



**FOURTH  
ROCKY MOUNTAIN  
BIOINFORMATICS  
CONFERENCE**



**SNOWMASS / ASPEN  
COLORADO**

**DECEMBER 1-3, 2006**



Conference Co-chairs:  
Lawrence Hunter, Ph.D.  
and Susan Trapp Ph.D.

University of Colorado School of Medicine





Dear Rocky '06 Participant,

Welcome to the fourth annual Rocky Mountain Bioinformatics Meeting, a conference of the International Society for Computational Biology (ISCB). The organizers hope that you enjoy the program, and find the meeting a productive opportunity to meet researchers, students and industrial users of bioinformatics technology. We think we have the best program yet, offering a remarkable cross-section of bioinformatics research.

The Rocky series began four years ago as a regional meeting, and has grown into an international program with a spotlight on regional development in the computational biosciences. Presenters at this year's meeting come from the Rocky Mountain regional states/provinces of Alberta, Colorado, Nevada, New Mexico, Utah and Wyoming, as well as from over 10 other states/provinces and at least six countries. These scientists represent a broad spectrum of universities, industrial enterprises, government laboratories, and medical libraries. Among our keynote speakers Prof. Ptitsyn just moved to the Rocky Mountain area to take a new faculty position, Prof. Wishart is a longstanding member of our regional community, and Dr. Myers got his Ph.D. in the area. The meeting is a chance to get to know your colleagues near and far, seek collaborative opportunities, and find synergies that can drive our field forward.

For 2006 we've expanded the science by extending the meeting to three full days, increased the duration of the short talks to 10 minutes, and enlarged the poster session, allowing everyone who wants to present to do so. We've retained the education workshop to discuss issues central to training the next generation of bioinformaticians, making it more relevant than ever to students and employers. And we've retained the lunchtime ski breaks so that those of you who want to ski, including beginners, will have a chance to do so among friends and colleagues.

We should all be grateful for the support of our sponsors: IBM, Affymetrix, HP, Bodesix and Dharmacon. It is only with the help of our corporate sponsors that we can make this meeting as affordable as it is. Even more important than the money, is the intellectual contribution of these companies. The keynotes from Bodesix and IBM lay out some of the challenges we face and some of the resources we will need to address them. We are also thankful to the 2006 ISCB Overton Prize winner, Mathieu Blanchette, who donated his financial award to this conference to enable us to keep student registration fees low. The meeting would simply not be possible without organizational help from the ISCB staff, as well as from Kathy Thomas at the University of Colorado School of Medicine.

Many of you have contributed \$5 to join the Rocky Mountain Regional Bioinformatics group, an affiliate of the ISCB. So far, the only activity of the Group has been to organize this meeting. If you are interested in volunteering to get something else going, please talk to Larry Hunter.

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

Larry Hunter and Susan Trapp, Rocky '06 co-chairs

# AGENDA

## THURSDAY

NOVEMBER 30

3:00 pm – 6:00 pm Registration

## FRIDAY

DECEMBER 1

7:30 am – 8:30 am Registration

8:30 am – 9:15 am **INVITED KEYNOTE 1: PETER HAUG**, MD, Senior Informaticist for Intermountain Health Care and Professor, Department of Biomedical Informatics, University of Utah  
*Data Mining, Clinical Modeling and the Future of Healthcare Computing*

9:15 am – 9:25 am **ORAL PRESENTATION 1: Computer Aided Genetic Engineering**  
Presenter: **DAN MCSHAN**, University of Colorado School of Medicine  
Author(s): Dan McShan

9:25 am – 9:35 am **ORAL PRESENTATION 2: Designing C2H2 Zinc Finger Proteins to Target Specific Genomic Sites**  
Presenter: **PETER ZABACK**, Iowa State University/Bioinformatics and Computational Biology  
Author(s): P. Zaback, J. Sander, F. Fu, D. Voytas, D. Dobbs

9:35 am – 9:45 am **ORAL PRESENTATION 3: Biological Network Inference and the SEBINI Software Environment**  
Presenter: **RONALD TAYLOR**, Pacific Northwest National Laboratory  
Author(s): Ronald Taylor, Anuj Shah, Charles Treatman, and Meridith Blevins

9:45 am – 9:55 am **ORAL PRESENTATION 4: The BioCyc Collection of Pathway/Genome Databases**  
Presenter: **ALEXANDER G. SHEARER**, SRI International  
Author(s): Alexander G. Shearer, Ron Caspi, Carol A. Fulcher, Pallavi Kaipa, Peter D. Karp

9:55 am – 10:05 am **ORAL PRESENTATION 5: Spatio-Temporal Modelling of Biochemical Pathways: Introducing Cell++**  
Presenter: **JOHN PARKINSON**, Hospital for Sick Children/University of Toronto  
Author(s): Chris Sanford, Matthew Yip and John Parkinson

10:05 am – 10:15 am **ORAL PRESENTATION 6: Minimal Oscillatory Currents in Neurons**  
Presenter: **TOM MCTAVISH**, UCHSC  
Author(s): Tom McTavish

10:15 am – 10:25 am **ORAL PRESENTATION 7: Assessment of siRNA Inhibition Prediction Algorithms**  
Presenter: **KEVIN SULLIVAN**, Dharmacon, Inc  
Author(s): Kevin Sullivan

10:25 am – 10:45 am **BREAK**

10:45 am – 10:55 am **ORAL PRESENTATION 8: The Balance Between Activating and Inhibiting Connections Controls the Dynamics of Biological Networks**  
Presenter: **DANIEL MCDONALD**, University of Colorado at Boulder  
Author(s): Daniel McDonald, Meredith Betterton, Laura Waterbury, Rob Knight

- 10:55 am – 11:05 am **ORAL PRESENTATION 9:** *Beyond Ontologies: The WayCraft Scientific Knowledge Integration Framework*  
 Presenter: **FRANK D. RUSSO**, WayCraft Biosoftware  
 Author(s): Frank D Russo, Steven T Connolly, Ranga Chandra Gudavida, Angela Qu, Anil Jegga, Bruce J Aronow
- 11:05 am – 11:15 am **ORAL PRESENTATION 10:** *An Experiment in Mining Gene-Disease Relationships from Biomedical Literature*  
 Presenter: **GRACIELA GONZALEZ**, Arizona State University — School of Computing and Informatics, Dept of Biomedical Informatics  
 Author(s): Graciela Gonzalez, Juan C. Uribe, Luis Tari, Colleen Brophy, Chitta Baral
- 11:15 am – 11:25 am **ORAL PRESENTATION 11:** *SESSION CANCELLED*
- 11:25 am – 11:35 am **ORAL PRESENTATION 12:** *Bioinformatics as a Tool to Predict Gene Function in Williams-Beuren Syndrome*  
 Presenter: **HANNAH TIPNEY**, University of Colorado at Denver Health Sciences Center  
 Author(s): Hannah Tipney, Andy Brass, May Tassabehji
- 11:35 am – 11:45 am **ORAL PRESENTATION 13:** *Unified Maximum Separability Analysis for Molecular Profile Analysis and Biomarker Discovery*  
 Presenter: **QIUYAN HUO**, Johns Hopkins University School of Medicine  
 Author(s): Qiuyan Huo, Zhen Zhang
- 11:45 am – 11:55 am **ORAL PRESENTATION 14:** *Data Integration as a Foundation for Semantic Integration*  
 Presenter: **DANIEL J. MCGOLDRICK**, University of Colorado, Department of Pharmacology  
 Author(s): Daniel J McGoldrick, Lawrence Hunter
- 11:55 am – 12:40 pm **INVITED KEYNOTE 2:** **JUDITH COHN**, PhD, Technical Staff Member, Bioscience Division, Los Alamos National Lab  
*Dynamics Perturbation Analysis of SCOP Domains*
- 12:40 pm – 4:00 pm **BREAK**
- 4:00 pm – 5:00 pm **WORKSHOP SESSION 1: CREATING THE RIGHT BIOINFORMATICS RESUME TO LAND THE INTERVIEW**
- 5:00 pm – 5:10 pm **ORAL PRESENTATION 15:** *Dissecting Genetics of Host-Pathogen Interactions*  
 Presenter: **PETER T. HRABER**, Theoretical Biology & Biophysics, LANL  
 Author(s): Peter T. Hraber
- 5:10 pm – 5:20 pm **ORAL PRESENTATION 16:** *Phylogenetic Profiling via Partial Genomes: Applications to the Apicomplexa*  
 Presenter: **JAMES WASMUTH**, Hospital for Sick Children  
 Author(s): Jennifer Daub, Matthew Fagnani, Jose Peregrin-Alvarez, Chris Sanford, James Wasmuth, John Parkinson
- 5:20 pm – 5:30 pm **ORAL PRESENTATION 17:** *Genome-Wide Co-Expression Based Prediction of Differential Expressions*  
 Presenter: **YINGLEI LAI**, The George Washington University, Department of Statistics  
 Author(s): Yinglei Lai

# AGENDA

- 5:30 pm – 6:00 pm **BREAK**
- 6:00 pm – 10:00 pm **DINNER AND INVITED KEYNOTE 3:** Kirk Jordan, PhD,  
Emerging Solutions Executive, IBM Strategic Growth Business/  
Deep Computing  
*(dinner tickets available for purchase at registration desk)*

## SATURDAY

DECEMBER 2

- 7:30 am – 8:30 am **REGISTRATION**
- 8:30 am – 9:15 am **INVITED KEYNOTE 4: HEINRICH RODER**, DPhil, Chief Technology  
Officer, Biodesix Inc., Colorado  
*Comparative Mass Spectrometry for Clinical Applications*
- 9:15 am – 9:25 am **ORAL PRESENTATION 18:** *A Semi-Manual Method Using the  
ANCOVA Framework to Identify Expression Profiles in Time-Course  
Microarray Experiments*  
Presenter: **TZU LIP PHANG**, University of Colorado Health Sciences  
Center, Department of Medicine  
Author(s): Tzu Lip Phang, Katherina Kechris
- 9:25 am – 9:35 am **ORAL PRESENTATION 19:** *ARB — A Comprehensive Phylogenetic  
Sequence Analysis and Probe Design Software Environment*  
Presenter: **HARALD MEIER**, Technische Universitaet Muenchen, Informatik  
Author(s): W. Ludwig, R. Westram, Y. Kumar, H. Meier
- 9:35 am – 9:45 am **ORAL PRESENTATION 20:** *Does the Region of rRNA Sequenced Affect  
Conclusions from Microbial Community Analysis*  
Presenter: **ZONGZHI LIU**, University of Colorado  
Author(s): Zongzhi Liu, Catherine Lozupone, Micah Hamady, Rob Knight
- 9:45 am – 9:55 am **ORAL PRESENTATION 21:** *Tight Clusters F Proteins*  
Presenter: **ROMAN L. TATUSOV**, NCBI NLM NIH  
Author(s): Boris Kiryutin
- 9:55 am – 10:05 am **ORAL PRESENTATION 22:** *A Bayesian Network Model of  
Stromatolite Formation*  
Presenter: **JACK K. HORNER**, Science Applications  
International Corporation  
Author(s): Jack K. Horner
- 10:05 am – 10:15 am **ORAL PRESENTATION 23:** *Genomic Supertrees of Life*  
Presenter: **DAVIDE PISANI**, The National University of Ireland Maynooth  
Author(s): Davide Pisani, James A. Cotton, James O. McInerney
- 10:15 am – 10:25 am **ORAL PRESENTATION 24:** *Modeling Evolution of Gene and  
Protein Networks*  
Presenter: **TODD A. GIBSON**, University of Colorado Health Sciences Center  
Author(s): Todd A. Gibson, Debra S. Golderg
- 10:25 am – 10:45 am **BREAK**

- 10:45 am – 10:55 am **ORAL PRESENTATION 25:** *The Unique-eome: What Makes Species Different?*  
 Presenter: **MARTIN GOLLERY**, University of Nevada, Reno  
 Author(s): Martin Gollery, John Cushman, Jeff Harper, Taliah Mittler, Ron Mittler
- 10:55 am – 11:05 am **ORAL PRESENTATION 26:** *Monte Carlo EM Algorithm for Sequence Motif-finding*  
 Presenter: **CHENGPENG BI**, Children’s Mercy Hospitals  
 Author(s): Chengpeng Bi
- 11:05 am – 11:15 am **ORAL PRESENTATION 27:** *Biomarker Discovery in Genomic Data with Partial Clinical Annotation*  
 Presenter: **COLE HARRIS**, Exagen Diagnostics, Inc.  
 Author(s): Cole Harris, Noushin Ghaffari
- 11:15 am – 11:25 am **ORAL PRESENTATION 28:** *Systematic Gene Selection for Dynamic Gene-Network Refinement*  
 Presenter: **CHRISTIAN V. FORST**, Los Alamos National Laboratory  
 Author(s): Nicole Radde, Jutta Gebert, Christian V. Forst
- 11:25 am – 11:35 am **ORAL PRESENTATION 29:** *Using an Anatomy Ontology to Predict the Spread of Tumor Cells to Regional Metastatic Sites*  
 Presenter: **IRA KALET**, University of Washington  
 Author(s): Ira Kalet
- 11:35 am – 11:45 am **ORAL PRESENTATION 30:** *Analyzing Time Course Microarray Data With Temporal Uncertainty*  
 Presenter: **STEPHEN BILLUPS**, University of Colorado at Denver and Health Sciences Center, Dept. of Mathematical Sciences  
 Author(s): Stephen Billups
- 11:45 am – 11:55 am **ORAL PRESENTATION 31:** *Energy Landscape Calculations for a DNA-Rotaxane using Milestoning Technique*  
 Presenter: **SHAHID QAMAR**, Arizona State University  
 Author(s): Shahid Qamar
- 11:55 am – 12:40 pm **INVITED KEYNOTE 5:** **DAVID S. WISHART**, PhD, Department of Computing Science and Department of Biological Science, University of Alberta  
*Metabolomics: The Next Frontier for Bioinformatics?*
- 12:40 pm – 4:00 pm **BREAK**
- 4:00 pm – 5:00 pm **WORKSHOP SESSION II: THE INTERVIEW AND NEGOTIATIONS**
- 5:00 pm – 5:10 pm **ORAL PRESENTATION 32:** *The SINE of the Opossum*  
 Presenter: **DAVID D. POLLOCK**, University of Colorado  
 Author(s): Wanjun Gu, David A. Ray, Jerilyn A. Walker , Erin Barnes, Andrew J. Gentles, Paul B. Samollow, Jerzy Jurka, Mark A. Batzer, David D. Pollock
- 5:10 pm – 5:20 pm **ORAL PRESENTATION 33:** *Strong Negative and Positive Selection Can Obscure Ancestral Signal in Phylogenetic Analysis*  
 Presenter: **ALEXANDER TCHURBANOV**, University of Wyoming  
 Author(s): Alexander Tchurbanov, Katherine Harris, Shruti Rastogi, David Liberles

