable information on human diseases by examining the genomes of other organisms. Given user preferences described with input parameters, current algorithms compute an optimal comparative genome map, which labels each chromosomal region of one species with a syntenic chromosomal region of the other. However, there are often a number of other maps that are also optimal. Knowing the variability between differing optimal maps can inform us about the robustness of the maps, help us reconcile differences needed to create symmetrical maps (where data from both species are used equally), and help us compute multi species maps. We have

modified comparative mapping software to indicate all optimal labelings for each chromosomal region. Although there may be an exponential number of optimal maps, this can be done in polynomial time.

LARGE-SCALE SEQUENCING OF T-CELL RECEPTOR REPERTOIRES IN DIABETIC NOD MICE

Presenter: Vijetha Vemulapalli, University of Colorado, Denver

Authors: Vijetha Vemulapalli, Todd A Castoe, Maki Nakayama, George Eisenbarth, David D Pollock

ABSTRACT: Type 1 Diabetes (T1D) is a T-cell mediated autoimmune disease where pancreatic beta cells are destroyed due to false identification of certain epitopes presented on these cells as non-self proteins. Previous studies have shown that insulin and pro-insulin tend to be among the most frequently targeted self molecules in cases of T1D. At a molecular level, however, little is known about the ways in which such autoimmunity develops in terms of which T-cell receptors (TCRs) target self-peptides in combination with which major histocompatibility complex (MHC) molecules. To study this, we implemented a high-throughput approach to sequence large samples of the TCR repertoire in various tissues in NOD mice 'a model of T1D with a single MHC phenotype' using the 454 sequencing platform. To analyze these large datasets of TCR cDNA sequences, we also developed a fast and efficient algorithm to identify and classify variants based on their exons and mutations. We are currently developing computational approaches to understand the diversity of the repertoire and to compare repertoires of different tissues. The overall goals are to be able to identify the TCR sequences that are diabetogenic and to further identify particular common features of such sequences that are responsible for their diabetogenic nature. Understanding this may shed light on the causes and progression of T1D, and other autoimmune diseases.

BOOLEAN NETWORK MODELS OF HUMAN AGING

Presenter: Michael Verdicchio, Arizona State University

Authors: Michael Verdicchio, Seungchan Kim

ABSTRACT: The systems biology of human aging is a complex, quantitative process. Many theories regarding senescence involve the roles of cellular components, such as mitochondria and lysosomes, as well the transportation and accumulation of various entities within and without the cell. In recent years, work by John Furber has amalgamated the research of many prominent aging biologists into a large chart illustrating many of the leading theories on human aging. The chart is organized into cellular components and describes many intricate, quantitative processes, along with their input and output entities. The representation, however, is not formalized as it is designed to be read and interpreted by humans. Boolean networks, which were first introduced by Kauffman and more recently applied to systems biology by Shmulevich et al., are a model well-suited in application to aging studies. The

ability to interpret attracting states and their basins of attraction in light of their biological meanings could greatly simplify our understanding of essential processes in human aging. We construct a Boolean network representation of part of Furber's chart and perform a new type of analysis on the systems biology of aging through the exploration of Boolean network attractors and their basins. Preliminary results have shown a division of attractor states into healthy and unhealthy sides and analysis of essential variables within the large basins of attraction have revealed key entities and processes responsible for leading to these healthy and unhealthy attractors. Our current expansion of the model and collaboration with biologists will facilitate further understanding.

