

**ROCKY '05**

# **Third Annual Rocky Mountain Regional Bioinformatics Meeting**

**Silvertree Hotel**

**Aspen/Snowmass**

**December 9-11, 2005**

Chaired by Lawrence Hunter, Ph.D. and Stephen C. Billups, Ph.D.

Sponsored by the International Society for Computational Biology



## WELCOME TO ROCKY '05

Dear Rocky '05 Participant,

Welcome to the third annual Rocky Mountain Regional Bioinformatics Meeting a regional 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 in our area. We think we have the best program yet, offering a remarkable cross-section of bioinformatics research.

There is clearly a lot of computational bioscience going on in the region. Presenters at this meeting come from New Mexico, Arizona, Nevada, Colorado, Utah, Alaska and Alberta, representing universities, industrial enterprises, government laboratories, medical libraries and even a biological field station. Two of our keynote speakers (Profs. Ben-Hur and Kechris) have just moved to the area to take new faculty positions. The meeting is a chance to get to know your local colleagues, look for collaborative opportunities, and find synergies that can drive our field forward.

We've listened carefully to comments made about the previous Rocky meetings, and tried to be responsive. We've moved the meeting space to a hotel so we can stay together through the full three days, and so that transportation is less of an issue. We've retained (and expanded) the education workshop to discuss issues relevant to training the next generation of bioinformaticians. And we've arranged for small group ski lessons so that those of you who want to learn to ski will have a chance to do so among friends and colleagues. We welcome any further suggestions that you have to make next year's meeting even better.

We should all be grateful for the support of our sponsors: ISCB, IBM, Affymetrix, AMD, Apple, Ariadne Genomics, Dharmacon, John Wiley & Sons, Cancer Informatics, Exagen Diagnostics, and PLoS. The meeting would simply not be possible with organizational help from the ISCB, as well as from Kathy Thomas and Susan Trapp at the University of Colorado School of Medicine. 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 Affymetrix and IBM lay out some of the challenges we face and some the resources we will need to address them. We're grateful for the valuable data and the opportunity for supercomputer time that they offer us as well.

Many of you have contributed \$5 to join the Rocky Mountain Regional Bioinformatics group, an affiliate of the ISCB. We have to charge dues of some sort in order for the ISCB to recognize us as an affiliate. 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 Steve Billups, *Rocky '05 co-chairs*

## FRIDAY, DECEMBER 9, 2005

2-6pm Registration

3-6pm **WORKSHOP PANEL DISCUSSION:** *Mining for Bioinformatics Jobs: Training students with the appropriate skill set for a successful career in the Bioinformatics Field*

6:30-10pm **DINNER AND OPENING KEYNOTE:** *Computational Bioscience in The Rockies*  
**Larry Hunter**, University of Colorado School of Medicine

## SATURDAY, DECEMBER 10, 2005

8am-Noon Registration

9-9:45am **INVITED KEYNOTE:** *A Kernel Method for Predicting Protein-Protein Interactions*  
**Asa Ben-Hur**, Colorado State University

9:45-9:52am **ORAL PRESENTATION 1:** *Mathematical Techniques for Predicting and Analyzing Ontological Protein Function Annotations*  
 Presenter: **Cliff Joslyn**, Los Alamos National Laboratory  
 Author(s): Cliff Joslyn, Karin Verspoor, Judith Cohn and Sue Mniszewski

9:52-9:59am **ORAL PRESENTATION 2:** *Mathematical Modeling of DNA Damage Response Pathway in Saccharomyces Cerevisiae*  
 Presenter: **Farzin Imani**, Yale University School of Medicine  
 Author(s): Farzin Imani, David Tuck

9:59-10:06am **ORAL PRESENTATION 3:** *Whole Genome Transcript Analysis with Affymetrix Exon Microarrays*  
 Presenter: **Charles Sugnet**, Affymetrix  
 Author(s): Charles Sugnet, Alan Williams, Yaron Turpaz, Jim Veitch, Tyson Clark, Anthony Schweitzer, Melissa Cline, Hui Wang, Raymond Wheeler, John Blume

10:06-10:13am **ORAL PRESENTATION 4:** *Integration of Expression Data and Transcriptional Control Network: Significant Regulators Driving Expression Changes*  
 Presenter: **Andrey Y. Sivachenko**, Ariadne Genomics, Inc.  
 Author(s): A. Y. Sivachenko, A. Yuryev, N. Daraselia, I. Mazo

10:15-10:35am Coffee Break

## AGENDA

- 10:35-10:42am **ORAL PRESENTATION 5:** *The Bootstrap and Genomic Data: The Potential for Over-Confidence*  
Presenter: **James McInerney**, *Bioinformatics Laboratory*  
Author(s): David A. Fitzpatrick, Christopher J. Creevey, James O. McInerney
- 10:42-10:49am **ORAL PRESENTATION 6:** *Protdist: An Analysis of Error*  
Presenter: **Chad Wagner**, *San Diego State University*  
Author(s): Chad Wagner, Anna Salamon, Pat McNairnie, Rob Edwards, Peter Salamon
- 10:49-10:56am **ORAL PRESENTATION 7:** *Integration and Enrichment of OBO Ontologies*  
Presenter: **Michael Bada**, *University of Colorado at Denver and Health Sciences Center*  
Author(s): Michael Bada and Lawrence Hunter
- 10:56-11:03am **ORAL PRESENTATION 8:** *Bioinformatics Research: An Introduction to NCBI Tools*  
Presenter: **Dana Abbey**, *Denison Memorial Library*  
Author(s): Dana Abbey
- 11:03-11:10am **ORAL PRESENTATION 9:** *Unison: Integrated Feature-Based Mining for Target Discovery*  
Presenter: **Reece Hart**, *Genentech, Inc.*  
Author(s): Reece Hart
- 11:10-11:17am **ORAL PRESENTATION 10:** *A Sentence Recognizer for Mutant Protein Structure Studies: Toward Intelligent Systems for the Management of Structural Biology Data*  
Presenter: **J. Gregory Caporaso**, *University of Colorado Health Sciences Center*  
Author(s): J. Gregory Caporaso, K. Bretonnel Cohen, Lawrence Hunter
- 11:17-11:24am **ORAL PRESENTATION 11:** *Statistical Data Visualization of Two Important Proteins as Phylogenetic Tools in Chloroplast Genomes*  
Presenter: **Beatrice Kilel**, *George Mason University*  
Author(s): Beatrice Kilel
- 11:24-11:31am **ORAL PRESENTATION 12:** *Inferring Dynamical Gene Regulatory Networks Subject to Noise*  
Presenter: **Andre S. Ribeiro**, *Institute for Biocomplexity and Informatics*  
Author(s): Stuart A. Kauffman, André S. Ribeiro, Jason Lloyd-Price

# AGENDA

11:31-11:38am **ORAL PRESENTATION 13:** *UIMA as a Platform for Biomedical Natural Language Processing*  
Presenter: **William A. Baumgartner, Jr.**, *University of Colorado Health Sciences Center*  
Author(s): William A. Baumgartner, Jr., Andrew E. Dolbey, K. Bretonnel Cohen, and Lawrence Hunter