# AGENDA

- 5:20 pm – 5:30 pm **ORAL PRESENTATION 34:** *Cyberenvironment for Computational Bioscience*  
Presenter: **THANH N. TRUONG**, Department of Chemistry, University of Utah  
Author(s): Thanh N. Truong
- 5:30 pm – 5:40 pm **ORAL PRESENTATION 35:** *A Computational Method to Identify RNA Binding Sites in Proteins*  
Presenter: **JEFF SANDER**, Iowa State University  
Author(s): J. Sander, M. Terribilini, J.H. Lee, R. Jernigan, V. Honavar, D. Dobbs
- 5:40 pm – 5:50 pm **ORAL PRESENTATION 36:** *Information Retrieval, Question-Answering, Machine Learning, and Concept Recognition in TREC Genomics 2006*  
Presenter: **J. GREGORY CAPORASO**, University of Colorado Health Sciences Center, Dept. of Biochemistry and Molecular Genetics  
Author(s): J. Gregory Caporaso, William A. Baumgartner, Jr., Hyunmin Kim, Zhiyong Lu, Helen L. Johnson, Olga Medvedeva, Anna Lindemann, Lynne Fox, Elizabeth K. White, K. Brettonel Cohen, and Lawrence Hunter
- 5:50 pm – 6:00 pm **ORAL PRESENTATION 37:** *FluKB: an Integrated Knowledge Base for Influenza Viruses*  
Presenter: Guoqing Lu, University of Nebraska at Omaha  
Author(s): Guoqing Lu, Kashi R Buyyani, Thaine W Rowley, Ruben Donis, Zhengxin Chen
- 6:00 pm – 6:10 pm **ORAL PRESENTATION 38:** *Improved Algorithms for Reaction Mapping*  
Presenter: **JOHN D. CRABTREE**, Colorado School of Mines  
Author(s): John D. Crabtree, Dinesh P. Mehta, J. Thomas McKinnon, Anthony M. Dean
- 6:10 pm – 6:20 pm **ORAL PRESENTATION 39:** *Biased Support Vector Machine and Kernel Methods for Tumor Classification*  
Presenter: **SRINIVAS MUKKAMALA**, New Mexico Tech  
Author(s): Andrew H Sung, Krishna Yendrapalli, Ram Basnet
- 6:20 pm – 6:30 pm **ORAL PRESENTATION 40:** *Identification of HTH Motifs from Amino Acid Sequence*  
Presenter: **CHANGHUI YAN**, Utah State University  
Author(s): Changhui Yan, Jing Hu
- 6:30 pm – 8:30 pm **RECEPTION AND POSTER SESSION**

## SUNDAY

- 7:30 am – 8:30 am **REGISTRATION**
- 8:30 am – 9:15 am **INVITED KEYNOTE 6: EUGENE MYERS**, PhD, Group Leader, Howard Hughes Medical Institute, Janelia Farms  
*Computer-Assisted Forensic Analysis of Mass Disasters*
- 9:15 am – 9:25 am **ORAL PRESENTATION 41:** *Improving Protein Function Prediction Methods with Improved Network Weighting and Integrated Literature Data*  
Presenter: **AARON GABOW**, UCDHSC  
Author(s): Aaron Gabow, Sonia Leach, Larry Hunter, Debra S. Goldberg

## DECEMBER 3

- 9:25 am – 9:35 am **ORAL PRESENTATION 42:** *A Topology-Based Clustering Algorithm for Analysis Very Large Biological Networks*  
 Presenter: **XIAOWEI XU**, University of Arkansas at Little Rock  
 Author(s): Xiaowei Xu, Zhidan Feng, Nurcan Yuruk
- 9:35 am – 9:45 am **ORAL PRESENTATION 43:** *The Phylogenetic Position of the Mitochondrion*  
 Presenter: **JAMES MCINERNEY**, National University of Ireland  
 Author(s): David A. Fitzpatrick, Christopher J. Creevey
- 9:45 am – 9:55 am **ORAL PRESENTATION 44:** *Co-conservation Analysis of Bacterial Genes Across Phyla Predict Gene Function*  
 Presenter: **ANIS KARIMPOUR-FARD**, University of Colorado Health Science Center  
 Author(s): Anis Karimpour-Fard, Corrella S. Detweiler, Ryan T. Gill, Lawrence Hunter
- 9:55 am – 10:05 am **ORAL PRESENTATION 45:** *De Novo Signaling Pathway Reconstruction From Multiple Data Sources*  
 Presenter: **DONGXIAO ZHU**, Stowers Institute for Medical Research  
 Author(s): Dongxiao Zhu, Michael Rabbat, Alfred O Hero, Robert Nowak, Mario Figueiredo
- 10:05 am – 10:25 am **BREAK**
- 10:25 am – 10:35 am **ORAL PRESENTATION 46:** *The Role of Discretization in Modeling Signal Transduction Networks*  
 Presenter: **DAVID J. JOHN**, Wake Forest University  
 Author(s): David J. John, Edward E. Allen, Leslie B. Poole, Richard F. Loeser, Jacquelyn Fetrow
- 10:35 am – 10:45 am **ORAL PRESENTATION 47:** *A Generalized Algorithm of Unsupervised Learning*  
 Presenter: **ANCA RADULESCU**, University of Colorado at Boulder  
 Author(s): Paul Adams, Kingsley Cox, Anca Radulescu
- 10:45 am – 10:55 am **ORAL PRESENTATION 48:** *Principal Component Models to Identify Co-and Differential- Gene Expression in Time-Course Microarray Data*  
 Presenter: **RAJAGOPALAN SRINIVASAN**, National University of Singapore  
 Author(s): Rajagopalan Srinivasan, Sudhakar Jonnalagadda
- 10:55 am – 11:05 am **ORAL PRESENTATION 49:** *Finding Informative Sentences in Full-Text Journal Articles*  
 Presenter: **ZHIYONG LU**, University of Colorado School of Medicine  
 Author(s): Zhiyong Lu, William A. Baumgartner, Jr., J. Gregory Caporaso, K. Bretonnel Cohen, Lawrence Hunter
- 11:05 am – 11:50 am **INVITED KEYNOTE 7: ANDREY PTITSYN**, PhD, Assistant Professor, Colorado State University  
*Life Works on AC Power*
- 11:50 am – Noon **ROCKY '06 CLOSING COMMENTS**

# EDUCATION PANEL

## TOPIC: “THE INTERVIEW PROCESS” TAILORED TO LANDING THAT BIOINFORMATICS COMPANY POSITION.

This year’s panel discussion workshop continues to build on previous years. The Rocky ’05 Education Panel Discussion focused on the skill set needed for a successful career in bioinformatics. This years Education panel will be interactive focusing on the entire interview process from resume to negotiations once you are offered the position, including “on stage” student mock interviews with bioinformaticians from industry. Below is a set of questions we hope you will gain insight on how to tackle from participating in this workshop. We encourage all participants of Rocky to take part in the Education Panel as it will be very interactive and your ideas, opinions, and thoughts will be invaluable to the success of this panel discussion.

- What makes you and your resume shine among the crowd of other Bioinformatics applicants?
- How to market yourself and use your resume and cover letter to your advantage that lands you the interview with your organization of choice.
- What are the interview skills within the interview process that gets you the job offer?
- How and when to negotiate, once they have offered you the position?

Facilitators for Sessions: KIRK JORDAN, IBM STRATEGIC GROWTH BUSINESS/DEEP COMPUTING  
SUSAN TRAPP, UNIVERSITY OF COLORADO SCHOOL OF MEDICINE

---

FRIDAY, DECEMBER 1, 2006 (4–5 PM), 1 HOUR

### Session I: CREATING THE RIGHT BIOINFORMATICS RESUME TO LAND THE INTERVIEW A. THE RESUME B. THE COVER LETTER

- 10 min The Resume & The Clever letter presentation by industry leader (*Kirk Jordan, IBM*)
- 10 min The Resume & The Cover letter presentation by industry leader (*Jack Horner, SAIC*)
- 10 min The Resume & The Cover letter presentation by industry leader (*Martin Gollery, Univ of Nevada, Reno*)
- 30 min Question answer period & resumes critique using student/postdoc participants (student postdoc participants)

---

SATURDAY, DECEMBER 2, 2006 (4–5 PM), 1 HOUR

### Session II: THE INTERVIEW & NEGOTIATIONS (AFTER THE OFFER): A MOCK INTERVIEW WILL BE PERFORMED LIVE ON STAGE

Followed by a Panel Discussion of the Interview Process

- 10 min Interview overview by Facilitator Kirk Jordan
- 10 min Mock Interview: student (interviewee) with industry leader (*Mike Lelivelt, Affy*)
- 10 min Mock Interview: (interviewee) with industry leader (*Kirk Jordan, IBM*)
- 5 min The Negotiations after job offer Overview by Facilitator Kirk Jordan
- 25 min Panel Discussion with Industry leaders & questions from audience (*Steve Lincoln, Affymetrix, Judith Cohn, Los Alamos Laboratories, Martin Gollery, Univ of Nevada Reno*)

## **Judith Cohn, PhD,**

Technical Staff Member, Bioscience Division, Los Alamos National Lab

### **Dynamics Perturbation Analysis of SCOP Domains**

**ABSTRACT:** We have developed an algorithm that uses analysis of protein dynamics to predict functional sites. The algorithm performs an approximate version of Dynamics Perturbation Analysis (DPA), which can predict ligand-binding sites in protein structures (D. Ming, M.E. Wall, 2006. *J Mol Biol* 358:213). The present algorithm decorates the surface of the protein with test points, and uses approximate calculations of entropies to characterize the degree to which each point perturbs the protein's thermal vibrations. Residues near points that cause a large change in entropy are predicted to reside in functional sites. We used the algorithm to analyze more than 50,000 SCOP domains; and predictions were integrated with residue-conservation statistics obtained from the HSSP database. The analysis was performed using a flexible, distributed software architecture recently developed for this and other computationally intensive tasks.

## **Peter Haug, MD,**

Senior Informaticist for Intermountain Health Care and Professor,  
Department of Biomedical Informatics, University of Utah

### **Data Mining, Clinical Modeling and the Future of Healthcare Computing**

**ABSTRACT:** The evolution of the modern electronic health record (EHR) has created an environment in which large amounts of medical data are collected reflecting the character of disease processes and the response of caregivers to these processes. This data has historically had an episodic character but now is rapidly becoming longitudinal in nature, allowing it to reflect the course of health over time. The presence of this data facilitates both the study of disease processes over time and the development of novel, experience-based approaches to support the delivery of care.

The data available is captured both as structured, readily-computable information and as textual reports, which require approaches based on natural language processing (NLP) paradigms for effective, computer use. In this presentation we will discuss our ability to extract this data and to use it to develop computable models of disease capable of informing care over time. The conjunction of data mining and artificial intelligence techniques holds promise to change the way we deliver, document, and evaluate the clinical care process.



## INVITED KEYNOTE SPEAKERS

### **Eugene Myers, PhD,**

Group Leader, Howard Hughes Medical Institute, Janelia Farms

### **Computer-Assisted Forensic Analysis of Mass Disasters**

**ABSTRACT:** We examine the problem of identifying remains in mass disasters such as the World Trade Center, Waco, and airplane crashes. Typically, the problem is closed or nearly so, in that the individuals that could be involved are known. Depending on the state of the remains, nuclear DNA profiles, typically the 13 CODIS loci used by the FBI, are produced for each sample, and in cases where the remains are significantly degraded, as in the case of severe heat or fire, one may also sequence mitochondrial DNA from the hyper-variable control region. The problem is to determine the individual from whom each sample came from, given the genetic profiles of near relatives and possibly direct evidence from personal effects of the victim.

The talk will elaborate on the nature of the data, develop the necessary background on computing the probability of a pedigree, and formulate the overall goal as a series of algorithmic problems with a preliminary progress report on each.

### **Andrey Ptitsyn, PhD,**

Assistant Professor, Colorado State University

### **Life Works on AC Power**

**ABSTRACT:** Multiple studies indicate that 10–15% of all genes in the hypothalamus and multiple peripheral tissues in mammals oscillate in a daily (circadian) rhythm. In our recently published studies we have applied three alternative algorithmic approaches to identify circadian oscillation in metabolically active peripheral tissues in mice and reported unexpectedly high number of oscillating genes. Our studies also detect no steady non-oscillation fraction of actively expressed genes. This leads to the conclusion that the accepted null-hypothesis in tests for gene expression periodicity is formulated on the unfounded assumption that all genes display a default steady-line expression. We propose a new approach that allows application of Digital Signal Processing (DSP) algorithms separately to each phase class of genes. Combined with Kolmogorov-Smirnov test this method identifies circadian baseline oscillation in almost 100% of all expressed genes. We conclude that such prominence of circadian oscillation in gene expression must be taken into account in all studies related to biological pathways. The importance of oscillation in signal transduction is demonstrated on the example of insulin signaling. This suggests that the loss of synchronization is likely to be among the causes or aggravating factors in metabolic disorders such as obesity and diabetes.

## **Heinrich Roder, DPhil,**

Chief Technology Officer, Biodesix Inc., Colorado

### **Comparative mass spectrometry for clinical applications**

**ABSTRACT:** The direct use of comparative mass spectrometry as an unlabeled probe for the differentiation of disease states, for prognostic stratification of patients according to treatments, and for disease progression monitoring faces special challenges intricately connected with the physics of mass spectrometry of biological molecules. In this presentation I will give an overview of statistical techniques specific to the comparative analysis of mass spectrometry data.