EVOLUTION OF A PLACENTA SPECIFIC REGULATORY NETWORK

Presenter: Thomas Walsh, Dublin City University

Authors: Thomas Walsh, Kieran Holohan, Anna O'Brien, Elinor Velasquez, Mary O'Connell

ABSTRACT: Here we examine the respective roles played by gene duplication and loss, natural selection and regulation in the evolution of a placental network. We have investigated the selective pressures exerted on a gene regulatory network and the influence of protein interactions on those evolutionary rates. Our dataset consists of a set of placenta essential genes that are co-expressed with epidermal growth factor receptor (EGFR). Each of the major developmental time points for placenta are represented in our dataset, and the network has been defined using Bayesian methods. The protein interaction network for each of these placental proteins has also been determined and included in our analysis. We have inferred the evolutionary history of each member of the regulatory network and the interacting network across a number of completed genomes (i.e. placental mammals, opossum, platypus and chicken). In doing so, we establish whether the selective pressures acting on the genes in this EGFR network are interdependent. With respect to identifying putative regulatory elements controlling the expression of these genes at specific time points, we focused specifically on identifying putative microRNA target sites. Our results show that the elements of this placental regulatory network have undergone functional shift in the ancestral placental mammal, notably in the EGFR hub protein itself. We show that the gene expression in this network is tightly regulated by a small subset of placenta-specific microRNAs.

PATTERN-BASED EXTRACTION OF ARGUMENTATION FROM THE SCIENTIFIC LITERATURE

Presenter and Author: Elizabeth White, University of Colorado, Boulder

ABSTRACT: As the number of publications in the biomedical field continues its exponential increase, techniques for automatically summarizing information from this body of literature have become more diverse. In addition, the targets of summarization have become more subtle: initial work focused on extracting the

factual assertions from full-text papers, but more recently, interest has shifted to recovering speculations and agreements or disagreements with other research. Scientific writing is rife with such argumentation, and the premises, evidence, conjectures, objections and rebuttals that writers use to persuade the reader represent a rich vein of expert knowledge for summarization. Agreement, disagreement, and conjecture are often expressed in highly scripted ways; likewise, the higher-order discourse structures that underpin multisentence arguments tend to assume particular forms into which claims and evidence can be nested. These features make these kinds of arguments readily recoverable by pattern-based search. Here, I present PARROT, which uses OpenDMAP patterns in combination with a Protégé ontology. PARROT first matches simple argumentative claims using a set of concepts relevant to scientific discourse and then exploits discourse cues and inference to combine these claims recursively into higher-order argument trees. PARROT outperforms an SVM classifier system in identifying statements of support and conflict at the sentence level. Additionally, PARROT provides a graphical representation of the arguments it finds, which makes it an valuable tool for summarizing the reasoning behind scientists' conclusions and identifying areas of consensus and contention.

AN INDIVIDUAL BASED MODELLING APPROACH TO STUDYING THE EVOLUTION OF MATE CHOICE STRATEGY

Presenter: Robert Williamson, Rose-Hulman Institute of Technology

Authors: Robert Williamson

ABSTRACT: Sexual selection is a key force driving morphological and behavioral evolution. It can lead to alternate life strategies within a species (for example some males fighting for mates and some sneaking copulations), or the development of exaggerated traits (like large size or elaborate ornaments). Key factors determining how sexual selection will affect a species' evolutionary trajectory are the mate choice strategies of female choice and male-male competition. Female choice encompasses the good gene, resource supply, and sensory exploitation strategies; while male-male competition consists of the direct confrontation, sperm competition, and infanticide strategies. Because each of these strategies affects the life histories of species that exhibit them, it is important to understand what factors may influence the development of each strategy. I developed an individual based model (IBM) system using the Swarm java library to investigate how various biotic and abiotic factors influence the evolution of different mate choice strategies, and which factors can cause shifts between the strategies. The factors included features like resource availability, gestation length, sex ratio, and infanticide rate. I will discuss the design and implementation of the system, including architecture and built in assumptions. I will also address the success of the IBM method to this type of scientific inquiry.