11:38-11:45am **ORAL PRESENTATION 14:** *Computational Identification of Binding Sites In Proteins*  
Presenter: **Drena Dobbs**, *Iowa State University*  
Author(s): M Terribilini, C Yan, F Wu, J-H Lee, J Sander, P Zaback, V Honavar, D Dobbs

11:45am-12:30pm **INVITED KEYNOTE:** *Blue Gene/L Impacting Computational Biology—One Year Later*  
**Kirk Jordan**, *IBM Strategic Growth Business/Deep Computing*

12:30-4pm **BREAK**

4-4:45 pm **INVITED KEYNOTE:** *Visualizing Bioinformatics Results*  
**Christoph Sensen**, *University of Calgary*

4:45-4:52pm **ORAL PRESENTATION 15:** *Dissection of the Human Genome Using Repeat Probability Clouds*  
Presenter: **David Pollock**, *Louisiana State University*  
Author(s): Wanjun Gu, Dale J. Hedges, Mark A. Batzer, David D. Pollock

4:52-4:59pm **ORAL PRESENTATION 16:** *MoBioS: A Conventional, Unconventional Database Management System Infrastructure for Biological Discovery*  
Presenter: **Daniel Miranker**, *University of Texas, Austin*  
Author(s): Daniel P. Miranker, Rui Mao, , Smriti Ramakrishnan, Willard S. Willard and Weijia Xu

4:59-5:06 pm **ORAL PRESENTATION 17:** *Genome-wide Analysis of the Distances Between Human Transcription Factor Binding Sites*  
Presenter: **Hyunmin Kim**, *University of Colorado School of Medicine*  
Author(s): Hyunmin Kim and Lawrence Hunter

5:06 -5:13pm **ORAL PRESENTATION 18:** *BioFPGA- An Open Source Hardware Project*  
Presenter: **Martin Gollery**, *University of Nevada, Reno*  
Author(s): Brian Beck, Martin Gollery

5:13-5:20pm **ORAL PRESENTATION 19:** *The Network Inference Testbed Software Environment*  
Presenter: **Ronald Taylor**, *Pacific Northwest National Laboratory*  
Author(s): Ronald Taylor

## AGENDA

- 5:20-5:27pm **ORAL PRESENTATION 20:** *Novel Approaches in Peptide and Protein Identification in Shotgun Proteomics*  
Presenter: **Karen Meyer-Arendt**, *University of Colorado*  
Author(s): Karen Meyer-Arendt, Shaojun Sun, Chia-Yu Yen, Steven Russell, William M Old, Natalie G Ahn, Katheryn A Resing
- 5:27-5:34pm **ORAL PRESENTATION 21:** *The Ecology of Place and Place-based Database Management Systems*  
Presenter: **Ian Billick**, *Rocky Mountain Biological Laboratory*  
Author(s): Ian Billick
- 5:34-5:41pm **ORAL PRESENTATION 22:** *Analysis of Human Promoters and Gene Expressions by an Integrative Approach: Constructing an Index Toward Gene Expression Patterns*  
Presenter: **Kihoon Yoon**, *University of Texas, San Antonio*  
Author(s): Kihoon Yoon, Stephen Kwek
- 5:45-6:30pm **INVITED KEYNOTE:** *Sequence Analysis of Human Alternative Splices Predicted from Exon Junction Arrays*  
**Katerina Kechris**, *University of California, San Francisco*
- 6:30-8:30pm **RECEPTION AND POSTERS WITH AUTHORS**

---

### SUNDAY, DECEMBER II, 2005

- 9-9:45am **INVITED KEYNOTE:** *Emerging Technology and Applications of Affymetrix GeneChips: Implications for Data Management and Analysis*  
**Steve Lincoln**, *Affymetrix, Inc.*
- 9:45-9:52am **ORAL PRESENTATION 23:** *Molecular Dynamics Evidence for Enzymatic Enabling of the Re-Configuration of the HIV Coat Glycoprotein gp120*  
Presenter: **Jack K. Horner**, *Science Applications International Corporation*  
Author(s): Jack K. Horner
- 9:52-9:59am **ORAL PRESENTATION 24:** *SCANS: System for Compression and Analysis of Nucleotide Sequences*  
Presenter: **Valerio Aimale**, *SeiraD, Inc.*  
Author(s): Valerio Aimale, Callum Bell, Joe Gatewood
- 9:59-10:06am **ORAL PRESENTATION 25:** *Bacterial Genes in a Eukaryotes*  
Presenter: **Gayle Philip**, *National University of Ireland*  
Author(s): Gayle K. Philip and Dr. James O. McNerney

## AGENDA

- 10:06-10:13am **ORAL PRESENTATION 26:** *ATGC: A Web Tool for Multiple Genome Comparison*  
Presenter: **Guoqing Lu**, *University of Nebraska at Omaha*  
Author(s): Guoqing Lu, Liying Jiang, Etsuko Moriyama, Luwen Zhang
- 10:13-10:20am **ORAL PRESENTATION 27:** *Inferring Three-way Gene Interactions from Microarray Data Sets*  
Presenter: **Jiexin Zhang**, *University of Texas M.D. Anderson Cancer Center*  
Author(s): Jiexin Zhang, Yuan Ji, Li Zhang
- 10:20-10:27am **ORAL PRESENTATION 28:** *Computational Alchemy*  
Presenter: **Dan McShan**, *University of Colorado at Denver and Health Sciences Center*  
Author(s): Dan McShan
- 10:27-10:34 am **ORAL PRESENTATION 29:** *To be announced*
- 10:34-10:41am **ORAL PRESENTATION 30:** *Mining Protein Transport Data from GeneRIFs*  
Presenter: **Zhiyong Lu**, *University of Colorado School of Medicine*  
Author(s): Zhiyong Lu; Larry Hunter
- 10:45-11:05am **COFFEE BREAK**
- 11:05-11:12am **ORAL PRESENTATION 31:** *Transcriptional Control and Behavioral Changes in Selectively Bred Mice*  
Presenter: **Razvan Lapadat**, *University of Colorado at Denver and Health Sciences Center*  
Author(s): Razvan Lapadat, Sanjiv Bhawe, Lawrence Hunter, Paula Hoffman, Boris Tabakoff
- 11:12-11:19am **ORAL PRESENTATION 32:** *The Survival Value of Dormancy: Human Hematopoietic Stem Cells and Hibernating Ground Squirrels*  
Presenter: **Tom Marr**, *University of Alaska*  
Author(s): Tom Marr
- 11:19-11:26am **ORAL PRESENTATION 33:** *Gene Classification by Decision Tree Data Mining Methods: Identification of Genes Likely to be involved in Human Genetic Disease*  
Presenter: **Julius Goth**, *North Carolina State University*  
Author(s): Julius Goth, ClarLynda Williams-DeVane, Jihye Kim
- 11:26-11:33am **ORAL PRESENTATION 34:** *To be announced*