## **David S. Wishart, PhD,**

Department of Computing Science and Department of Biological Science, University of Alberta

### **Metabolomics: The Next Frontier for Bioinformatics?**

**ABSTRACT:** Metabolomics is a newly emerging field of “omics” research concerned with the high-throughput identification and quantification of the small molecule metabolites in the metabolome. Metabolomics is drawing considerable attention these days because it has the potential to substantially improve the speed and accuracy of many clinical tests and diagnoses. Metabolomics shares many of the same computational needs with other, much better established “omics” fields such as genomics, proteomics and transcriptomics. All four “omics” techniques require electronically accessible and searchable databases, all of them require software to handle or process data from their own high-throughput instruments, all of them require laboratory information management systems (LIMS) to manage their data, and all require software tools to predict properties, pathways, relationships or functions. Unfortunately, for metabolomics specialists, relatively few of these essential tools or resources exist. Fortunately, for bioinformaticians, this represents a wonderful opportunity to apply what they have learned from other “omics” endeavors to this newly emerging field. In this presentation I will highlight some of the efforts we, and others are making to bring modern bioinformatics practices to metabolomics – including the development of public metabolomic databases, the development of LIMS and the creation of software to interpret metabolomic data.

---

## ORAL PRESENTATION 1

### Computer Aided Genetic Engineering

Presenter: Dan McShan, University of Colorado School of Medicine

Author(s): Dan McShan

**ABSTRACT:** This talk will outline a vision for Computer Aided Genetic Engineering (CAGE). The vision is that within the next several years it will be possible to quite literally “program” biological systems as we do modern computer systems. With this in mind, CAGE endeavors to apply software engineering practices to genetic engineering. I will outline how concepts like “compiling”, “debugging”, and “version control” relate to the CAGE vision. I will briefly discuss the necessity for a high level programming language for biological systems, focusing on the need for a mixed declarative/imperative approach. Finally, I will introduce the Insitute for Computer Aided Engineering (iCAGE) to further develop these concepts as well as a journal (jCAGE) to collect and disseminate these ideas.

---

## ORAL PRESENTATION 2

### Designing C<sub>2</sub>H<sub>2</sub> Zinc Finger Proteins to Target Specific Genomic Sites

Presenter: Peter Zaback, Iowa State University/Bioinformatics and Computational Biology

Author(s): P. Zaback, J. Sander, F. Fu, D. Voytas, D. Dobbs

**ABSTRACT:** Zinc fingers, the most abundant DNA binding motifs in eukaryotes, offer perhaps the best understood protein-DNA recognition mechanism. Zinc finger proteins (ZFPs) promise to become valuable tools for gene regulation and genome modification because they can be used to target other proteins, including transcriptional activators and nucleases, to virtually any desired location in any genome. Consisting of multiple modular and interchangeable nucleic acid binding domains, C<sub>2</sub>H<sub>2</sub> ZFPs provide a convenient framework for engineering new sequence-specific DNA binding proteins. Using ZF modules characterized by others, we have developed a program, ZiFiT (<http://bindr.gdcb.iastate.edu/ZiFiT>), to identify optimal ZFP binding sites in any gene or chromosomal region of interest. ZiFiT allows users to choose from several different validated ZF module sets. Based on ZiFiT output, users can request reagents from the Zinc Finger Consortium (<http://www.zincfingers.org>), including cloned ZF modules in a framework for simple restriction enzyme-mediated assembly (*Nature Protocols*, in press). In ongoing work, we are using both computational and experimental approaches to directly test the efficacy of the modular approach to constructing novel ZFPs. Our goal is to accurately predict which combinations of ZF modules are most likely to function successfully in vivo.

---

## ORAL PRESENTATION 3

### **Biological Network Inference and the SEBINI Software Environment**

Presenter: Ronald Taylor, Pacific Northwest National Laboratory

Author(s): Ronald Taylor, Anuj Shah, Charles Treatman, and Meridith Blevins

**ABSTRACT:** Reconstruction of regulatory and signaling networks is a critical task in systems biology. High-throughput experiments are now producing mRNA expression data in quantities large enough for researchers to attempt to reconstruct the structure of gene transcription networks based primarily on state correlation measurements. This talk will give a very brief introduction to the concept of network reconstruction, and then will describe the Software Environment for Biological Network Inference (SEBINI), which has been created to provide an interactive environment for the deployment and evaluation of algorithms used to reconstruct the structure of biological regulatory networks. More recently, SEBINI has also been extended to handle inference of undirected protein-protein interaction networks. SEBINI can be used to score and compare network inference methods on artificial networks and simulated gene expression perturbation data. It also allows the analysis within the same framework of experimental high-throughput expression data; hence SEBINI should be useful to software developers wishing to evaluate, compare, refine, or combine inference techniques, and to bioinformaticians analyzing experimental data.

---

## ORAL PRESENTATION 4

### **The BioCyc Collection of Pathway/Genome Databases**

Presenter: Alexander G. Shearer, SRI International

Author(s): Alexander G. Shearer, Ron Caspi, Carol A. Fulcher, Pallavi Kaipa, Peter D. Karp

**ABSTRACT:** The BioCyc collection of Pathway/Genome Databases (PGDBs) provides integrated representations of pathway and genome information for more than 200 organisms, of which most are microbes. Most BioCyc PGDBs were computationally derived from annotated genomes using the MetaCyc database (DB), which describes more than 800 experimentally determined metabolic pathways from more than 700 organisms. MetaCyc is a highly curated, literature-based DB of metabolic pathways and enzymes that provides a quality reference for pathway prediction. It is also an encyclopedic reference source for metabolic engineering and for other studies of metabolism. The computationally generated BioCyc PGDBs include predicted metabolic pathways as well as predicted fillers of holes in those metabolic pathways. BioCyc data are encoded using the Pathway Tools ontology, which facilitates the representation of complex biological knowledge with high fidelity. Pathway Tools provides a variety of visualization and analysis capabilities. Those to be presented here include its new comparative pathway analysis capabilities and its many data access mechanisms. Pathway Tools

also has the ability to automatically generate metabolic map diagrams for each organism in the BioCyc collection. These diagrams can be used for analysis of omics datasets and can be enlarged to produce publication-quality metabolic wall charts.

---

## ORAL PRESENTATION 5

### **Spatio-Temporal Modelling of Biochemical Pathways: Introducing Cell+**

Presenter: John Parkinson, Hospital for Sick Children / University of Toronto

Author(s): Chris Sanford, Matthew Yip and John Parkinson

**ABSTRACT:** High throughput genomic technologies are leading to the generation of vast amounts of data on cellular components, detailing their expression, interactions and organization within biochemical pathways. To understand how these relationships result in a functional pathway, a variety of computational tools have been proposed that attempt to simulate the behavior of the molecular components. In general, these tools tend to neglect the spatial organization of molecules, typically treating the system as a homogenous mixture of components. However, there is increasing evidence that spatial factors such as the co-localization of components have the potential to significantly influence pathway function and efficiency. Here we present Cell++, a novel spatio-temporal modeling platform that performs three-dimensional simulations of biochemical pathways. Combining a cellular automata engine with Brownian dynamics, Cell++ is capable of simulating the bulk properties of large quantities of small molecules (e.g. pyruvate), while simultaneously allowing larger molecules such as enzymes to be treated as more complex entities. Applying Cell++ to the study of metabolic pathways, we demonstrate how the spatial organization of enzymes can alter pathway efficiency and control the production of substrate intermediates, features consistent with the phenomenon of metabolic channeling. Further details of Cell++ can be found at: <http://www.compsysbio.org/CellSim/>.

---

## ORAL PRESENTATION 6

### **Minimal Oscillatory Currents in Neurons**

Presenter: Tom McTavish, UCHSC

Author(s): Tom McTavish

**ABSTRACT:** Many neurons exhibit oscillatory behavior. With a variety of neuron models, we obtained the minimal repeating current necessary to persistently maintain their oscillatory behavior. With these impulses, we show paradoxes of inhibitory neural network systems which can induce such signals and create oscillatory and synchronous behavior in the network. Compared to a network which receives a constant barrage of signals, these inhibitory networks

- 1) have action potentials with greater amplitudes,
- 2) can fire at higher frequency ranges, and
- 3) require less synaptic drive to induce firing.

---

## ORAL PRESENTATION 7

### Assessment of siRNA Inhibition Prediction Algorithms

Presenter: Kevin Sullivan, Dharmacon, Inc

Author(s): Kevin Sullivan

**ABSTRACT:** Accurate prediction of the inhibition potential for synthetic small interfering RNAs (siRNA) is an important step in the use of RNA interference (RNAi). Here we use a large data set published by Huesken et al. to investigate the application of Neural Networks to predict siRNA inhibition. Through this work we were able to reproduce results by Huesken et al., and identify well-performing topology and feature extraction methods. We use a descriptive and robust performance indicator to compare these results to other prediction algorithms. Our results suggest that linear regression techniques may be equal in performance to more complex classification, while providing explicit feature weighting and insight to the mechanism of RNAi.

---

## ORAL PRESENTATION 8

### The Balance Between Activating and Inhibiting Connections Controls the Dynamics of Biological Networks

Presenter: Daniel McDonald, University of Colorado at Boulder

Author(s): Daniel McDonald, Meredith Betterton, Laura Waterbury, Rob Knight

**ABSTRACT:** Many recent studies of biological networks, including transcription regulation and protein-protein interaction networks, have analyzed network ‘wiring diagrams’ (topological features). Such work has suggested that specific types of network wiring, such as small-world network structure, may increase network robustness. However, knowledge of topological features does not allow one to predict dynamical behavior. Here we report that the balance between activating and inhibiting connections is crucial in determining whether a network reaches steady state or oscillates. We use a mathematical model that captures the essential behavior of a network of interacting genes or proteins to study randomly sampled networks, networks subjected to selection for specific properties, and examples of real biological networks. Remarkably, in all cases the ratio of activating to inhibiting connections controls whether the network reaches steady state or oscillates. Indeed, real biological networks which are expected to reach steady state (signaling or developmental switches) and those which are expected to oscillate (circadian oscillators) have a fraction of activating connections consistent with our model. The fraction of connections which are activating is a previously unrecognized parameter which plays a major role in determining network dynamical behavior. Therefore, the fraction of activating connections may be a control parameter that cells use to predispose a network to oscillate or reach steady state.

---

## ORAL PRESENTATION 9

### **Beyond Ontologies: The WayCraft Scientific Knowledge Integration Framework**

Presenter: Frank D. Russo, WayCraft Biosoftware

Author(s): Frank D Russo, Steven T Connolly, Ranga Chandra Gudavida, Angela Qu, Anil Jegga, Bruce J Aronow

**ABSTRACT:** Ontologies provide an excellent basis for the description and organization of concepts and terminologies, and have proven particularly useful for categorizing the results of large-scale data acquisition. However, ontologies generally lack good representation of complex relationships, causality, or dynamic behavior, properties that will be crucial for next-generation applications that model complex biological systems, and predict the impact of perturbations on their behavior. To approach this, WayCraft has developed the Scientific Knowledge Integration Framework (SKIF) as a platform to construct computationally rigorous models of biological systems, and to develop applications based on those models. At its core, SKIF is a model of entities, processes, polymorphisms, and states, which is used to represent biological systems in fundamentally the same way that biologists think about them. The framework itself makes extensive use of open-source systems such as Hibernate and Eclipse. The first application to be built on this platform consolidates knowledge around one disease or disease family, including a model of the underlying disease process at both a molecular and clinical level.

---

## ORAL PRESENTATION 10

### **An Experiment in Mining Gene-Disease Relationships from Biomedical Literature**

Presenter: Graciela Gonzalez, Arizona State University — School of Computing and Informatics, Dept of Biomedical Informatics

Author(s): Graciela Gonzalez, Juan C. Uribe, Luis Tari, Colleen Brophy, Chitta Baral

**ABSTRACT:** The promises of the post-genome era disease-related advances have yet to be realized, with opportunities for discovery hiding in millions of biomedical papers. Public databases have data extracted from the literature by experts, but their coverage is limited and lags behind recent discoveries. We present a computational method that combines data extracted from the literature with data from curated sources to uncover gene-disease relationships not directly stated or missed by the initial mining. An set of genes is obtained from gene-disease relationships extracted from PubMed abstracts using NLP. Interactions involving the corresponding proteins are similarly extracted and integrated with interactions from curated databases. Each protein is then ranked combining two scores: one on the strength of its relationship with the initial set and another on the importance of the gene in maintaining the connectivity of the network. We applied the method to

atherosclerosis to assess its effectiveness. Ranked proteins reached 100% accuracy for the top 20 and 80% for the top 90 (duplicates ignored). Thus, though the initial gene set and interactions are subject to the impreciseness of automatic extraction, their use for further hypothesis generation is valuable given adequate computational analysis to boost the accuracy of automatic extraction.

---

**ORAL PRESENTATION 11**  
**SESSION CANCELLED**

---

**ORAL PRESENTATION 12**

**Bioinformatics as a Tool to Predict Gene Function in Williams-Beuren Syndrome**

Presenter: Hannah Tipney, University of Colorado at Denver Health Sciences Center

Author(s): Hannah Tipney, Andy Brass, May Tassabehji

**ABSTRACT:** Williams-Beuren Syndrome (WBS) is a sporadic microdeletion disorder, effecting ~1/20,000 live births and presenting as a complex multisystem phenotype. Over 24 genes reside in the WBS deletion, however only one (ELN) is associated with a facet of the WBS phenotype (SVAS). The aim of this research was to increase understanding of the WBS genotype-phenotype relationship. As part of a WBS region mapping effort, 'new' genes residing in the deletion were identified and characterised through standard bioinformatics approaches. Of particular interest was *GTF2IRD2*, a member of the novel I-Repeat domain containing TFII-I protein family, which has been formed by the fusion of a TFII-I-like gene and a CHARLIE8 transposase-like element. I-Repeat sequences across all species were collated and through alignment and phylogenetic analysis it was concluded that the TFII-I family of proteins have evolved from a single ancestral I-Repeat via numerous duplications, deletions and rearrangements. This work also highlighted problems associated with the complex nature of publicly available biological data. A successful collaboration with the myGrid project focused on the problem of monitoring incomplete and incorrectly assembled genomic contigs, a vital yet time consuming task. This produced the first application of emerging grid technology to a real world, biologically significant problems.