EVALUATING GENE-DISEASE ASSOCIATION PREDICTIONS

Presenter: Laura Wojtulewicz, Arizona State University

Authors: Graciela H. Gonzalez, Fabian Spinnherin, Juan C. Uribe, Annie Skariah, Laura Wojtulewicz, Hailong Cui

Many methods have been proposed for facilitating the uncovering of genes that underlie the pathology of different diseases. Some are purely statistical, resulting in a (mostly) undifferentiated set of genes that are differentially expressed (or co-expressed), while others seek to prioritize the resulting set of genes through comparison against specific 'known' targets. Evaluation of these methods is usually done through the use of disease-specific data sets. It is advantageous to generalize evaluation methods in order to compare prediction methods across diseases and, most importantly, to get a sense of the value of the method to suggest potentially novel targets from non-empirical knowledge sources automatic extraction from the literature and curated databases). We present here a quantitative and comparative evaluation of GeneRanker across 34 different diseases, using an adaptation of take-one-out cross validation that can be more suitable for prediction-based methods.

FACTORS THAT CONTROL FUNCTIONALITY OF SESQUITERPENE SYNTHASES FROM PHYLOGENETIC AND BIOPHYSICAL SIMULATIONS

Presenter: Troy Wymore, National Resource for Biomedical Supercomputing

Authors: Troy Wymore, Brian Y. Chen, Hugh B. Nicholas Jr, Alexander J. Ropelewski, Charles L. Brooks III

ABSTRACT: The attainment of new catalytic functions from an existing protein scaffold is a major force guiding evolutionary change but one that is perhaps only beginning to be understood. Understanding the evolution of enzymatic function at a physiochemical level requires first that probable evolutionary paths that interconvert an enzyme's specific function from one to another through accessible mutational changes be discovered and thus a catalytic landscape be defined. Through a landmark study of terpene synthases (O'Maille et al, Nature Chemical Biology, 2008) in which one with a specific activity was engineered to obtain a different specific activity through mutational swaps of nine residues as well as characterization of 418 proteins with different combinations of these nine residues, a catalytic landscape underlying the evolution of sesquiterpene chemical diversity was revealed. In this presentation, we will describe both phylogenetic analyses and biophysical simulations of sesquiterpene synthases to begin elucidating at the residue and atomic level the factors that control the function of these enzymes. Specifically, we will present a mechanism of 5-epi-aristolochene synthase and a hypothesis for how residues (many of them outside the active site) can convert the enzyme to a 4-epi-eremophilene and premenaspirodiene synthase.

STRUCTURE-BASED KERNELS FOR THE PREDICTION OF CATALYTIC RESIDUES AND THEIR INVOLVEMENT IN DISEASE

Presenter: Fuxiao Xin, Indiana University, Bloomington

Authors: Fuxiao Xin, Steven Myers, Yong Fuga Li, David N. Cooper, Sean D. Mooney, Predrag Radivojac

ABSTRACT: We propose a new kernel-based algorithm for the prediction of catalytic residues based upon protein sequence, structure, and evolutionary information. The method relies on explicit modeling of similarity between residue-centered structural neighborhoods. We constructed oriented structural neighborhoods for each residue based on the directionality of covalent bonds between backbone atoms. The volume within the spherical spatial neighborhood was then divided into cells and the similarities between pairs of corresponding cells in two structural neighborhoods were used to calculate the kernel matrix. The kernel function is a product of three kernels, each addressing a separate aspect of protein function: (i) geometric kernel addresses shape similarity, (ii) chemical kernel addresses similarity between physicochemical properties, and (iii) evolutionary kernel addresses similarity of conservation patterns. The method favorably evaluates against state-of-the-art approaches and also provides insights into the relative importance of geometry, physicochemical properties, and evolutionary conservation of catalytic residue activity. We used our method to identify disease-causing mutations whose molecular mechanism is predicted to be the loss or gain of catalytic residues. We suggest that it is a viable approach for identifying the molecular basis of disease, and this approach can be applied to functional sites prediction in general.