## AGENDA

- 11:33-11:40am **ORAL PRESENTATION 35:** *Challenges in Analyzing Methylation cDNA Microarray*  
Presenter: **Mihail Popescu**, *University of Missouri*  
Author(s): Mihail Popescu, Gerald Arthur, Charles Caldwell
- 11:40-11:47 am **ORAL PRESENTATION 36:** *Simple Disease Prediction Models for Cancer Screens?*  
Presenter: **James Lyons-Weiler**, *University of Pittsburgh Cancer Institute*  
Author(s): James Lyons-Weiler
- 11:47-11:54am **ORAL PRESENTATION 37:** *A Molecular Concept Map for Human Biology and Disease*  
Presenter: **Daniel Rhodes**, *University of Michigan*  
Author(s): Daniel R. Rhodes, Shanker Kalyana-Sundaram, Vasudeva Mahavisno, Nicole Kasper, Terrence R. Barrette, Debashis Ghosh, Arul M. Chinnaiyan
- 11:54-12:01pm **ORAL PRESENTATION 38:** *Development of a Proteome Database from a Genome Sequence or ESTs for Protein Identification and in silico Predictions*  
Presenter: **Dong Chen**, *Utah State University*  
Author(s): Dong Chen, Jon L. Pearson, Desai Prerak, Jake Michaelson, Bart C. Weimer
- 12:01-12:08pm **ORAL PRESENTATION 39:** *Identifying Temporally Dependent Variation in Noisy Gene Expression Data Without Replicates*  
Presenter: **Stephen Billups**, *University of Colorado at Denver and Health Sciences Center*  
Author(s): Stephen Billups
- 12:08-12:15pm **ORAL PRESENTATION 40:** *A Study to Identify Interactions Between Transcription Factors and Non-housekeeping Genes to Determine Expression Regulation of cancer genes*  
Presenter: **Kihoon Yoon**, *University of Texas at San Antonio*  
Author(s): Amitava Karmaker, Kihoon Yoon, Stephen Kwek
- 12:15-1pm **CLOSING KEYNOTE:** *Computer-Assisted Forensic Analysis of Mass Disasters*  
**Gene Myers**, *Howard Hughes Medical Institute*

## EDUCATION PANEL

**TOPIC:** *Mining for Bioinformatics Jobs: Training students with the appropriate skill set for a successful career in the Bioinformatics Field*

**FOCUS:** Last year the Education Panel discussion focused on introductions and exchange of ideas on best practices for Bioinformatics Education and course curriculum with in the regional west. This was accomplished by having representatives from each of the academic institutions introduce their Bioinformatics Programs and present the progress and challenges. This year we will focus on the bioinformatics skills needed to be successful in today's bioinformatics market in the academic, industry, and other non-academic environments.

An article in *Nature Reviews | Drug Discovery* (2004) on bioinformatics careers states, "the growing number of courses and experience workers in bioinformatics means that the recruitment market has become tougher for new graduates." According to a survey of academic training programs and an analysis of advertised job openings published in 2005 (*Biochemistry and Molecular Biology Education*), it was concluded that the labor market in bioinformatics has changed dramatically from 1990's to the early 2000's. The number of training programs, as well as enrollment, expanded rapidly during this period. The expansion has created a substantial pipeline of students who will matriculate from these programs in the near future. Results of the survey state: "while the expansion of training programs has occurred, the demand in bioinformatics market declined and its origins have shifted largely from industry to academe. Unless conditions in industry change dramatically in the next few years, it is likely that trainees from these programs will have difficulty finding the expect jobs in industry."

IF this proposed trend is accurate, this raises a number of issues, which are the topic of Rocky 2005 Education Panel Discussion. In Session I: we will have a panel of academics professionals who will discuss their perspective. In Session II: we will have a panel of non-academics describe their perspective. In Session III: we will open up the discussion to participants in a round table discussion.

Are we training our students appropriately for careers in academia and what is the required skill set needed?

How do we train our student to be successful in careers outside of academia given the current bioinformatics job market?

**SESSION I:** Bioinformatics skill set from Academic Perspective

**SESSION II:** Bioinformatics skill set from Industry Perspective

**SESSION III:** Round table discussion

## INVITED KEYNOTE SPEAKERS

Larry Hunter, *University of Colorado School of Medicine*

### COMPUTATIONAL BIOSCIENCE IN THE ROCKIES

---

Asa Ben-Hur, *Colorado State University*

### A KERNEL METHOD FOR PREDICTING PROTEIN-PROTEIN INTERACTIONS

Most proteins perform their function by interacting with other proteins. Therefore, information about the network of interactions that occur in a cell can greatly increase our understanding of protein function. We present a kernel method for predicting protein-protein interactions using a combination of data sources, including protein sequences, annotations of protein function, local properties of the network, and interactions in different species. We propose a pairwise kernel that provides a similarity between pairs of proteins, and illustrate its effectiveness in conjunction with a support vector machine classifier. We obtain improved performance by combining several sequence-based kernels and by further augmenting the pairwise sequence kernel with features that are based on additional sources of data.

---

Kirk Jordan, *IBM Strategic Growth Business/Deep Computing*

### BLUE GENE/L IMPACTING COMPUTATIONAL BIOLOGY - ONE YEAR LATER

Last year at Rocky '04, I asked the question is Blue Gene a System for Computational Biology? One year has passed and Blue Gene has been placed at several sites, many of which have a computational biology component. In this talk, I will review some of the results obtained at some of these sites. I will elaborate on work underway with collaborators and colleagues in the area of computational biology. For those unfamiliar with the Blue Gene System, I will very briefly describe the hardware and software environment. I will describe how Blue Gene might be used to tackle multi-scale problems, many existing in computational biology. While progress is being made, there remain many challenges for the computational biology community to apply the Blue Gene resource to "Big" science problems with impact on society that until now or in current implementations have fallen short of the mark. Finally, I will elaborate on opportunities that exist for the community to get access to Blue Gene.

---

Christoph Sensen, *University of Calgary*

### VISUALIZING BIOINFORMATICS RESULTS

My laboratory has long been involved in the creation of tools for the integration and visualization of Bioinformatics results. Early efforts focused on the characterization of Genomes, building tools such as MAGPIE and Bluejay in the process. While we are still working on tools for 2D Bioinformatics, we have embarked on the visualization of data with a spatial and temporal aspect, such as the results of

## INVITED KEYNOTE SPEAKERS

Gene Expression or Protein modification studies. Our approach is to build the computational infrastructure for 4D Bioinformatics "top down". This includes the use of Virtual Reality environments, the creation of a middleware layer, which allows scientists to explore their data space without scripting or programming, and the creation of "top down" models of the organisms that we work with (mainly humans, mice and rats). We are now beginning to conduct case studies using the new environment, which are focused on the characterization of genetic diseases and developmental patterns.