---

## ORAL PRESENTATION 13

### **Unified Maximum Separability Analysis for Molecular Profile Analysis and Biomarker Discovery**

Presenter: Qiuyan Huo, Johns Hopkins University School of Medicine

Author(s): Qiuyan Huo, Zhen Zhang

**ABSTRACT:** Unified Maximum Separability Analysis (UMSA) modifies support vector machine (SVM) learning algorithm to allow training samples to be weighted in the soft-margin SVM learning algorithm. The weights are typically calculated according to the relative agreement of individual samples with a statistical classifier constructed using the same data. An adjustable parameter seamlessly unifies the results from two traditionally different approaches (density estimation vs. empirical risk minimization). Results from real and simulated data as well as theoretical proofs have suggested UMSA's superiority for small sample size learning problems for which the effective extraction and use of information by a learning algorithm is often more important than its asymptotic performance. UMSA has been successfully used for molecular profile data analysis and for biomarker discovery. We have implemented the UMSA algorithm in multiple analysis modules for web-based deployment with a user interface designed to fit tightly into the workflow of clinical genomics and proteomics research for biomarker discovery. We will present examples of real data analysis from our own work. In addition, for comparison with commonly used methods such as SAM, we analyzed a public set of molecular profiling data of myeloid leukemia cell lines to show the relative advantage of UMSA-based tools.

---

## ORAL PRESENTATION 14

### **Data Integration as a Foundation for Semantic Integration**

Presenter: Daniel J. McGoldrick, University of Colorado, Department of Pharmacology

Author(s): Daniel J McGoldrick, Lawrence Hunter

**ABSTRACT:** The post-genomic era has created an explosion of valuable gene-centered data. Bioinformaticists are now moving to classify in standardized ways “what the genes mean” using ontologies, the Semantic Web and OWL. In this pursuit, text corpora and natural language processing approaches are well complimented by machine driven “connect the dot” approaches using public database identifiers. Descriptron is presented here as a framework for establishing cross references among gene identifiers from public databases as well linkages to ontology terms. Argued is how the generation of ontology connections is more efficient, more reliable, and more complete using a data integration approach. Covering some 35 species, over 7 million biological facts and 5 million maps of inter-connected facts, we show how Descriptron can reduce the limiting size of the triplet universe and establish a basis for synthesis and normalization, in mapping meaning to gene

identifiers. Provenance and homonymy — two major data integration challenges are defined and addressed with tables and examples here. Descriptron then as a data integration tool is improving our understanding of genes, their cross references and ultimately their meaning.

---

**ORAL PRESENTATION 15****Dissecting Genetics of Host-Pathogen Interactions**

Presenter: Peter T. Hraber, Theoretical Biology & Biophysics, LANL

Author(s): Peter T. Hraber

**ABSTRACT:** Methods in computational and molecular biology generally study the genetic processes of one model organism at a time. In reality, life is anything but a pure culture. Pathosystems extend the single-species paradigm across multiple replicating organisms and present new problems, because infected cells house both host and pathogen. Computational solutions to challenges in the analysis of pathosystem data will be discussed.

---

**ORAL PRESENTATION 16****Phylogenetic Profiling via Partial Genomes: Applications to the Apicomplexa**

Presenter: James Wasmuth, Hospital for Sick Children

Author(s): Jennifer Daub, Matthew Fagnani, Jose Peregrin-Alvarez, Chris Sanford, James Wasmuth, John Parkinson

**ABSTRACT:** The availability of 450+ expressed sequence tag (EST) datasets complements the 42 completed genome sequences enabling a systematic and comprehensive analysis of the Eukaryotes. We have created PartigeneDB to process, cluster and annotate these ESTs into ‘partial genomes’. Protein families can be generated from these datasets and the species distribution of each family used to create phylogenetic profiles. These reveal the spectrum of diversity, from highly conserved proteins, to those that are species-specific. Applying these methods to the study of the Apicomplexa, a medically important phylum that includes parasites such as Plasmodium (responsible for the death of one million children per year) and Toxoplasma (a major food-bourne pathogen). We compared 87,000 protein sequences from the Apicomplexa with both complete and partial genomes. Here, we report that 42% of proteins were specific to the Apicomplexa, with 2,100 conserved across four or more species, suggesting possible targets for broad spectrum therapeutics. Using more sensitive domain analyses, we confirm that 70% of these families are only found in the Apicomplexa. Other, more widely distributed, protein families, are shown to have apicomplexan-specific expansions. These findings highlight the advantage of including partial genomes in comparative genomic studies.

---

## ORAL PRESENTATION 17

### **Genome-Wide Co-Expression Based Prediction of Differential Expressions**

Presenter: Yinglei Lai, The George Washington University, Department of Statistics  
Author(s): Yinglei Lai

**ABSTRACT:** Microarrays have been widely used for various disease studies to detect novel disease related genes. They enable us to study differential gene expressions at a genomic level. They also provide us with informative genome-wide co-expressions. Although numerous methods have been proposed for identifying differentially expressed genes, genome-wide co-expressions have not been well considered for this issue. Incorporating genome-wide co-expression information in the differential expression analysis may improve the detection of disease related genes. In this study, we propose a statistical method for predicting differential expressions through the local linear regression between differential expression tests and expression correlation measures. A mixture normal quantile based method is used to transform data. We use the gene-specific permutation procedure to evaluate significance of a prediction. Two published microarray data sets were analyzed for applications. For the data set collected for a prostate cancer study, the proposed method identified numerous genes with weakly differential expressions. These genes have been shown in literature to be associated with the disease. For the other data set collected for a type 2 diabetes study, no significant genes could be identified by the traditional methods. However, the proposed method identified numerous genes with significantly low false discovery rates.

---

## ORAL PRESENTATION 18

### **A Semi-Manual Method Using the ANCOVA Framework to Identify Expression Profiles in Time-Course Microarray Experiments**

Presenter: Tzu Lip Phang, University of Colorado Health Sciences Center, Department of Medicine  
Author(s): Tzu Lip Phang, Katherina Kechris

**ABSTRACT:** The decline of microarray prices has motivated more complicated multi-treatment time-course experimental designs with proper sample replication at each time point. In one particular design, researchers may be interested in both the differential expression with respect to time and treatment, as well as their interaction effects. At first glance, the ANCOVA model seems like an ideal solution for attacking this analysis (Park *et al*, 2003, Conesa *et al*, 2006). Careful

examination, however, revealed that the implementation by these authors faces serious statistical limitations: i) The algorithm only compares treatment factors with respect to a baseline, both for evaluating the factor, and factor and time interaction. In their framework, performing a non-baseline comparison is problematic. ii) Only one factor is allowed, which poses a limitation for many experimental designs that have multiple treatment factors, in addition to measurements in time. iii) There is no convenient tool for biologists to select their expression profiles of interest based on the results from the modeling. In this work, we addressed these shortcomings, and introduced a new workflow for analyzing multi-treatment microarray time-course data analysis using the Tukey Honest Significant Differences (TukeyHSD) and Johnson-Neyman procedures.

---

## ORAL PRESENTATION 19

### **ARB — A Comprehensive Phylogenetic Sequence Analysis and Probe Design Software Environment**

Presenter: Harald Meier, Technische Universität Muenchen, Informatik

Author(s): W. Ludwig, R. Westram, Y. Kumar, H. Meier

**ABSTRACT:** Comparative phylogenetic sequence analysis is widely used for illuminating the relationships of organisms taking their evolutionary history into consideration. For microorganisms such as bacteria even the current taxonomy bases mainly on phylogenetic trees inferred from sequences of conserved universal genes. Methods based on additive sequence treeing or on in silico designed phylogenetic oligonucleotide probes on diagnostic microarrays are widely applied for the identification of pathogens as well as in terms of profiling of microbial communities for structure/function analyses in microbial ecology. ARB is an integrative bioinformatic software package which facilitates the performance of such complex analyses significantly by joining all required tools and functions under a common graphical user interface. A full central database system allows the creation and maintenance of large genome, gene or protein sequence databases. The secondary and tertiary structure alignment and visualization components as well as integrated high performance phylogenetic analysis programs such as RAXML, allow a fast and solid phylogenetic clustering of even large datasets. Interactive tree visualization and client/server based search infrastructure data structures as well as algorithms for proper phylogenetic probe design and evaluation complete the functionality required for comprehensive in silico molecular microbial diagnostic analyses.

---

## ORAL PRESENTATION 20

### Does the Region of rRNA Sequenced Affect Conclusions from Microbial Community Analysis

Presenter: Zongzhi Liu, University of Colorado

Author(s): Zongzhi Liu, Catherine Lozupone, Micah Hamady, Rob Knight

**ABSTRACT:** 16S rRNA sequence data from uncultured environmental samples are accumulating rapidly. Different research groups use different sequencing primers, thus obtaining rRNA sequences starting at different positions and with different lengths. We tested whether this variety of primer choices affects the results of sequence-based community analysis. We focus on UniFrac, a method for comparing microbial communities we recently developed. We performed three tests of primer effects. First, starting with near-full-length sequences in a recent study of microbial communities in obese mice, we tested all possible combinations of standard primers, used parsimony insertion to add these clipped sequences to an existing microbial tree, and tested how much of the clustering structure in the data was recovered (using the clusters from full-length sequences as a reference standard). Second, we repeated this study for short sequences (100- and 200-base reads) resembling those produced by pyrosequencing. Third, we used sequences from over 200 globally dispersed environments to test whether studies that used similar sequencing primers clustered together mistakenly. The results show that sequencing effort is best focused on gathering more short sequences than fewer longer ones, and that UniFrac is surprisingly robust to variations in the region sequenced.

---

## ORAL PRESENTATION 21

### Tight Clusters F Proteins

Presenter: Roman L. Tatusov, NCBI NLM NIH

Author(s): Boris Kiryutin

**ABSTRACT:** Tight clusters of proteins. Boris Kiryutin, Roman L Tatusov The task of gathering the orthologous proteins from completely sequenced genomes raises numerous problems due to complicated nature of evolution. The significant fraction of clusters as long as the essence of the majority of clusters can be determined automatically. The proposed procedure groups homologous proteins efficiently and robustly. The cluster is defined as a group of proteins more close to members of the group than to non-members. This notion is similar to complete linkage clustering with variable cut-off value. Thus, the constructed clusters contain proteins from the same subtree regardless of the tree constructing procedure. The strict criteria allows for fast algorithmic implementation. Surprisingly, about 40% of COGs are identical to tight clusters while the majority (68%) of COGs have 80% of the proteins covered. The stringent criteria appear to be most useful while constructing clusters limited to close organisms from a simple taxon such as order

or family. The proteins are more similar and the number of abnormal evolutionary events is limited. The multiple alignments constructed for each cluster facilitate tracing the recent evolution of close genomes. The inspection of tight clusters in COGs is viewable: <http://www.ncbi.nlm.nih.gov/COG/grace/budiew.cgi>

---

ORAL PRESENTATION 22

**A Bayesian Network Model of Stromatolite Formation**

Presenter: Jack K. Horner, Science Applications International Corporation

Author(s): Jack K. Horner

**ABSTRACT:** Stromatolites are attached, lithified sedimentary growth structures, accretionary away from a point or limited surface of initiation. Whether stromatolites have a biotic origin is vigorously debated[1]. If biotic, the oldest (~3.5 billion years BP) were created by some of the first terrestrial life forms. The outcome of the debate is thus fundamental to our understanding of the development of early life on Earth and, potentially, elsewhere. Because no single piece of evidence at present could decide the issue, the debate depends significantly on how to interpret the evidence “as a whole”. Here I describe requirements on, and the design and implementation of, a Bayesian network[2] model of the stromatolite-origin dispute that provides a unified realization of the relation of the hypothesis to each of the individual types of evidence and to the evidence “as a whole.”

Keywords: stromatolite, Bayesian network — [1] A. C. Allwood *et al.* Stromatolite reef from the Early Archaean era of Australia. *Nature* 441 (8 June 2006). pp. 714–718. [2] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Second Edition. Morgan Kaufmann. 1988.

---

ORAL PRESENTATION 23

**Genomic Supertrees of Life**

Presenter: Davide Pisani, The National University of Ireland Maynooth

Author(s): Davide Pisani, James A. Cotton, James O. McInerney

**ABSTRACT:** How to recover the tree (or the network) of life is still debated. One possibility would be generating gene trees for each protein family in the genome of each considered species, and use either consensus trees or networks to obtain eloquent synopses of these trees. Because the gene-trees will be partially overlapping, supertrees (or supernetworks) should be used. Here, I will present results obtained from analyses of up to 169 complete genomes from the three domains of life, thus illustrating how supertrees could be implemented to (1) recover trees based on complete genomic information, and (2) identify multiple phylogenetic signals in complete genomes.

---

## ORAL PRESENTATION 24

### **Modeling Evolution of Gene and Protein Networks**

Presenter: Todd A. Gibson, University of Colorado Health Sciences Center

Author(s): Todd A. Gibson, Debra S. Golderg

**ABSTRACT:** Human physiology is the product of evolutionary mechanisms common to all life forms. Evolutionary models of other organisms are invaluable analogs to human evolution and could eventually help researchers intervene in the evolutionary development of viruses and other fast-evolving pathogens. We have modeled the evolution of *Arabidopsis thaliana*'s protein interaction network from a putative ancestor to the modern organism and are currently analyzing the results and tuning the model. Subsequently, the model will be extended to *Saccharomyces cerevisiae*. Our approach differs significantly from current models which are either decontextualized or static. Decontextualized models evolve protein interaction networks as graph-theoretical constructs; genes duplicate and diverge as homogenous entities unrelated to any specific genes from actual organisms. Approaches which associate protein interaction networks with modern organisms derive meaningful inferences by comparing static network structures between different species. Our model introduces dynamic evolution to the same organism-specific networks. The model architecture will be presented including construction of the ancestral network, dating gene and genome duplication events, evolving the ancestral network to the modern organism, theoretical considerations, and issues around available data and inference tasks.

---

## ORAL PRESENTATION 25

### **The Unique-eome: What Makes species Different?**

Presenter: Martin Gollery, University of Nevada, Reno

Author(s): Martin Gollery, John Cushman, Jeff Harper, Taliah Mittler, Ron Mittler

**ABSTRACT:** Proteins with obscure features (POFs), which lack currently defined motifs or domains, represent between 18% and 38% of a typical eukaryotic proteome. To evaluate the contribution of this class of proteins to the diversity of eukaryotes, we performed a comparative analysis of the predicted proteomes derived from 10 different sequenced genomes, including budding and fission yeast, worm, fly, mosquito, *Arabidopsis*, rice, mouse, rat, and human.