EVOLUTIONARY STUDY AND PREDICTION OF PROTEIN-PROTEIN INTERACTIONS IN CHROMATIN MODIFICATION COMPLEXES

Presenter: Xuejian Xiong, Hospital for Sick Children

Authors: Xuejian Xiong, Tuan On, Shuye Pu, Andrei Turinsky, Yunchen Gong, Andrew Emili, Zhaolei Zhang, Jack Greenblatt, Shoshana J. Wodak, John Parkinson

ABSTRACT: Chromatin modification (CM) comprises an array of broadly conserved biological processes that modify chromatin to control access to DNA. Due to its fundamental role in processes such as transcription, DNA replication and repair, many components of CM are thought to play an important role in various forms of cancer and cancer related developmental processes. While CM machinery has been extensively characterized in yeast, less is known about its operation in other eukaryotes. Here 111 eukaryotic genome sequences are used to systematically profile the conservation and evolution of CM by applying the InParanoid algorithm based on the comprehensive surveys of literature and database resources. Furthermore, we examined the distribution of the evolutionary

distinct groups within the yeast protein-protein interaction network. Many established CM complexes were readily identifiable. Particularly striking is the mosaic nature of the conservation patterns of the components of each complex. In addition, we exploit the conserved interactions among five model species, i.e. yeast, worm, fly, mouse and human for CM complexes. In RSC/SWI/SNF complexes, the interaction of SWI3-SNF5 is conserved in all five models, while that of SWI3-SNF2 is conserved in four models except worm. Based on the evolutionary study of the complexes, we can predict that the triangle interactions among SWI3, SNF2, and SNF5 are conserved across model species. Together these findings provide insights into the evolution of CM and will help facilitate an improved annotation of CM across eukaryotes, and the prediction of the protein-protein interactions.

SIMPLEMHC: A NOVEL METHOD FOR IN-SILICO PREDICTION OF PEPTIDE BINDING TO MHC CLASS II MOLECULES

Presenter: Li Xue, Iowa State University

Authors: Li Xue, Arthur Fridman

ABSTRACT: Immunogenicity of therapeutic proteins, including monoclonal antibodies, is a significant concern during drug development. Protein immunogenicity is believed to be at least in part due to the presence of helper T-cell epitopes in the protein sequence. Accordingly, screening protein therapeutic candidates in-silico for the presence of T cell epitopes has potential to be an important step in the clinical development. Because binding to major histocompatibility complex Class II receptors (MHCII) is a necessary condition for T-cell immunogenicity, reliable prediction of binding affinity to MHCII plays a key role in this task. We propose a simple classifier, called SimpleMHC, based on single-state Hidden Markov Model (HMM), to predict peptide binding affinity to MHCII. We tested SimpleMHC on a recently available large dataset and observed comparable performance with state-of-art MHC class II prediction methods ProPred, ARB, and SMM-align. and. We further developed a hybrid model that combines the predictions of SimpleMHC and ProPred. The hybrid model had significantly better accuracy identifying MHCII binders than ProPred, ARB, SMM-align, and SimpleMHC alone.

PROTEIN-PROTEIN INTERACTIONS ARE DRIVEN BY FUNCTIONAL EVOLUTION

Presenter: Yiqiang Zhao, Buck Institute for Age Research

Authors: Yiqiang Zhao, Sean Mooney

Many fundamental biological processes involve protein-protein interactions. It is an interesting question that how and why proteins are becoming interacted and contributes to the genetic complexity of organisms. The preferential attachment model, when applied to protein-protein interaction, asserts that a protein is more

likely to evolve to have many interaction partners if it has many interaction partners before the evolution. Consistent with the model we find that older genes tended to have more interactions than newer ones. Two findings, on the other hand, were not consistent with the preferential attachment model. One analysis divided human genes into 7 temporal groups. Most interactions, both new and old, involved gene partners generated not in the oldest temporal groups, but rather in the temporal group representing the evolution from vertebrates to warm-blooded animals. A second analysis found that there was not an increased number of interactions in 645 gene clusters with historic duplication events, which is inconsistent with the hypothesis if the preferential attachment governs the evolution of protein-protein interaction. Based on these two analyses, it does not appear that the preferential attachment model adequately explains the evolution of protein-protein interactions and we suggest that the interactions are selected to increase species-special complexity and achieve species-special functions in the evolution history instead of simply being driven by “rich get richer”.

NOTES

NOTES



NOTES

NOTES