---

Katerina Kechris, *University of California, San Francisco*

### **SEQUENCE ANALYSIS OF HUMAN ALTERNATIVE SPLICES PREDICTED FROM EXON JUNCTION ARRAYS**

Alternative splicing of exons in a pre-mRNA transcript is an important mechanism that contributes to the protein complexity found in humans. Alternative splicing is thought to be regulated by various protein-protein and protein-RNA interactions, including those that involve specific sequence elements that act as enhancers or suppressors. The recent adaptation of DNA microarray technology to measure splice variants is providing new directions for the high-throughput study of alternative splicing. In this talk, methods will be presented for predicting alternatively spliced exons from splicing arrays that consist of exon-junction probes. I will then illustrate how contrast word counts and regression-based methods can be used to identify candidate enhancers and silencers that may regulate splice site choice.

---

Steve Lincoln, *Affymetrix, Inc.*

### **EMERGING TECHNOLOGY AND APPLICATIONS OF AFFYMETRIX GENECHIPS: IMPLICATIONS FOR DATA MANAGEMENT AND ANALYSIS**

Over the last few years, microarray technology has made significant contributions to biomedical research and development by allowing high-quality gene expression and genotyping data to be generated in volume cost-effectively. Of the technology components needed to successfully generate and exploit such data, arguably computational analysis remains both critical today as well as a vital ongoing area of research. With current evolution of microarray technology, we believe this situation is soon to be amplified.

Using new manufacturing and instrumentation technologies which scale down to 5 micron features, Affymetrix GeneChip cartridges with over 6 million different probes, as well as 96-well plates with 150 million probes, can be reliably manufactured, hybridized and scanned. In this very brief overview we will describe various new GeneChips which have been designed using these methods and we will illustrate biological data from these chips in a variety of applications: gene expression, splice

## INVITED KEYNOTE SPEAKERS

variant analysis, transcriptome analysis, genetic variation and copy number analysis. We will also provide a brief introduction to changes being implemented in instrumentation and software to better accommodate the size and complexity of these data. We will discuss data sets and other materials available to the research community who may wish to pursue any one of a number of open problems in the field.

---

Gene Myers, *Howard Hughes Medical Institute*

### **COMPUTER-ASSISTED FORENSIC ANALYSIS OF MASS DISASTERS**

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.

# PRESENTATION ABSTRACTS

## ORAL PRESENTATION 1

### CLIFF JOSLYN, LOS ALAMOS NATIONAL LABORATORY

Author(s): Cliff Joslyn, Karin Verspoor, Judith Cohn and Sue Mniszewski

#### *Mathematical Techniques for Predicting and Analyzing Ontological Protein Function Annotations*

The Protein Function Inference Group (PFIG) at the Los Alamos National Laboratory (LANL) has developed an approach to automatically produce novel Gene Ontology (GO) functional annotations of proteins based on categorizing the regions of the GO to which similar (in some sense) proteins are annotated. Current input spaces include proteins which are near neighbors in BLAST space, or which are described by similar terms occurring close together in documents. Validation of this or similar methods depends on the development of evaluation metrics which are appropriate to the mathematical structure of the space of annotations, in this case, the directed acyclic graph (DAG) structure of the GO. In this talk, we first outline the LANL architecture for ontological function annotation, implemented within our POSet Ontology Laboratory Environment (POSOLE), and for both the CASP and BioCreative evaluations. We present some test results from the BLAST-based effort. We then discuss our novel evaluation metrics, and conclude with a consideration of their applicability to the general problem of measuring the consistency of annotation sets.

---

## ORAL PRESENTATION 2

### FARZIN IMANI, YALE UNIVERSITY SCHOOL OF MEDICINE

Author(s): Farzin Imani, David Tuck

#### *Mathematical Modeling of DNA Damage Response Pathway in *Saccharomyces Cerevisiae**

In response to DNA damage, cells recruit several signaling complexes at the site of DNA damage which activate checkpoint signaling cascades to halt cellular replication and elicit DNA repair. Mathematical models account for these processes can provide insights into understanding how cells maintain genomic stability and integrity. Our models are capable of simulating dynamics of this process using two conceptually distinct formalisms, continuous deterministic and discrete stochastic methods. The deterministic model consists of a set of differential equations that describe biochemical reactions including activation of Mec1 to autonomous dimerization of Rad53 via several steps. The reaction rate constants and initial concentrations have been measured in the laboratory or taken from the literature. The stochastic model implements the Gillespie's algorithm and simulates the same biochemical reactions. Considering the nanomolar concentration of reactants and the fact that the cascade may be initiated by a single site of Mec1 binding to DNA, a stochastic method that allows simulating discrete numbers of molecules yields more realistic model. Using this model we can follow the time course of the vari-

## PRESENTATION ABSTRACTS

ous complex formation processes and of the phosphorylation states of the proteins involved. We compare the two formalisms, and identify how they might complement each other.

---

### ORAL PRESENTATION 3

#### CHUCK SUGNET, AFFYMETRIX

Author(s): Sugnet, Charles; Williams, Alan; Turpaz, Yaron; Veitch, Jim; Clark, Tyson; Schweitzer, Anthony; Cline, Melissa; Wang, Hui; Wheeler, Raymond; Blume, John.

#### *Whole Genome Transcript Analysis with Affymetrix Exon Microarrays*

Ongoing technology developments have enabled the tiling of over 6 million probes on a single Affymetrix oligo microarray. This technology push has enabled the development of commercial whole genome exon arrays. Current exon microarray implementations consist of a single microarray covering roughly 1 million exons with 1.4 million probesets consisting of 4 perfect match probes per probeset for the human genome. Whole genome exon microarrays present researchers with additional insight into transcriptional complexity, such as alternative splicing, but also raise the need for improved and new analysis algorithms and improved computational efficiency. We have developed new algorithms for identifying alternative spliced exons based on ANOVA. We have applied these algorithms to a publicly available 11 tissue data set (3 replicates each) and 7 paired colon normal/cancer tissues data sets to identify exons whose splicing is differentially regulated.

---

### ORAL PRESENTATION 4

#### ANDREY SIVACHENKO, ARIADNE GENOMICS, INC.

Author(s): A. Y. Sivachenko, A. Yuryev, N. Daraselia, I. Mazo

#### *Integration of Expression Data and Transcriptional Control Network: Significant Regulators Driving Expression Changes*

Microarrays provide an invaluable insight into the biomolecular mechanisms, however raw results are disjoint genome-wide “one-gene-at a time” datasets with high levels of noise. Placing data into the biological context through integration with different data sources is critical both for noise reduction and for objectively quantifiable system-level hypothesis formulation. We analyze differential expression (DE) data in the context of large network of known transcription regulation events. DE data sample downstream of a regulator is compared to the sampling distribution derived from the network, with network connectivity taken into account. The analysis is aimed at elucidating regulators with statistically significant patterns of downstream expression changes and explaining DE data in terms of activated/suppressed regulatory cascades. The set of plausible regulatory events provides conceptual data reduction and a step towards elucidating/building extended pathways. We apply our analysis to a few disease datasets, demonstrate

## PRESENTATION ABSTRACTS

robustness and statistical significance of the results, and show that the sets of regulators suggested as putatively involved in the differential response are potentially interesting biologically and exhibit statistically significant overlap with sets of known disease associated genes. Assembling significant regulators into a putative signaling pathway and applications of our procedure to other networks (metabolic, binding) are also discussed.