## ORAL PRESENTATION 26

**Monte Carlo EM Algorithm for Sequence Motif-finding**

Presenter: Chengpeng Bi, Children's Mercy Hospitals

Author(s): Chengpeng Bi

**ABSTRACT:** The EM-type motif-finding algorithm is one of the most popular de novo motif discovery methods. However, as pointed out in literature, EM algorithms largely depend on its initialization and can be easily trapped in local optima. In this presentation a Monte Carlo EM-type algorithm is proposed and implemented to overcome the drawbacks inherent in EM motif-finding algorithms. The new motif-finder starts from a set of random seeds, then it performs simulation in the entire multiple local alignment space by drawing motif sites according to the conditional probability distributions and sequence data, and progressively approximates the optimal alignment solution whereby the local optimal estimation of motif model is achieved. The new algorithm is compared with other popular motif algorithms using simulated, annotated prokaryotic and eukaryotic motif sequences. Results showed that the new motif-finder performs equal to or better than other algorithms.

## ORAL PRESENTATION 27

**Biomarker Discovery in Genomic Data with Partial Clinical Annotation**

Presenter: Cole Harris, Exagen Diagnostics, Inc.

Author(s): Cole Harris, Noushin Ghaffari

**ABSTRACT:** Standard approaches to biomarker discovery in microarray gene expression data require that data be collected from clinically annotated samples. The availability of such human samples with the requisite clinical annotation is generally limited. However, samples from similar patient populations, but without annotation, are often available. We propose a new technique for mining of gene expression datasets where the clinical information may not be available for most datasets. Our method supplements standard supervised learning with clustering of data lacking clinical annotation to estimate the predictive power of gene subsets. As a demonstration, we have applied this technique to the well-known MIT ALL/AML dataset. In this dataset all samples are known to be from patients with either ALL or AML. However in our test, the disease subtype was blinded for half of the training samples. We employed a genetic algorithm search, in combination with an objective function consisting of terms for training classification accuracy and cluster separation, to uncover diagnostic subsets of genes. Classifiers built from these combinations were then evaluated on the test samples. We find that the addition of the unannotated data in training significantly improves test set classification accuracy. On average, test set classification accuracy increased by 9%.

---

ORAL PRESENTATION 28

**Systematic Gene Selection for Dynamic Gene-Network Refinement**

Presenter: Christian V. Forst, Los Alamos National Laboratory

Author(s): Nicole Radde, Jutta Gebert, Christian V. Forst

**ABSTRACT:** A quantitative description of interactions between cell components is a major challenge in computational Biology. ODEs are used for this purpose, by providing detailed insight into the dynamic behavior. Typically, the number of time points of experimental time series is usually too small to estimate parameters of a whole gene regulatory network model based on ODEs, such that one needs to focus on subnetworks consisting of only a few components. For most approaches, the set of components of the subsystem is given in advance and only the structure has to be estimated. However, the set of components that influence the system significantly are not always known in advance, making a method desirable that determines both, the components and the parameters. We have developed a method that uses gene expression data as well as interaction data between cell components to define a set of genes that we use for our modeling. In a subsequent step, we estimate the parameters of our model of piecewise linear differential equations and evaluate the results simulating the behavior of the system with our model. We have applied our method to the DNA repair system of *M.tuberculosis*. Our analysis predicts a new gene playing an important role in this system.

---

ORAL PRESENTATION 29

**Using an Anatomy Ontology to Predict the Spread of Tumor Cells to Regional Metastatic Sites**

Presenter: Ira Kalet, University of Washington

Author(s): Ira Kalet

**ABSTRACT:** Advances in radiation treatment machinery and dose modeling software now enable the planning and delivery of radiation to very precise volumes within any particular patient's body. The usefulness of this capability critically depends on knowing accurately where the tumor cells are, not only for the primary tumor, but for local and regional spread, mainly to nearby lymph nodes. It is impossible to visualize these involved lymph nodes on any present day imaging systems, so a theoretical model to predict which nodes are likely to be involved would be invaluable. The lymphatics are the primary pathways of spread for many tumor sites. We have constructed a theoretical model that can predict the extent of involved nodes from three ingredients: The Foundational Model of Anatomy (FMA), a rich and detailed ontology of canonical anatomy of the human body, including the lymphatic system, a simple path search algorithm for querying the FMA to determine drainage patterns for any particular tumor site, and a Markov model that describes the transition probabilities for propagation of tumor cells from location to location

in the lymphatic system. With some reasonable guesses at uniform transition probabilities, the model predictions agree reasonably well with reported surgical data for tumors in the head and neck. This simple idea is yet another example of how logical pathway models can other formal ontologies can be leveraged as the basis for building biological theories.

---

**ORAL PRESENTATION 30**

**Analyzing Time Course Microarray Data With Temporal Uncertainty**

Presenter: Stephen Billups, University of Colorado at Denver and Health Sciences Center, Dept. of Mathematical Sciences

Author(s): Stephen Billups

**ABSTRACT:** When conducting time-course microarray experiments, one is interested in measuring time relative to some biological process of interest. Unfortunately, it is often difficult to measure this “biological” time accurately. Such temporal uncertainty negatively impacts both the accuracy and power of data analysis. In this talk, I will describe a method for handling temporal uncertainty based on the premise that microarray data itself can be used to refine our estimates of “biological” time.

---

**ORAL PRESENTATION 31**

**Energy Landscape Calculations for a DNA-Rotaxane using Milestoning Technique**

Presenter: Shahid Qamar, Arizona State University

Author(s): Shahid Qamar

**ABSTRACT:** We are working on a technique to sequence the DNA with an atomic force microscope. Motivated by the experiment, we used an efficient technique to compute free energies of a DNA rotaxane molecule composed of a single strand DNA and a cyclodextrin molecule. Quantitative free energy computation involves milestoning technique with Arrhenius rate equation. The algorithm used computes the time scales of complex processes following the predetermined milestones along a reaction coordinate. A Markovian hopping mechanism was used. We performed the large scale molecular dynamics simulations at micro second level to compute the rare event kinetics which involves large scale distributed computational resources. We performed the molecular dynamics simulations for a DNA rotaxane in the absence of external force to compute the free energy differences among them. All the simulations were performed in aqueous solvent. The theoretical estimation of free energies qualitatively agrees with the experimental data obtained for nano pore DNA sequencing with an atomic force microscope. Initial results show the thermal fluctuations are dominant and the free energy differences between purine and pyrimidine is of the order of  $k_B T$  so the modification of DNA rotaxane is required to suppress the thermal fluctuations.

---

## ORAL PRESENTATION 32

### The SINE of the Opossum

Presenter: David D. Pollock, University of Colorado

Author(s): Wanjun Gu, David A. Ray, Jerilyn A. Walker, Erin Barnes, Andrew J. Gentles, Paul B. Samollow, Jerzy Jurka, Mark A. Batzer, David D. Pollock

**ABSTRACT:** We analyzed SINEs and other repetitive elements in the newly sequenced *Monodelphis domestica* (opossum) genome. Based on a recently developed heuristic algorithm, a high copy number repeat family fitting the SINE<sub>I</sub> profile was identified and the most probable subset of recently duplicated members predicted. The family appears to be derived from a tRNA gene and is similar to the Mar SINEs. Over one hundred predicted loci were amplified from a panel of 43 *Monodelphis* individuals from five diverse geographic locations. The family has expanded recently enough that many members are polymorphic among these *Monodelphis* populations, and three loci were polymorphic only in the source population. Genetic distance among populations reflected geographic distance. High SINE density is associated with high GC content in the genome, and also with altered mutational patterns and rates. This appears to largely explain a mutation rate increase in genes on the X chromosome. Young fixed SINEs and polymorphic SINEs are not associated with biased adjacent GC content, however, and so as with Alus in humans, SINE<sub>I</sub> appears to be preferentially removed from low GC content regions over evolutionary time. A 9 MY gap in SINE<sub>I</sub> diversification may have been caused by high rates of SINE removal.

---

## ORAL PRESENTATION 33

### Strong Negative and Positive Selection Can Obscure Ancestral Signal in Phylogenetic Analysis

Presenter: Alexander Tchurbanov, University of Wyoming

Author(s): Alexander Tchurbanov, Katherine Harris, Shruti Rastogi, David Liberles

**ABSTRACT:** Paralogs and less frequently orthologs contain increased ratio of non-synonymous ( $K_a$ ) to synonymous ( $K_s$ ) substitutions. There are several approaches to reconstruct phylogenetic relationships based on gene sequences such as maximum likelihood, maximum parsimony and neighbor joining algorithm. We investigate phylogenetic signal contained in sequences used for phylogenetic tree reconstruction and conclude that multiple sequence alignment fragments with  $K_a/K_s$  ratio slightly less than one produce the best phylogenetic reconstruction. Fragments under strong negative and strong positive selection obscure the phylogenetic signal and are not handled well by the current evolutionary models based on random (neutral) processes.

## ORAL PRESENTATION 34

**Cyberenvironment for Computational Bioscience**

Presenter: Thanh N. Truong, Department of Chemistry, University of Utah

Author(s): Thanh N. Truong

**ABSTRACT:** The need for efficient and secure access to information and resources on remote computing systems and data sources is critical for many areas of computational science including bioinformatics. The demonstration will illustrate how the cyberenvironment called Computational Science and Engineering Online (CSE-Online, <http://cse-online.net>) accessing these remote resources including the computing grid. CSE-Online is based on a unique client-server software architecture that will introduce a paradigm shift in that data, resources, and applications on distributed remote servers are 'delivered' to the user desktop environment rather than the users having to go to these servers to access them. CSE-Online was recently released with more than 30 tools for research and education in Chemistry, Chemical Kinetics, Bio-Molecular Modeling and Simulations, and Computer-Aided Drug Design. CSE-Online also provides a software framework for vertical integration of tools in different domains to facilitate multidisciplinary research. Focus of the presentation will be on the applicability of the cyberenvironment for research and education in bioscience.

## ORAL PRESENTATION 35

**A Computational Method to Identify RNA Binding Sites in Proteins**

Presenter: Jeff Sander, Iowa State University

Author(s): J. Sander, M. Terribilini, J.H. Lee, R. Jernigan, V. Honavar, D. Dobbs

**ABSTRACT:** RNA-protein interactions are vitally important in a wide range of biological processes. We are developing machine learning approaches for predicting which amino acids of an RNA-binding protein mediate RNA-protein interactions, using either sequence alone or a combination of sequence and structure-derived information as input. Interfaces from RNA-protein complexes in the PDB were extracted to generate a non-redundant set of 147 RNA-binding protein chains. Using this dataset, we trained and tested several individual and ensemble classifiers. Our best ensemble classifier identifies interface residues with 87% overall accuracy, correlation coefficient of 0.37, specificity of 57% and sensitivity of 33%. In recent work (Terribilini *et al* 2006), we tested the performance of sequence-based classifiers on clinically important proteins for which experimental structure information has been elusive, including the HIV-1 Rev and human telomerase reverse transcriptase (hTERT) proteins. In both cases, predicted RNA-binding residues are in good agreement with experimental results. Ongoing work is directed at exploiting new experimental results and additional types of information to improve classification performance. The ability to reliably predict which residues of a protein contribute to RNA binding — even in the absence of structural information — could potentially contribute to new therapies for both genetic and infectious diseases.

---

## ORAL PRESENTATION 36

### **Information Retrieval, Question-Answering, Machine Learning, and Concept Recognition in TREC Genomics 2006**

Presenter: J. Gregory Caporaso, University of Colorado Health Sciences Center, Dept. of Biochemistry and Molecular Genetics

Author(s): J. Gregory Caporaso, William A. Baumgartner, Jr., Hyunmin Kim, Zhiyong Lu, Helen L. Johnson, Olga Medvedeva, Anna Lindemann, Lynne Fox, Elizabeth K. White, K. Bretonnel Cohen, and Lawrence Hunter

**ABSTRACT:** TREC Genomics 2006 presented a challenge to automatically identify document passages that answer questions on twenty-seven topics from a collection of 162,259 full-text biomedical journal articles. Questions were derived from actual information needs of biomedical researchers, and performance was based on human evaluation of the retrieved passages. The Center for Computational Pharmacology approached this task by generating a candidate result set which was subsequently expanded (to improve recall) and then pruned (to improve precision). First, we created a Lemur search index and converted questions into term-expanded queries. Pseudo-relevance feedback was employed to expand search results. Next, from a pool of zone-filtered documents, we further expanded the result set using an LSA-like approach. In the final, false-positive eliminating step, we used naïve Bayesian classifiers trained on human-labeled data with features including bigrams, normalized semantic concepts, and OpenDMAP-style pattern matches. Three separate experiments allow us to compare the utility of the different techniques in this task.

---

## ORAL PRESENTATION 37

### **FluKB: an Integrated Knowledge Base for Influenza Viruses**

Presenter: Guoqing Lu, University of Nebraska at Omaha

Author(s): Guoqing Lu, Kashi R Buyyani, Thaine W Rowley, Ruben Donis, Zhengxin Chen

**ABSTRACT:** Influenza viruses present formidable scientific and public health challenges. Meanwhile, these challenges have brought bioinformatics enormous opportunities in the development of biological databases, computational algorithms, and analysis tools. Several databases such as the NCBI Influenza Virus Resource and the LANL Influenza Sequence Database have been created to store genomic sequence data. In order to have a better understanding of how influenza viruses evolve and cause infectious diseases, there is a critical need for mining knowledge embedded within a variety of data (e.g., genetic data, epidemiological data, vaccine information, and literature) and creating an integrated database to store not only data but also knowledge.

We present here a knowledge base for influenza viruses, named as FluKB, in its design and implementation and describe a set of web tools developed for retrieving data, generating reports, and displaying results of data mining. FluKB as a consolidated database will facilitate both basic research as well as applied research, providing a better understanding of evolutionary mechanisms of influenza viruses and new knowledge for the development of vaccines in order to prevent and control pandemic or epidemic influenza.

---

## ORAL PRESENTATION 38

### Improved Algorithms for Reaction Mapping

Presenter: John D. Crabtree, Colorado School of Mines

Author(s): John D. Crabtree, Dinesh P. Mehta, J. Thomas McKinnon, Anthony M. Dean

**ABSTRACT:** This paper describes algorithms for mapping the migration of atoms in a chemical reaction, a graph-theoretic problem that has numerous applications in biology and in the chemical industry. Unlike recent approaches that are confined to certain types of reactions and do not specify optimality criteria, this paper formulates the problem in general terms and seeks to identify mappings that minimize the number of bonds broken or formed during the reaction. A family of three algorithms (FBF) that guarantees optimal solutions, but may take exponential time in the worst case, is presented. Two of these take polynomial time if the number of bonds broken/formed is a constant. Experiments confirm that the best of the FBF methods is fast in practice on most reactions. A second algorithm (CSL), which takes polynomial time in the worst case and finds optimal solutions 84% to 94% of the time, is also presented. We present the results of applying our algorithms to several databases including the Kyoto Encyclopedia of Genes and Genomes chemical database which contains over 5,000 biomolecular reactions.

---

## ORAL PRESENTATION 39

### Biased Support Vector Machine and Kernel Methods for Tumor Classification

Presenter: Srinivas Mukkamala, New Mexico Tech

Author(s): Andrew H Sung, Krishna Yendrapalli, Ram Basnet

**ABSTRACT:** This paper describes results concerning the robustness and generalization capabilities of kernel methods in classification accuracy on Leukemia, Lymphoma and Prostate cancer datasets of broad institute and Colon cancer dataset from Princeton gene expression project. We use traditional support vector machines (SVM), biased support vector machine (BSVM) and leave-one-out model selection for support vector machines (looms) for model selection. We also evaluate the

impact of kernel type and parameter values on the accuracy of a support vector machine (SVM) performing tumor classification. Through a variety of comparative experiments, it is found that SVM performs the best. We show that classification accuracy varies with the kernel type and the parameter values; thus, with appropriately chosen parameter values, SVMs have the potential of assisting in diagnosis of the malignancy of a tumor with higher accuracy and lower rates of false alarms.

---

## ORAL PRESENTATION 40

### Identification of HTH Motifs from Amino Acid Sequence

Presenter: Changhui Yan, Utah State University

Author(s): Changhui Yan, Jing Hu

**ABSTRACT:** We have developed a Hidden Markov model method (Referred as HMM\_AA\_SS) that combine amino acid sequence and predicted secondary structure for predicting HTH motifs from amino acid sequence. The results show that the method can identify HTH motifs that share only low sequence similarities (<25%) with the HTH motifs available in the training set. To reduce the complexity of protein sequence, we tried three reduced alphabets of various sizes to encode protein sequences. The results show that using reduced alphabets results in a big improvement in identifying HTH motifs. We compared HMM\_AA\_SS with FFAS, a profile-profile comparison method that has been shown successful in detecting remote homology. The predictions of the two methods have a big overlap. Besides that, HMM\_AA\_SS identifies 24 HTH motifs that are missed by FFAS, and FFAS identifies 71 missed by HMM\_AA\_SS. This indicates that HMM\_AA\_SS can provide helpful information in identifying HTH motifs. It also suggests the possibility of achieving better performance by combining the two methods. We applied HMM\_AA\_SS to identify HTH motifs in a small bacterial genome. The method predicted 6 putative HTH motifs: 4 of them are supported by evidences obtained from previous publications and 2 are new putative HTH motifs.

---

## ORAL PRESENTATION 41

### Improving Protein Function Prediction Methods with Improved Network Weighting and Integrated Literature Data

Presenter: Aaron Gabow, UCDHSC

Author(s): Aaron Gabow, Sonia Leach, Larry Hunter, Debra S. Goldberg

**ABSTRACT:** Motivation: Determining the function of uncharacterized proteins is a major challenge in the post-genomic era, due to the complexity of the problem and the large number of proteins. Identifying a protein's function contributes to

an understanding of the involved pathways' processes, how they can vary, and how they can be modified. Several graph-theoretic approaches predict unidentified functions of proteins by using the functional annotations of better-characterized proteins in protein-protein interaction networks. We improve the performance of these by including literature co-occurrence data and using new methods for predicting edge strength. We also quantify changes in performance as the algorithms work with an increasing level of annotation detail. Results: We find that including information on the co-occurrence of proteins within an abstract greatly boosts performance in both the global Functional Flow and local Majority algorithms. We were able to achieve similar performance gains by better integrating data from various databases.

---

**ORAL PRESENTATION 42****A Topology-Based Clustering Algorithm for Analysis Very Large Biological Networks**

Presenter: Xiaowei Xu, University of Arkansas at Little Rock

Author(s): Xiaowei Xu, Zhidan Feng, Nurcan Yuruk

**ABSTRACT:** Biological networks such as protein regulatory networks and metabolic pathways are very large complex networks with huge amount of information need to be analyzed. In this paper we present a novel topology based clustering algorithm for the analysis of very large networks including social networks, scientific collaboration networks, customer networks and biological networks. We first introduce the basic concepts of topology-based clustering algorithm. Then a hierarchical topology-based clustering algorithm, which is preferable because of the hierarchical structural nature of the networks, is presented. The topology is defined based on the neighborhood of vertices and a similarity measure defined by the fraction of shared neighbors. Since the topology-based algorithm needs exactly one neighborhood search for each vertex and the neighborhood search has a constant cost using the adjacency list for the sparse graphs, which are the most popular networks in the real world, the complexity of topology-based clustering algorithm is linear to the number of vertices in the networks. We conducted extensive experiments to evaluate our algorithm by using both synthetic and real network datasets. The results show that our approach outperforms the modularity-based clustering algorithm in both the speed and the accuracy.

---

ORAL PRESENTATION 43

**The Phylogenetic Position of the Mitochondrion**

Presenter: James McInerney, National University of Ireland

Author(s): David A. Fitzpatrick, Christopher J. Creevey

**ABSTRACT:** The evolutionary history of life on the planet is difficult to resolve. This is partially due to methodological difficulties and shortcomings of current evolutionary models. It is also due to horizontal gene transfer (HGT). As a result, the origins of the mitochondrial endosymbiont have proved difficult to resolve. Specifically, the place for the mitochondrion among the alpha-proteobacteria is contentious and requires careful analysis. We demonstrate that HGT has not completely wiped out the phylogenetic history of the alpha-proteobacteria and that we can confidently place the mitochondrion as a sister group of the Rickettsiales.

---

ORAL PRESENTATION 44

**Co-conservation Analysis of Bacterial Genes Across Phyla Predict Gene Function**

Presenter: Anis Karimpour-Fard, University of Colorado Health Science Center

Author(s): Anis Karimpour-Fard, Corrella S. Detweiler, Ryan T. Gill, Lawrence Hunter

**ABSTRACT:** Co-conservation is a well studied method for predicting functional relations and interaction between proteins. Although this method is widely used, the patterns of co-conservation that arise are found for individual protein pairs with a single target species. To improve protein function prediction, we developed and tested a new method, Multiple Species Cluster Co-Conservation, that incorporates interaction data between groups, instead of pairs, of proteins across multiple bacteria species. We demonstrate that clustered co-conservation of large sets of genes delineates groups of proteins that function in a coherent biological process and is more informative than pairs of proteins alone. We also found that the co-conservation method can be used to identify genes that function in the same process, genes that function in related processes and to predict the function of previously uncharacterized genes. Thus, the co-conservation method can be used as a tool for the blind annotation of genomes as well as to speculate about the function of gene-clusters containing genes with unknown function.

## ORAL PRESENTATION 45

**De Novo Signaling Pathway Reconstruction From Multiple Data Sources**

Presenter: Dongxiao Zhu, Stowers Institute for Medical Research

Author(s): Dongxiao Zhu, Michael Rabbat , Alfred O Hero, Robert Nowak , Mario Figueiredo

**ABSTRACT:** Signaling pathways are the primary means of regulating cell growth, differentiation and apoptosis. The de novo signaling pathway reconstruction problem can be divided into two sub-problems: discovery of pathway components and ordering the pathway components. While the literature abounds with computational and biological approaches for discovering pathway components, there has only been limited research on ordering pathway components, despite its importance. The main biological approach, genetic epitasis analysis, is limited by the cost and unavailability of mutants. Existing computational approaches reconstruct the network from numerical data which may be unreliable. Consequently, these approaches are sensitive to data selection. Here we describe a new statistical approach to signaling network reconstruction exploiting information about which genes belong to each pathway to reconstruct the “gene regulation network topology” in the form of a first-order Markov chain transition matrix. The approach naturally integrates information from multiple data sources such as text literature and biological expert knowledge, and is not limited to the numerical and categorical data used by previous approaches. The performance and stability of this approach follow directly by the scale-free property of the biological networks. We demonstrate the advantages of this approach over previous approaches using three well-known signaling pathways.

## ORAL PRESENTATION 46

**The Role of Discretization in Modeling Signal Transduction Networks**

Presenter: David J. John, Wake Forest University

Author(s): David J. John, Edward E. Allen, Leslie B. Poole, Richard F. Loeser, Jacquelyn Fetrow

**ABSTRACT:** An important aspect of computational algebraic based modeling of signal transduction networks involves the discretization of experimental data. Discretization is recognized by researchers as a fundamentally important problem that poses significant challenges with respect to the identification of biological signal. Given a set of experimental data, there are many discretizations that can be applied. Each individual discretization can be used to produce a model which contains some biological signal, amidst noise introduced both by experiment and discretization. This research produces and compares models using individual discretizations as well as those produced by an algorithmic consensus technique across various discretizations. In the consensus model each discretization technique will contribute its interpretation of biological information or signal. This consensus

modeling technique is applied to experimental data from human chondrocytes. These data follow the time course of phosphorylation of a small number of proteins (twelve proteins with six time points) following stimulation by insulin-like growth factor I (IGF-I). Depending on the cell type, this perturbation has been shown to stimulate two major signaling pathways, the phosphoinositide-3-kinase (PI3K) and the mitogen-activated protein kinase (ERK/MAPK) pathways. Comparison between known and predicted pathways is presented.

---

**ORAL PRESENTATION 47**

**A Generalized Algorithm of Unsupervised Learning**

Presenter: Anca Radulescu, University of Colorado at Boulder

Author(s): Paul Adams, Kingsley Cox, Anca Radulescu

**ABSTRACT:** Recent work on Long Term Potentiation in brain slices shows that Hebb's rule is not completely synapse-specific, possibly due to intersynapse calcium diffusion. We extend the classical Oja unsupervised learning model of a single linear neuron to include Hebbian infidelity, by introducing an error matrix  $T$ , which expresses the crosstalk between Hebbian updating at different connections.  $T$  has off-diagonal elements that depend on  $n$  (the input dimension) and  $q$  (a measure of the "quality", or accuracy of the Hebbian rule at each synapse). We show the modified algorithm converges to the leading eigenvector of the matrix  $TC$ , where  $C$  is the input covariance matrix. We study, analytically as well as through Matlab simulations, how the accuracy of the answer depends on the quality value, for different degrees of input correlations and for various network sizes. We find that the output error increases (learning becomes less useful) with decreases in quality or increases in size. We argue that this is because the linear model is sensitive only to pairwise statistics, and that a more powerful model that uses higher order correlations would show an error catastrophe under biologically plausible conditions.

---

**ORAL PRESENTATION 48**

**Principal Component Models to Identify Co- and Differential-Gene Expression in Time-Course Microarray Data**

Presenter: Rajagopalan Srinivasan, National University of Singapore

Author(s): Rajagopalan Srinivasan, Sudhakar Jonnalagadda

**ABSTRACT:** Time-course microarray experiments are useful for studying evolution of biological processes. In a typical experiment, expression levels of thousands of genes are measured at different time-points. There is a vast potential to use this data to understand the functioning of organisms. However, suitable data analysis techniques are essential to extract useful information from this data. In this paper, we propose a unified methodology based on multivariate statistics to group genes into clusters,

compare gene clusters and identify genes that are differentially expressed in different conditions. We model the expression data using Principal Components Analysis (PCA) and extract Principal Components (PCs) that represent the fundamental patterns (regulatory programs) of the cells. The expression profile of each gene is represented as a linear combination of dominant PCs with gene specific scores. Genes can be grouped into different clusters using their scores on different PCs. Further, PCA models of different clusters can be compared using a similarity metric to identify distinct clusters. Also, by projecting the expression dataset from a different biological condition on the developed PCA model and comparing scores, we can identify genes that are differentially expressed between the conditions. We illustrate the proposed method using publicly available microarray datasets.

---

**ORAL PRESENTATION 49****Finding Informative Sentences in Full-Text Journal Articles**

Presenter: Zhiyong Lu, University of Colorado School of Medicine

Author(s): Zhiyong Lu, William A. Baumgartner, Jr., J. Gregory Caporaso, K. Bretonnel Cohen, Lawrence Hunter

**ABSTRACT:** “Informative” sentences make assertions about the function of a gene or gene product. When linked to specific genes, they provide a mechanism for aggregating information about those genes that would otherwise be scattered throughout multiple articles and journals. Previous approaches to this solution for the “fragmentation problem” have focussed on extracting sentences from abstracts. For example, the NLM’s Gene References Into Function (GeneRIFs) are mostly pulled verbatim, or with slight modifications, from paper abstracts or titles. This mechanism is limited because it misses many important experimental results that are not included in abstracts. Therefore, it is important to be able to identify informative sentences in the body of full-text journal articles. Not only can this complement GeneRIFs, but it has the potential to serve as input for more sophisticated processing. For example, such sentences can serve as the evidence sentences for various annotation projects (e.g., Gene Ontology annotation). In this talk, I will present the methodologies we developed based on observational data from gene function annotations, as well as our own work on predicting GeneRIFs using text mining techniques. I will show how we applied them to a real-world task: finding interaction sentences in the BioCreAtIvE 2006 shared task.

## Exploring Inconsistencies in Genome-Wide Protein Function Annotations

Presenter: Carson Andorf, Iowa State University

Author(s): Carson Andorf, Dr. Drena Dobbs, Dr. Vasant Honavar

**ABSTRACT:** Incorrectly annotated sequence data are becoming more commonplace as databases increasingly rely on automated techniques for annotation. Hence, there is an urgent need for computational methods for checking consistency of such annotations against independent sources of evidence and detecting potential annotation errors. We show how a machine learning approach designed to automatically predict a protein's Gene Ontology (GO) functional class can be employed to identify potential gene annotation errors. In a set of 211 previously annotated mouse protein kinases, we found that greater than 95% of the GO annotations returned by AmiGO appear to be inconsistent with the UniProt functions assigned to their human counterparts. In contrast, 97% of the predicted annotations generated using a machine learning approach were consistent with the UniProt annotations of the human counterparts, as well as with available annotations for these mouse protein kinases in the Mouse Kinome database. We conjecture that most of our predicted annotations are, therefore, correct and suggest that the machine learning approach developed here could be routinely used to detect potential errors in GO annotations generated by high-throughput gene annotation projects.