---

### ORAL PRESENTATION 5

**JAMES MCINERNEY, BIOINFORMATICS LABORATORY**

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

*The Bootstrap and Genomic Data: The Potential for Over-Confidence*

We can now use genome-scale data for the inference of phylogenetic relationships. Therefore, the question of the most appropriate analysis method is important. We address whether the methods that hold for shorter datasets are appropriate for extremely long datasets. We also address the potential for extremely large datasets with no phylogenetic information returning a result that indicates the presence of a robust phylogeny. We present the result of simulation studies that show that bootstrapping may not be the most appropriate method of analysis for large concatenated datasets. We also address the issue of how bootstrapping behaves for genome-scale datasets when lateral gene transfer is present and demonstrate a particular problem with data concatenation among strains of the bacterium *Neisseria meningitidis*. Historically, it has been assumed that data can be concatenated when the phylogenetic signal from individual genes does not conflict. However, again, we demonstrate that this may be a dangerous paradigm.

---

### ORAL PRESENTATION 6

**CHAD WAGNER, SAN DIEGO STATE UNIVERSITY**

Author(s): Chad Wagner, Anna Salamon, Pat McNairnie, Rob Edwards, Peter Salamon

*ProtDist: An Analysis of Error*

The present study proposes new data-based methods predicated on the assumption of approximate ultrametricity for estimating the accuracy of phylogenetic distance measures for particular sets of proteins. Using a database of over 19,000 phage proteins, we find good validity for approximate ultrametricity up to a PROTDIST value of about 1.5 where the behavior makes a clear transition. The structure is more evident using pairwise alignments than multiple alignments and over 640,000 pairs of proteins were aligned in a pairwise manner as part of this study. We present several ways of seeing the ultrametricity and the transition around the PROTDIST score of 1.5. Our findings demonstrate the utility of these methods for estimating the accuracy of PROTDIST for phage proteins at different distances. The implications for phylogenetic inference are considered.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 7

MICHAEL BADA, UNIVERSITY OF COLORADO AT DENVER AND HEALTH SCIENCES CENTER

Author(s): Michael Bada, Lawrence Hunter

#### *Integration and Enrichment of OBO Ontologies*

We have begun investigating the merging and enrichment of the disparate biomedical ontologies of the Open Biological Ontologies (OBO), a set of over 50 open-source, orthogonal ontologies represented in a shared text-file format. While many of these ontologies are quite extensive, they are generally structurally simple, typically using at most a few relationships apart from the fundamental taxonomic is-a relationship. Furthermore, OBO was explicitly designed to be orthogonal so as to enable modular use of the ontologies, so there are no links among terms from separate ontologies. We have converted the Gene Ontology, Chemical Entities of Biological Interest, and the Cell Type Ontology into frame-based representations to be used in Protege, the preeminent tool for the authoring of frame-based ontologies, and these three ontologies have been merged into one project. We have additionally begun to add further associations among terms both within a given ontology and between separate ontologies by mining ontological terms for pattern elements within the terms. These pattern elements can be mapped to added attributes of the terms so that a given term containing a given pattern element can be linked to a term referenced by the text matching the pattern element.

---

### ORAL PRESENTATION 8

DANA ABBEY, DENISON MEMORIAL LIBRARY

Author(s): Dana Abbey

#### *Bioinformatics Research: An Introduction to NCBI Tools*

Established in 1988, the National Center for Biotechnology Information became the focal point for bioinformatics at the National Institutes of Health. They cull the best resources and make them available free to computer users worldwide. Major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data. The NCBI is committed to the development and implementation of tools that enable efficient access to, and use and management of, various types of information. The NCBI databases are linked through a unique search and retrieval system, called Entrez, which allows users to not only access and retrieve specific information from a single database but to access integrated information from many NCBI databases. Highlighted in this session: Evolutionary Biology, Protein Modeling with 3-D applications, Genome Mapping, Online Mendelian Inheritance in Man,

## PRESENTATION ABSTRACTS

Bioinformatics Support Network, and Bioinformatics Grants and Funding.

---

### ORAL PRESENTATION 9

**REECE HART, GENENTECH, INC.**

Author(s): Reece Hart

*Unison: Integrated Feature-Based Mining for Target Discovery*

Unison is a database of, and web interface to, precomputed sequence and structure predictions for a comprehensive set of protein sequences. The integration of these data enables the mining of sequences based on holistic protein feature criteria, the synthesis of predictions for individual sequence analysis, and the refinement of hypotheses regarding the composition of protein families. Unison includes prediction results for signal peptides, transmembrane domains, GPI anchoring, subcellular localization, secondary structure, sequence motifs, HMM, PSSM, and threading alignments, and genomic localization. SCOP, GO, HomoloGene, patent, and other auxiliary information permit richer queries and interpretations of prediction results. The PDB schema enables reliable structural localization of sequence features. Unison was designed to be kept up-to-date easily and the build process is automated. Sequence "run histories" enable incremental updates of precomputed results. The Unison schema, code, web interface, and non-proprietary data are released under the Academic Free License. They are available online and for local installation at <http://unison-db.org/>.

We have used Unison for mining projects involving TNF ligands, helical cytokines, death domains, and other protein families. I will demonstrate Unison's utility by describing our search for proteins containing Immunoreceptor Tyrosine Inhibitory, Activation, and Switch Motifs (ITIMs, ITAMs, ITSMs).

---

### ORAL PRESENTATION 10

**GREGORY CAPORASO, UNIVERSITY OF COLORADO HEALTH SCIENCES CENTER**

Author(s): J. Gregory Caporaso, K. Bretonnel Cohen, Lawrence Hunter

*A Sentence Recognizer for Mutant Protein Structure Studies: Toward Intelligent Systems for the Management of Structural Biology Data*

We are developing an information extraction system to identify and compile results of site-specific mutagenesis structural biology studies. Structural analyses of site-specific mutants are often performed to determine the significance of individual amino acid residues in the higher order structures of proteins. The quantity of literature documenting these mutant protein structures is already overwhelming and rapidly increasing. An automated literature search for mutation/structural change (M/SC) mappings could populate and keep current a database of the results of these studies which would constitute a valuable asset in many areas of

## PRESENTATION ABSTRACTS

structural biology research. We are experimenting with various machine learning approaches to build a sentence classifier which identifies discussion of M/SC findings in PUBMED abstracts, and have received promising preliminary results using support vector machines. We present data on the performance of M/SC sentence classifiers constructed using varied feature types and classification algorithms, in addition to information on the growth of the structural biology literature base and the role of sentence classification in our larger goal.