## Phylogenetic Analysis of Teeth Decaying Oral Bacteria Through CLUSTAL W

Presenter: Abdul Arif Khan, Department of Microbiology, College of Life Sciences, Cancer Hospital & Research Institute

Author(s): Abdul Arif Khan

**ABSTRACT:** Various oral bacteria have suspected role in dental caries causation. Bacterial isolates from oral bacteria has been confirmed many times for their potential to produce acid, because acid production is cardinal for the development of dental caries. However the role of specific bacterial species or group of bacteria in dental caries is still controversial. Some researcher proposed the *Streptococcus mutans* as a major cause of dental caries, while some proposes role of other bacteria in this event. During present study, total sixty one different bacterial strains were chosen for phylogenetic analysis. Total twelve *Streptococcus sobrinus* strains, thirteen strains of *S. mutans*, ten strains of *S. salivarius*, six strains of *S. mitis*, one strain of *S. gordonii*, three strains of *S. sanguis*, three strains of *S. sanguinis*, one strain of *Uncultured Enterococci*, one strain of *S. italicus*, seven strains of *Actinomyces viscosus*, two strains of *Actinobacillus actinomycetemcomitans*, and two strains of *Lactobacillus acidophilus* were analyzed. 16S RNA sequences of each strains was retrieved through Genebank and their multiple sequence alignment was performed through clustal W.

Multiple sequence alignment revealed that *Streptococcus salivarius* accession AY581142 and *Actinomyces viscosus* accession no. M33908 are more related to each other, rather than their other strain of same genera. So, multiple sequence alignment can answer various enigma related to involvement of oral bacteria in dental caries causation, it can also give clue about the involvement of mixed bacteria in dental caries, as similarities between bacteria can also inform us about any methods to counteract these all caries bacteria by only a single drug by finding out their similarities and differences.

### Using GoogleRanks™ for Phage Phylogeny

Presenter: Annalinda Arroyo, San Diego State University, Department of Mathematics and Statistics

Author(s): Annalinda Arroyo, Chrystian Irazoque

**ABSTRACT:** Using protein similarity, we define an approach to molecular based phylogeny for bacteriophages. The phylogeny uses phage-phage distances obtained by weighted averaging of the protein-protein distances between every pair of proteins. The weights in the averaging use GoogleRanks™ of the associated random walks. We expect to find an interesting and significant depiction of the relatedness between bacteriophage species.

### Modeling Mixtures of Cells using Flow Cytometry

Presenter: Mike J. Boedigheimer, Amgen

Author(s): M.J. Boedigheimer

**ABSTRACT:** Flow Cytometry has become a standard technique for measuring physical attributes of single cells in a suspended cell mixture such as size, granularity and other indirectly labeled attributes by virtue of fluorescent labeled tags. By using different labels multiple attributes can be measured simultaneously allowing different subsets of cells to be studied independently. Currently, this is common done using a process of gating. Despite being commonly used, gating has some well-known problems, such as being inefficient and prone to artifacts. Here we introduce a multivariate non-gating technique that accomplishes the same goals as gating while eliminating many of its weaknesses and an application to automate the analysis. The results are as good as expert manual gating with greatly reduced effort and provides for more robust statistical inference.

### **Information Retrieval, Question-Answering, Machine Learning, and Concept Recognition in TREC Genomics 2006**

Presenter: J. Gregory Caporaso, University of Colorado Health Sciences Center, Dept. of Biochemistry and Molecular Genetics

Author(s): J. Gregory Caporaso, William A. Baumgartner, Jr., Hyunmin Kim, Zhiyong Lu, Helen L. Johnson, Olga Medvedeva, Anna Lindemann, Lynne Fox, Elizabeth K. White, K. Bretonnel Cohen, and Lawrence Hunter

**ABSTRACT:** TREC Genomics 2006 presented a challenge to automatically identify document passages that answer questions on twenty-seven topics from a collection of 162,259 full-text biomedical journal articles. Questions were derived from actual information needs of biomedical researchers, and performance was based on human evaluation of the retrieved passages. The Center for Computational Pharmacology approached this task by generating a candidate result set which was subsequently expanded (to improve recall) and then pruned (to improve precision). First, we created a Lemur search index and converted questions into term-expanded queries. Pseudo-relevance feedback was employed to expand search results. Next, from a pool of zone-filtered documents, we further expanded the result set using an LSA-like approach. In the final, false-positive eliminating step, we used naïve Bayesian classifiers trained on human-labeled data with features including bigrams, normalized semantic concepts, and OpenDMAP-style pattern matches. Three separate experiments allow us to compare the utility of the different techniques in this task.

### **Glycosylation Site Prediction Using Machine Learning Approaches**

Presenter: Cornelia Caragea, Iowa State University

Author(s): C. Caragea, J. Sinapov, A. Silvescu, D. Dobbs and V. Honavar

**ABSTRACT:** Protein glycosylation is one of the most complex and important post-translational modifications (e.g., some proteins cannot fold unless they are glycosylated). It is a site-specific, enzymatic process of addition of saccharides to proteins, occurring on S and T (O-Linked), N (N-Linked) and W residues (C-Linked). In this work, we address the problem of predicting glycosylation sites using different machine learning approaches applied to sequences for which different alternative representations are used. We investigate it on glycosylation data at O-GlycBase, a resource containing experimentally verified glycosylation sites. We build models on windows of different lengths centered at S, T, N, and W. Because our dataset is large and unbalanced, we use an ensemble classifier approach: we train a bag of classifiers, with each being trained on a balanced fraction of the data. Three types of classifiers were used: Support Vector Machine with different kernels: *o*/1 String kernel, BLOSUM62 Substitution Matrix kernel, and Polynomial kernel, Naïve Bayes and Decision Tree. Our ensemble classifier predicted O-Linked sites with 0.89 accuracy, 0.57 correlation coefficient (CC),

0.65 precision, 0.62 recall, N-Linked sites with 0.93 accuracy, 0.76 CC, 0.75 precision, 0.86 recall, and C-Linked sites with 0.78 accuracy, 0.63 CC, 0.65 precision, 0.98 recall.

## Structure-Activity Relationship Analysis of Chemical Compounds with Antioxidant Activity

Presenter: Weiguo Fan, Kent State University

Author(s): Weiguo Fan, Xin Lin, Boren Lin, Johnnie Baker, Chun-Che Tsai

**ABSTRACT:** A topological approach is used for structure-activity relationship (SAR) analysis of chemical compounds with antioxidant activity through structural activity maps (SAMs). SAMs are graphical maps plotting a molecular descriptor such as NAB (number of non-hydrogen atoms and bonds in a molecule) or MSI (molecular similarity index) against biological activity such as antioxidant activity. NAB or MSI is used to quantify the chemical structures. A program called TOPSIM was designed to find the maximal common substructure (MaCS) of a pair of molecules, each represented as a two-dimensional (2D) labeled graph. The molecular similarity index (MSI) and the topological distance (TD) define the similarity/dissimilarity of compounds. The algorithm uses a subgraph isomorphism approach to finding the maximum common subgraph (MCSG). TD and MSI calculated based on MaCS are used to build SAMs. SAMs provide a very efficient way of representing and visualizing SAR information in a chemical database. SAMs also allow comparison of grouped topological isomers and exploration of important trends in activity and sites of modification through structural orderings. A structural ordering is a set of structurally related compounds. It is utilized in this study to examine the effects of systematic modification of compounds on their biological activities.

## A Customizable Evaluation Platform for Biomedical Information Extraction Systems

Presenter: Graciela Gonzalez, Arizona State University / School of Computing and Informatics, Dept of Biomedical Informatics

Author(s): Graciela Gonzalez, Anthony Gitter, Craig Teegarden, Chitta Baral

**ABSTRACT:** The ever growing number of biomedical articles published has created a dire need to access information hidden in text. Many biomedical information extraction systems have been developed to extract protein-protein interactions and other relationships such as gene-disease or gene-drug interactions. It is difficult to compare the performance of the extraction strategies because there is no standard for evaluation of such systems. In general, precision and recall are the performance measures used, but it is meaningless to compare these values when the tests

performed vary so greatly. Developers make different decisions for critical aspects, and render the comparison of systems practically impossible, even when using the same articles and facts. We created a web-based platform that allows collaboration in extraction system evaluation and in creating “gold standards” that can be shared by multiple users. The platform allows evaluators to load PubMed abstracts or other references while viewing the extracted facts. The software is openly available and intuitive to use. Using the evaluation platform, developers are able to more accurately compare their system with competing systems, and identify weaknesses with greater ease. It is also a suitable platform for evaluating different systems, as is done for shared-tasks challenges like Biocreative and TREC.

### **New Method to De Novo Identify Repetitive Structures in Large Genomes**

Presenter: Wanjun Gu, Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center

Author(s): Wanjun Gu, Dale J Hedges, Mark A Batzer, David D Pollock

**ABSTRACT:** The annotation of repeat structure in newly sequenced eukaryotic genomes can be difficult because of the lack knowledge of repeat consensus sequence in RepBase and the large amount of genome data that needs to be processed and compared. Here, we introduce a novel heuristic approach to *de novo* identify the repetitive sequence in large genomes, which avoids similarity searches and sequence alignments, the most two time-consuming steps in comparative analysis. We use the word counts in the genome to get the excessive observed words and then cluster them to get the oligonucleotide excess probability clouds, or “P clouds”. After construction, P clouds were mapped back onto the genome, and regions of high P cloud density were identified as repetitive segments. This method is at least 10 times faster than current approaches for whole genome analysis. Comparing with the annotation of this method in human genome to RepeatMasker annotation, this method can mask most of the repeat elements RepeatMasker identified. Analysis of simulated genome with short Alu-segments shows our method have higher sensitivity to annotate short segments of known repeat elements than RepeatMasker. This method is useful in comparative analysis of eukaryotic genomes and *de novo* repeat analysis in new genomes.

## Comparative Studies of Elastin: Genomics of a Highly Repetitive Protein

Presenter: David He, University of Toronto

Author(s): David He, Fred W. Keeley, John Parkinson

**ABSTRACT:** Elastin is a polymeric, self-assembling, extracellular protein with remarkable durability, capable of withstanding billions of stretch-and-recoil cycles over decades without significant loss of performance or mechanical failure. As a result, elastin is an especially promising candidate for the development of novel biomaterials with wide ranging applications. However, while it is known that elastin's physical properties depend on its unusual sequence and domain structure, there is no clear understanding of how the sequence of elastin directly impacts these properties. The most significant challenge in unraveling these sequence-function relationships is the difficulty of analyzing elastin's repetitive yet highly variable sequence using traditional sequence analysis methods. We have applied an innovative application of tools previously developed for the study of protein interaction on a high quality, in-house elastin dataset, and have identified both common and species-specific 'core' structural elements within the elastin sequences. We further exploited this dataset to delineate a potential evolutionary path of elastin exons, suggesting duplicated regions which may have influenced the evolutionary-functional relationship of the elastin protein. Our methods have proven useful for the analysis of elastin sequences, and can be readily modified for other similarly repetitive sequences.

## An Isometric Strength Testing Device for Use With Elderly: Validation Compared With Isokinetic Measures

Presenter: Alexander T. Hutchison, University of Houston

Author(s): Alexander T. Hutchison, Mark S.F. Clarke

### ABSTRACT:

*Objective:* To estimate the concurrent validity of a new isometric strength testing device, the Leg Press Sled (LPS), using a criterion measure of isokinetic strength among a frail elderly population.

*Design:* Prospective validation study.

*Participants:* Eleven elderly subjects ( $81.7 \pm 7.0$  yrs) from an assisted living facility.

*Measurements:* Measurements of isometric leg strength using the LPS were compared with measurements of isokinetic leg strength using the Biodex System 3, as a reference standard. For isometric measures, a sample of elderly subjects (mean  $81.7 \pm 7.0$  yrs) was measured at knee angles  $60^\circ$  and  $90^\circ$ .

*Results:* Significant relationships ( $R_2 > 0.50$ ,  $P < .05$ ) between lower limb strength measures performed using the LPS and the Biodex System 3 were observed at both angles measured.

*Conclusion:* The LPS is a valid measure of leg strength in elderly populations. The device is portable, inexpensive, easy to use, and specifically suited for special-needs populations such as the home bound elderly due to its supine mode of operation.

## **Analysis of Ethnomedical Plants with a Potential for Drug Discovery — A Biomedical Informatics Approach**

Presenter: Beatrice Kilel, George Mason University

Author(s): Beatrice Kilel

**ABSTRACT:** Little information is currently known on some of the ethnomedical plants that have a potential for drug discovery. Hoodia cactus plant has recently captured headlines as an important resource with the potential as a biological pill for the obese based on the ingredient P57 that is known to suppress hunger. Some of the ethnomedical plants have been known to have a potential to cure some of the ailments that are not easy to cure using current drugs such as cancer. In order to ensure that these plants (such as *Prunus africana*, *Warburgia salutaris*) are not depleted, there has to be a way to regulate the harvesting regimes. Some of these resources have over time been pirated and patented in foreign countries without any profits to the rightful countries of origin.

*Prunus africana* is an ethnomedical plant that is traditionally used in Africa to treat ailments like chest pain, malaria and fevers. It has also been sought by large pharmaceutical companies for the manufacture of products to treat prostate gland hypertrophy (enlarged prostate gland) and benign prostatic hyperplasia (BPH).

Bark extracts are pulverized and incorporated into capsules and sold under various trade names, including Pygenil, produced in Italy, and Tadenan, produced in France. *Warburgia salutaris* is an important antimicrobial for treating yeast, fungal, bacterial and protozoal infections. It has been determined that the leaves and bark of this plant contains numerous drimane sesquiterpenoids, including warburganal and polygodial.

Interesting results were found in these less studied plants, and their potential for drug discovery. By having a more centralized repository to query from and make more scientific discoveries, more knowledge can be obtained about these ethnomedical resources.

## Interpreting Gene Expression Data using Protein Interaction Networks

Presenter: Sonia Leach, University of Colorado, Pharmacology  
Author(s): Sonia Leach

**ABSTRACT:** Protein interaction networks are a graphical means of representing the growing amount of information describing potential interaction between genes. Nodes in the graph correspond to genes and edges between the nodes represent a combination of physical, genetic, biochemical, functional, evolutionary or computational evidence of interaction between the corresponding genes. The challenge is to integrate the sources of interaction evidence while remaining sensitive to their relative reliability and coverage. We present methods which assign a reliability score to each evidence source and compute a probabilistic consensus likelihood of interaction. We demonstrate how the resulting consensus probabilities can be used to build visualization tools for the downstream analysis of gene expression data. Networks containing high confidence interactions provide a functional context for the genes and facilitate global interpretation of the observed changes in gene expression.

## Protein 3D Structure Classification Using Distance Images and Needleman-Wunsch Algorithm

Presenter: Sherif H. El Meligy, Teacher Assistant  
Author(s): Sherif H. El Meligy, Osama badawy Samah Sonbol

**ABSTRACT:** Since the 3D structure of a protein determines its function, the protein structural identification and classification is very important to biologists. Many protein representations and descriptors have been proposed. On the early days, the representation of protein structure to be used on classification consists of the position and the intra-molecule distance for all atoms consisting of protein. This representation had critical weak points due to high computational complexity and memory usage. In this paper we use an improved method by considering the position of alpha carbons (protein back bone) instead of using all atoms, Distance images are generated from the 3D coordinates of the alpha carbon atoms; then we apply a new technique to extract the information encoded in these images using block labeling and string alignment. Distance matrices or images are capable of representing specific protein structural topologies, and similar proteins in the same family tend to have similar distance matrices. A part of the SCOP database, which provides a structural classification of the proteins, is used as the ground truth in order to evaluate the classification accuracy of the proposed system; also a comparison is made between our system and other methods for further evaluation. The experimental results show that the proposed system achieves more than 99 percent classification accuracy.

## **Spatio-Temporal Modelling of Biochemical Pathways: Introducing Cell++**

Presenter: John Parkinson, Hospital for Sick Children / University of Toronto  
 Author(s): Chris Sanford, Matthew Yip and John Parkinson

**ABSTRACT:** High throughput genomic technologies are leading to the generation of vast amounts of data on cellular components, detailing their expression, interactions and organization within biochemical pathways. To understand how these relationships result in a functional pathway, a variety of computational tools have been proposed that attempt to simulate the behavior of the molecular components. In general, these tools tend to neglect the spatial organization of molecules, typically treating the system as a homogenous mixture of components. However, there is increasing evidence that spatial factors such as the co-localization of components have the potential to significantly influence pathway function and efficiency. Here we present Cell++, a novel spatio-temporal modeling platform that performs three-dimensional simulations of biochemical pathways. Combining a cellular automata engine with Brownian dynamics, Cell++ is capable of simulating the bulk properties of large quantities of small molecules (e.g., pyruvate), while simultaneously allowing larger molecules such as enzymes to be treated as more complex entities. Applying Cell++ to the study of metabolic pathways, we demonstrate how the spatial organization of enzymes can alter pathway efficiency and control the production of substrate intermediates, features consistent with the phenomenon of metabolic channeling. Further details of Cell++ can be found at: <http://www.compsysbio.org/CellSim/>.

## **The Conservation and Evolutionary Modularity of Metabolic Networks**

Presenter: Jose M. Peregrin-Alvarez, Hospital for Sick Children University of Toronto  
 Author(s): Jose M. Peregrin-Alvarez, John Parkinson

**ABSTRACT:** Comparative genomics of metabolic pathways across genomes produces invaluable information on their evolution, potential drug targets, and other biotechnological applications. Here we investigate the level of conservation of metabolic enzymes and pathways, and the extent to what orthologs of proteins involved in metabolic pathways are co-occurring together across genomes (referred as evolutionary modularity). We also analyse from a topological and evolutionary point of view the network of interactions where metabolic enzymes participate. We found that metabolic enzymes are highly conserved but not their participation in pathways. We generated a map of metabolic pathways that represents the similarity of metabolic pathway profiles and the extent of metabolic pathway similarity across taxa. Domain- and taxa-specific adaptations are identified and described. In addition, we found that metabolic pathways do not have a high

degree of evolutionary modularity, although their modularity becomes more evident when metabolic pathways are analysed in the context of individual taxa. Metabolic network reconstruction also revealed a scale-free topology. Finally, for the very first time, we reveal a core metabolic network highly conserved across all kingdoms of life which potentially represents an ancestral state of the metabolic network.

### **Characterization of Binding Site of Cyclin-Dependent Kinase (CDK9) by Homology Modeling, Molecular Docking and Pharmacophore Identification**

Presenter: Amresh Prakash, Jawaharlal Nehru University

Author(s): Amresh Prakash, Madhu Chopra

**ABSTRACT:** Cyclin-dependent kinase 9 (CDK9) has significant role in muscle differentiation, neuron, astrocytes maturation and is one of the causal factor of heart hypertrophy. The present work describes the structural requirements for designing inhibitors for CDK9, which is an attempt towards Insilico drug discovery. As the 3D structure of CDK9 is not known till date, we built a model structure of CDK9 by homology modeling with template 1GZ8, 1VoO and 1H4L, which have share sequence identity more than 40% with the target. Pharmacophores were generated on the basis of known inhibitors of CDK2 and CDK5 using the tool Catalyst/HipHop. The pharmacophores, in general, were found to be characterized of four features, namely, two hydrogen bond acceptors, one hydrogen bond donor and one aromatic ring. We conducted virtual screening<sup>2</sup> using compound libraries — NCI and MayBridge to generate potential CDK9 inhibitors. The initial high numbers of hits generated by 3D database search were further searched for drug likeliness properties using CAS SciFinder, which finally gave four lead molecules as CDK9 inhibitors. Keywords: CDK 9, heart hypertrophy, homology modeling, Pharmacophore, Virtual screening, Catalyst/HipHop

### **A Computational Approach to the Evolution of Cytochrome P450 Structures**

Presenter: Gowri Shankar, School of Information Technologies, University of Sydney

Author(s): Gowri Shankar, Michael Charleston, Michael Murray, David Hibbs

**ABSTRACT:** Cytochrome P450 (CYP) is an important enzyme in metabolizing drugs in humans and other species. This study aims at the evolution of this enzyme: How the structures have changed over time and between species and the significant changes in the drug binding site, which determines the function of the protein. We constructed evolutionary trees relating all 14 CYP proteins and used them to compare structures and sequences in a sound statistical framework. Structural differences were compared with the evolutionary distances between species. In the binding site of these proteins, there is a conserved pattern (F-G-X-G-X-H/R-X-C-L/I-G) with

a histidine to arginine change in mammals from archaea and bacteria at position 6. There are substantial changes in the structure among species and there is less change in the binding site, though the function of the proteins changes substantially with time. Many hydrophobic residues are present in the helices of these proteins; these residues have a greater influence on the ligand and have a weak interaction with the ligand and the heme present in the CYPs. In summary, we note a remarkable consistency of some sites over 2 billion years, but clear evolutionary changes particular to some species.

### **The BioCyc Collection of Pathway/Genome Databases**

Presenter: Alexander G. Shearer, SRI International

Author(s): Alexander G. Shearer, Ron Caspi, Carol A. Fulcher, Pallavi Kaipa, Peter D. Karp

**ABSTRACT:** The BioCyc collection of Pathway/Genome Databases (PGDBs) provides integrated representations of pathway and genome information for more than 200 organisms, of which most are microbes. Most BioCyc PGDBs were computationally derived from annotated genomes using the MetaCyc database (DB), which describes more than 800 experimentally determined metabolic pathways from more than 700 organisms. MetaCyc is a highly curated, literature-based DB of metabolic pathways and enzymes that provides a quality reference for pathway prediction. It is also an encyclopedic reference source for metabolic engineering and for other studies of metabolism. The computationally generated BioCyc PGDBs include predicted metabolic pathways as well as predicted fillers of holes in those metabolic pathways. BioCyc data are encoded using the Pathway Tools ontology, which facilitates the representation of complex biological knowledge with high fidelity. Pathway Tools provides a variety of visualization and analysis capabilities. Those to be presented here include its new comparative pathway analysis capabilities and its many data access mechanisms. Pathway Tools also has the ability to automatically generate metabolic map diagrams for each organism in the BioCyc collection. These diagrams can be used for analysis of omics datasets and can be enlarged to produce publication-quality metabolic wall charts.

### **Ultrametric Structure Exhibited in Phage Proteins**

Presenter: Chad Wagner, San Diego State University

Author(s): Chad Wagner

**ABSTRACT:** When phage proteins are compared to an ultrametric model, their phylogenetic structure adheres closely to the model for small neighborhoods. Statistical analysis of their deviations from ultrametricity reveals an interesting small scale structure and shows how this structure fits together into a larger geometric hierarchy.

## Linguistic Features of Conflict and Support Statements in Biomedical Papers

Presenter: Elizabeth K. White, University of Colorado Health Sciences Center  
Author(s): Elizabeth K. White

**ABSTRACT:** Scientific authors often note controversy in their field of research by way of introduction or discussion. Automatic recovery of conflicting and supporting evidence for a hypothesis can generate a bird's eye view of a field of study. Formal writing suggests several common rhetorical ways to express degree of agreement, and these can provide the basis for extracting support and conflict statements from scientific text. I analyzed six hundred support and conflict statements from biological papers to derive linguistic cues indicating corroboration or disagreement. I present a taxonomy of how these cues combine to form canonical conflict and support statements, and evaluate searches based on these cues against a test set of papers from the TREC 2006 corpus.

## Alu-enhanced Sequence Diversity in the Human Genome

Presenter: Hong Xue, Department of Biochemistry, HKUST  
Author(s): Siu-Kin NG, Hong Xue

**ABSTRACT:** Identifying features shaping the architecture of sequence variations is important for understanding genome evolution and mapping disease loci. In this study, high-resolution scanning of Alu-centered alignments of the human genome sequences has revealed a striking elevation of the frequency of single nucleotide polymorphisms inside Alu sequences (Alu-SNP) compared to their background sequence. This enhancement in SNP density is evident for all 24 chromosomes. Reduced levels of Alu-SNPs in the sex chromosomes, especially in the non-recombining region of the Y chromosome, are consistent with recombination events playing an important role in the enhancement. Variations in Alu-SNPs among the HapMap populations of European ancestry (CEU), Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT), and Yoruba from Ibadan, Nigeria (YRI) indicate that Alu-SNPs provide useful sequence markers, in addition to the Alu-insertion polymorphisms themselves, for the delineation of human genome evolution. That Alu-SNP levels are highest in the youngest Alu-Y, intermediate in the Alu-S of intermediate age, and lowest in the oldest Alu-J is consistent with the occurrence of not only genetic drift but also natural selection on the Alu-SNPs. Such evolutionary selection in turn suggests that Alu-SNPs might include potential sites of disease association, and therefore deserve detailed investigation.

## SPONSORS

### PLATINUM SPONSOR

#### IBM

1 Rogers Street; Cambridge MA 02142  
phone: 617-693-4581  
[www.ibm.com/servers/deepcomputing](http://www.ibm.com/servers/deepcomputing)



IBM collaborates with innovators and decision makers whose core business or research demands intense computation to advance technology to solve real business and industry problems. Together we're accelerating new uses of technology into areas previously limited by cost, knowledge or imagination. Deep Computing delivers powerful, innovative solutions to customers' most challenging and complex problems, enabling businesses and researchers to get results faster and gain a sustainable business advantage.

### SILVER SPONSORS

#### Affymetrix

3380 Central Expressway, Santa Clara, CA 95051  
Phone: 408-731-5000  
[www.affymetrix.com](http://www.affymetrix.com)



Affymetrix provides the industry standard platform for monitoring genomic information using microarray technology. Affymetrix offers an open bioinformatics platform allowing integration of its technology into any bioinformatics resource. For more information please visit [www.affymetrix.com/genechip/developer](http://www.affymetrix.com/genechip/developer)

#### Hewlett-Packard Company

200 Forest Street, Marlborough, MA 01752-3085  
Phone: 508-467-9748  
[www.hp.com/go/lifesciences](http://www.hp.com/go/lifesciences)



For researchers working to explore a theory, create new materials, or develop a cure, HP technology can help accelerate research, shorten product development time, and gain competitive advantage. HP solutions are ideally suited for the scientific and business challenges of life and materials sciences customers. HP offers the broadest choice of solutions — based on industry standards to maximize return on investment; the largest portfolio of life and materials sciences applications — well supported and with optimum performance; and HP's heritage of collaboration and innovation enables us to deliver the right solution for your research and budget.

HP is a technology solutions provider to consumers, businesses, and institutions globally. The company's offerings span IT infrastructure, global services, business and home computing, and imaging and printing. For the four fiscal quarters ended July 31, 2006, HP revenue totaled \$90.0 billion.

## CORPORATE SPONSORS

**Biodesix, Inc.**

1370 Bob Adams Drive, Steamboat Springs, CO 80477

Phone: 970-870-9041

[www.biodesix.com](http://www.biodesix.com)

biodesix

Biodesix intends to be the leading provider of mass spectrometry based clinical diagnostics. By applying its unique and proprietary data analysis on clinical samples, Biodesix's scientists are isolating signatures pertinent to disease management, drug discovery, and therapeutic stratification. Biodesix is collaborating with leading industrial and academic oncology centers of excellence.

**Dharmacon, Inc.**

2650 Crescent Drive, Suite 100; Lafayette, CO 80026

phone: 800-235-9880

[www.dharmacon.com](http://www.dharmacon.com)

DHARMACON

RNA TECHNOLOGIES

Dharmacon is the world's leading provider of synthetic RNA, siRNA and related RNA-interference products and technologies. Dharmacon's SMARTselection™ and SMARTpool® siRNA technologies provide the industry's highest level of guaranteed gene silencing. Dharmacon offers guaranteed siRNA reagents targeting all unique human genes in the NCBI RefSeq database.

# NOTES







Rocky '06 is an  
official conference of the  
International Society for  
Computational Biology



Rocky 06' is partially supported by the  
Computational Bioscience Program at the  
University of Colorado School of Medicine