---

### ORAL PRESENTATION II

**BEATRICE KILEL, GEORGE MASON UNIVERSITY**

Author(s): Beatrice Kilel

*Statistical Data Visualization of Two Important Proteins as Phylogenetic Tools in Chloroplast Genomes*

The current problem facing whole genomes is lack of good visualization techniques that would allow rapid identification of regions of possible structural and/or functional importance. Different statistical data visualization tools can be used to further study genome relationships. These visual graphics go beyond the usual 2-D views to stereoscopic 3-D, which provides a better feel for depth of the data being analyzed. In dealing with the large amounts of data like the chloroplast genome data, there are chances that over plotting can become an imminent problem. This is because it becomes hard to

represent pixels for single observations in comparison to hundreds or even thousands of observations. In order to solve this problem, saturation brushing is used to desaturate the data until a very small and distinct component of color is obtained. Statistical data

visualization is important for viewing data simultaneously, identifying clusters, frequencies, relationships and patterns in data, identifying outliers and anomalous data with the techniques used depending on data and objective. This study has used scatter plots, principal component analysis, density estimate, vista plots to show how common genes in chloroplast genomes can be visualized graphically while discovering how they relate to each other.

---

### ORAL PRESENTATION 12

**ANDRE RIBEIRO, INSTITUTE FOR BIocomplexity AND INFORMATICS**

Author(s): Stuart A. Kauffman, André S. Ribeiro, Jason Lloyd-Price

*Inferring Dynamical Gene Regulatory Networks Subject to Noise*

We present an inference algorithm for dynamical gene regulatory networks (IAD-GRN) applied to Boolean networks, based on the ARACNe algorithm. IADGRN computes mutual information (MI) between pairs of genes states of activity, accepting a connection if MI is higher than a threshold. Also, existence of noise

## PRESENTATION ABSTRACTS

allows the removal of indirect correlations using the generalized triangle inequality test. Unlike ARACNe, IADGRN uses time series of genes state transitions to infer directed connections. Since thresholds are computed from the same time series, true connections can be inferred for any states update rules. We introduce new procedures that infer more true connections using acquired data, in circumstances where previous algorithms were not efficient and infer the update rules, using the certainty of the function values for each possible inputs value, to look for undetected connections. Combining the algorithm with simulated experimental procedures of connections confirmation, one is able to deduce from 75% to 85% of the network connections and Boolean rules, for all tested topologies. Finally, we present results for several topologies and various initial conditions, to estimate the sensitivity of the algorithms several stages to the network properties and determine the cause of failed inferences using new accuracy measures.

---

### ORAL PRESENTATION 13

**WILLIAM BAUMGARTNER, UNIVERSITY OF COLORADO HEALTH SCIENCES CENTER**

Author(s): William A. Baumgartner, Jr., Andrew E. Dolbey, K. Bretonnel Cohen, and Lawrence Hunter

#### *UIMA as a Platform for Biomedical Natural Language Processing*

We have evaluated IBM's Unstructured Information Management Architecture (UIMA) as a framework for natural language processing (NLP) tasks. Currently freely available to end users and scheduled for open-source release later this year, UIMA enables large-scale text processing while promoting tool sharing, modular development, code re-use, and I/O abstraction. We have found the formalized, yet flexible, data structure that UIMA uses for passing information between components to be well-suited for representing text-based information for a variety of tasks, including entity identification, information retrieval, document classification, and general annotation of text spans, and have extended the framework with a persistence mechanism that allows the storage and retrieval of arbitrarily complex annotations. Here, we discuss an annotation analysis system that has been developed in the UIMA framework. We demonstrate a comparison of several automatic sentence boundary detectors, followed by a straightforward transformation of the system into a comparator of entity identification systems. UIMA's platform independence, recent community software development efforts (e.g. the Mayo UIMA Starter Kit), and its upcoming open source release all suggest that this emerging technology holds significant potential for the biomedical NLP community.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 14

**DRENA DOBBS, IOWA STATE UNIVERSITY**

Author(s): M Terribilini, C Yan, F Wu, J-H Lee, J Sander, P Zaback, V Honavar, D Dobbs

#### *Computational Identification Of Binding Sites in Proteins*

Characterization of macromolecular interactions is important for problems ranging from rational drug design to analysis of metabolic and signal transduction networks. We are developing machine learning algorithms for analyzing protein complexes and generating classifiers capable of predicting which amino acids of a target protein participate in protein-protein or protein-nucleic acid interactions. Using only amino acid sequence information or a combination of sequence and structure-derived information as input, these classifiers can identify interface residues in protein-protein complexes with >77% overall accuracy; classifiers trained on diverse sets of protein-DNA or protein-RNA complexes can predict interface residues with ~77% and ~83% overall accuracy, respectively. We will present results of recent studies in which: i) classifiers were tested on a protein-RNA complex for which the structure has not yet been determined, but for which we have performed site-directed mutagenesis and RNA-binding experiments to verify our predictions; and ii) different methods of encoding and/or additional sources of information have been used to enhance classification performance. The ability to reliably predict which residues of a protein directly contribute to its interactions with other macromolecules should contribute to our understanding of the molecular basis for specific cellular recognition events.

---

### ORAL PRESENTATION 15

**DAVID POLLOCK, LOUISIANA STATE UNIVERSITY**

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

#### *Dissection of the Human Genome Using Repeat Probability Clouds*

The analysis of repeat structure in eukaryotic genomes can be time-consuming and difficult because of the large amount of information (~3x 10<sup>9</sup> bp) that needs to be processed and compared. We developed a novel approach to this task that avoids similarity searches and sequence alignments, two of the more time-consuming components of comparative analysis. We designed and implemented a set of algorithms to quickly calculate the exact counts for any length oligonucleotide in large genomes, and then analyzed oligonucleotide excess probability clouds, or "P clouds." These P clouds are composed of clusters of related oligonucleotides that occur, as a group, more often than expected by chance. After construction, P clouds were mapped back onto the genome, and regions of high P cloud density were identified as repetitive segments. We used this method to analyze the repeat content of the human genome, and found that in addition to being at least 10 times faster than current approaches for whole genome analysis, the sensitivity for

## PRESENTATION ABSTRACTS

detecting repeat-derived regions is considerably higher. The combined speed and sensitivity of this method should make it extremely useful in comparative analysis of eukaryotic genomes and de novo repeat structure analysis in newly sequenced genomes.

---

### ORAL PRESENTATION 16

**DANIEL MIRANKER, THE UNIVERSITY OF TEXAS AT AUSTIN**

Author(s): Daniel P. Miranker, Rui Mao, , Smriti Ramakrishnan, Willard S. Willard, Weijia Xu

*MoBIOs: A Conventional, Unconventional Database Management System Infrastructure for Biological Discovery*

MoBIOs, (The Molecular Biological Information System, pronounced mobius) is a specialized database management system with built-in biological data types and integral support for nearest-neighbor queries. The system provides scalable performance for both biological sequence analysis and mass-spectroscopy databases. An extended SQL language (mSQL) enables the concise and flexible definition of solutions to bioinformatics problems where current practice often entails the complex scripting of utilities.

Central to MoBIOs is the development [metric] distance-based models of biological similarity and concomitant index structures. We have developed and validated a metric model of amino acid substitution i.e. a new derivation of the PAM substitution matrix that forms a metric. We have further established that an approximate version of cosine distance can be used to manage spectrum databases. Applications to date include a comparative study of the Rice and Arabidopsis genomes in  $O(m \log n)$ , where  $m$  and  $n$  are the lengths of the genomes ( $n, m \sim 10^8$ ) vs.  $O(mn)$  for BLAST-based analysis and a system for protein identification by database look-up of theoretical mass-spectrometer signatures. In addition to a simpler programming model, 1 and 2 orders of magnitude performance improvements were demonstrated.

---

### ORAL PRESENTATION 17

**HYUNMIN KIM, UNIVERSITY OF COLORADO SCHOOL OF MEDICINE**

Author(s): Hyunmin Kim and Dr. Lawrence Hunter

*Genome-wide Analysis of the Distances between Human Transcription Factor Binding Sites*

Genome-wide analysis of human transcription factor binding site pairs reveals more than 1000 novel long distance interactions. Eukaryotic transcription factors function combinatorially, with many interactions between enhancer and repressor elements and the core promoter complex. Some such interactions have previously been demonstrated to occur at relatively long distances of more than 1.5 kbps. We

## PRESENTATION ABSTRACTS

present a novel method for identifying and characterizing pairs of transcription factor binding sites whose distance distributions differ significantly from an empirical random model. Using a database of 153 position weight matrices, we obtained 12 million pairs of binding sites in 12kbp upstream regions of human genes. The degree of dissimilarity was quantified by Kolmogorov-Smirnov statistic. The test p-values were corrected for multiple testing, resulting in 988 significant pairs with an expected false discovery rate of 0.001. The significant pairs were fitted to Mixture model for finding low variance peaks in their distance distribution. We found the peaks are annotated with a statistically significant Gene Ontology terms, suggesting functional coherence. These results suggest that the extent of long distance interactions in the regulation of human genes is much more prevalent than previously suspected, that distance characteristics among TFBSs can be useful in identifying previously unsuspected interactions.

---

### ORAL PRESENTATION 18

**MARTIN GOLLERY, UNIVERSITY OF NEVADA, RENO**

Author(s): Brian Beck, Martin Gollery

*BioFPGA: An Open Source Hardware Project*

Many open-source development projects have become available in the past few years, with BioPerl, Biopython, BioRuby and BioJava all providing useful codes to the community. We will present a similar platform (BioFPGA) for accelerated bioinformatics algorithms that run on FPGA hardware for speeds that are hundreds to thousands of times faster than java or C programs running on conventional processors. This will benefit the global community of FPGA developers which has been growing rapidly with the explosion of new tools in this field, and also the end user bioinformaticist who will eventually be able to install a powerful system on the desktop for a much lower price than with current commercial solutions. Ultimately, we hope to have a complete package of High-Throughput sequence analysis programs along with an entirely new suite of algorithms that have never before been accelerated on FPGAs.

---

### ORAL PRESENTATION 19

**RONALD TAYLOR, PACIFIC NORTHWEST NATIONAL LABORATORY**

Author(s): Ronald Taylor

*The Network Inference Testbed Software Environment*

The "Network Inference Testbed" (NIT) project at DOE's Pacific Northwest National Laboratory is building a software platform that permits direct inference of genetic regulatory networks from high-throughput microarray mRNA expression data. There are currently no interactive environments available to (1) evaluate different algorithms for inference of biological regulatory network structure using common data sets, and (2) easily apply such state-of-the-art algorithms to experi-

## PRESENTATION ABSTRACTS

mentally generated high throughput data. The NIT project will fill this gap, as an aid in the reconstruction of the structure of mRNA regulatory networks. Briefly, methods using high-throughput data rely on searching for patterns of partial correlation or conditional probabilities that indicate causal (regulatory) influence in the input set of expression values. Such patterns of partial correlations found in the high-throughput data, possibly combined with other supplemental data on the organism of interest, are the basis upon which the algorithms in the NIT's toolkit infer regulatory networks. The NIT software environment will permit direct comparisons of different inference algorithms on dynamically generated artificial networks of different topologies, with simulated perturbations represented as expression sets. It is also designed to analyze experimental high-throughput expression data using the suite of (trained) inference methods. Hence the testbed will be useful both to software developers wishing to compare, refine, or combine inference techniques, and to bioinformaticians analyzing experimental data.

---

### ORAL PRESENTATION 20

**KAREN MEYER-ARENDT, UNIVERSITY OF COLORADO**

Author(s): Karen Meyer-Arendt, Shaojun Sun, Chia-Yu Yen, Steven Russell, William M. Old, Natalie G. Ahn, Katheryn A. Resing

#### *Novel Approaches in Peptide and Protein Identification in Shotgun Proteomics*

Shotgun proteomics for protein profiling of complex samples is a systems biology method made possible by the convergence of sensitive mass spectrometers (MS), near completion of protein databases, and advances in modeling peptide MS fragmentation. Search programs used to identify peptides from MS fragmentation data are improving, but are still plagued by low sensitivity and specificity. Our lab uses consensus among multiple search programs, filters out unlikely identifications based on chemical properties inconsistent with the sample preparation protocol, and validates peptide identifications based on the latest models for expected fragmentation and spectral intensity. This allows us to identify peptides with higher sensitivity and specificity, but the protein inference problem is far from solved. We have developed IsoformResolver, a program based on a novel restructuring of the protein database in a peptide-centric manner, thereby enabling automated grouping of protein isoforms. Together with removal of unlikely peptide isoforms – alternate amino acid sequences that cannot be differentiated by current MS – our program eliminates overcounting of proteins without loss of information. Protein identifications are presented along with GO and other annotations and can be data-mined to find expression trends in cells.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 21

IAN BILLICK, ROCKY MOUNTAIN BIOLOGICAL LABORATORY

Author(s): Ian Billick

#### *The Ecology of Place and Place-Based Database Management Systems*

A central paradox for field biologists is that while they search for general insights, the results of field studies are often highly dependent upon location and time. Field biologists address this paradox, at least in part, by using contextual information to interpret and generalize otherwise idiosyncratic results. While many database management systems are conceptually organized and geographically dispersed, place-based database management systems are geographically organized and conceptually dispersed. Consequently place-based database management systems face unique challenges. I will use the efforts at the Rocky Mountain Biological Laboratory as a case study of the informatic tools that can be used to manage field data and enhance field studies.

---

### ORAL PRESENTATION 22

KIHOON YOON, THE UNIVERSITY OF TEXAS AT SAN ANTONIO

Author(s): Kihoon Yoon, Stephen Kwek

#### *Analysis of Human Promoters and Gene Expressions by an Integrative Approach: Constructing an Index Toward Gene Expression Patterns.*

Identification of gene controlling elements in human is fundamental to the understanding of the mechanisms of diseases. Here, we present an integrative analysis of promoter sequences and gene expressions of normal human tissues to create a promoter complexity index (PCI) as the primary indications of tissue specificities and expression levels. To achieve this goal, our approach must be sensitive enough to detect subtle differences in the controlling regions. We applied a new sequence signal detection algorithm to promoter and downstream regions of transcription start sites (TSSs). Our approach considers two cases that typical pattern finding algorithms may not be able to handle, (1) patterns reside on non-fixed positions relative to TSSs, but yet the regions are limited to ~30 bp and (2) rare pattern signals which may be ignored easily by “over representation”-based methods. The patterns found were further refined by integrating the mRNA expression profiles of normal human tissues to minimize possible false positive pattern detections. We are also currently developing a better way of gene expression analysis schemes to draw more meaningful co-expression information. In summary, we have identified unique sequence patterns from the promoters of housekeeping and tissue-specific genes which may reflect different gene controlling mechanisms.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 23

**JACK HORNER, SCIENCE APPLICATIONS INTERNATIONAL CORPORATION**

Author(s): Jack K. Horner

#### *Molecular Dynamics Evidence for Enzymatic Enabling of the Re-Configuration of the HIV Coat Glycoprotein gp120*

gp120 is an HIV coat glycoprotein that plays an essential role in cell attachment and membrane fusion, both of which are required for the virus to successfully attack a host cell. gp120 exists in two kinetically stable conformations. One of these forms is non-antigenic and incapable of promoting cell attachment and membrane fusion; the other form is strongly antigenic and promotes attachment and fusion. During transport, the non-antigenic form is converted to the antigenic form. A natural question in the rational design of therapeutic agents for HIV is whether the Gibbs free energy of the re-configuration is negative, i.e., whether the non-antigenic form is thermodynamically unstable and therefore could, assisted by an enzyme, spontaneously re-configure. If so, that enzyme would be a candidate target for rational therapeutics design. A non-negative Gibbs free energy would discourage the search for such an agent. Here, I use molecular dynamics simulation of the two forms of gp120 to show that the Gibbs free energy of the gp120 re-configuration is negative. Keywords: HIV, gp120, molecular dynamics, rational therapeutics.

---

### ORAL PRESENTATION 24

**VALERIO AIMALE, SEIRAD, INC.**

Author(s): Valerio Aimale, Callum Bell, Joe Gatewood

#### *SCANS: System for Compression and Analysis of Nucleotide Sequences*

SeiraD has developed a storage and analysis system (SCANS (TM)) for very large collections of complete human genomes. The software is based on suffix trees, delta compression and dictionary-based compression. The redundancy among human genome sequences means that a genome can be efficiently stored using delta compression as a set of differences versus a reference genome. The differences are identified by comparing sequences using a suffix tree algorithm. We have developed a distributed genome alignment server in which multiple suffix trees can persist in memory. This overcomes the need to build a suffix tree for each alignment, allows multiple comparisons to be done simultaneously, and takes advantage of a computational cluster environment. Deltas are extracted from the alignment output and stored in a relational database or in a custom-designed storage system. SCANS (TM) was able to store, in a test case, 5,000 simulated chromosomes (approximately 1 Terabyte of total data) using less than 4 Gigabytes of physical space. The achieved compression ratio amounts to 0.034 bits/base - two orders of magnitude better than currently available nucleotide sequence compression systems. Our technology offers a space and time efficient alternative to traditional database and sequence alignment approaches.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 25

**GAYLE PHILIP, NATIONAL UNIVERSITY OF IRELAND**

Author(s): Gayle K. Philip and Dr. James O. McInerney

#### *Bacterial Genes in a Eukaryotes*

The malaria parasite (genus *Plasmodium*) is a unicellular eukaryote, which invades the erythrocytes of its vertebrate host through the course of a complex life cycle. The disease is estimated to give rise to 270-515 million clinical cases each year with 1-2.7 million deaths, mainly attributable to *Plasmodium falciparum*. It was our objective to identify the origin of each of the 5,295 *P. falciparum* protein-coding genes and in particular, cases where the nearest neighbour to the *P. falciparum* protein was a prokaryotic sequence.

Homologues were identified by performing a FASTA search of a sequence library set of 867,899 proteins made up from 20 Archaeal, 179 Bacterial and 25 Eukaryotic completed genomes. Phylogenetic trees were then reconstructed for each of the *P. falciparum* genes and were manually examined. A number of genes were found to be candidates for having undergone a lateral gene transfer event. In particular, 35 genes were identified where a query *P. falciparum* gene was found to have homologues from bacterial genomes, but not from any archaeal or other eukaryotic genomes. 16 of these *Plasmodium*-bacterial specific genes have a function that is unknown, while the remaining genes are involved in different pathways including pyrimidine metabolism and fatty acid biosynthesis.

---

### ORAL PRESENTATION 26

**GUOQING LU, UNIVERSITY OF NEBRASKA AT OMAHA**

Author(s): Guoqing Lu, Liying Jiang, Etsuko Moriyama, Luwen Zhang

#### *ATGC: A Web Tool for Multiple Genome Comparison*

Comparative genomics is an essential tool for charactering unique genes specific to a given organism and homolog genes appearing in multiple genomes but derived from a common ancestor. Two commonly used approaches include genome alignment and adding anchors on the genome. The first approach is exemplified as MUMmer whereas the second approach exemplified as BLAST. BLAST uses a heuristic algorithm for local sequence similarity search. The potential disadvantage of using BLAST for genome comparison is that it finds too many matches of short fragments. Here we modified the algorithm and used another parameter called coverage as blast search threshold. We implemented this algorithm in a web application. This application is particularly useful for comparison of small genomes, such as virus genomes.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 27

**JIEXIN ZHANG, UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER**

Author(s): Jieixin Zhang, Yuan Ji, Li Zhang

#### *Inferring Three-way Gene Interactions from Microarray Data Sets*

It has been an important and challenging problem to infer the network of gene interactions from microarray data. Conventional methods use correlation of expression profiles between two genes to look for signs of co-expression. However, patterns of co-expression are often obscured because they change depending on biological conditions such as tissue types, or diseases. In this study, we used a large microarray dataset of various human cancers to survey for three-way gene interactions, in which co-expression of two genes depends on the expression level of a third gene. Such three-body interactions cannot be derived from two-body interactions based on pair-wise correlations. We used a model-based clustering algorithm to identify genes with bimodal expression profiles, and partitioned the samples accordingly. We then identified the gene pairs of which correlation of expression changed significantly between the two partitions of samples. To perform cross validation, we randomly split our collection of 545 samples into a training-set of 360 and testing-set of 185. Our survey found ~83000 significant gene triplets (permutation test p-values  $< 10^{-9}$  in the training-set, of which 61% have p-values  $< 10^{-6}$  in the testing-set.). Our results may prove valuable in the construction of complex gene networks.

---

### ORAL PRESENTATION 28

**DAN MCSHAN, UNIVERSITY OF COLORADO AT DENVER AND HEALTH SCIENCES CENTER**

Author(s): Dan McShan

#### *Computational Alchemy*

I will discuss the transmutation of base materials into valuable ones, utilizing biochemistry. Synthetic chemistry has in many ways been the cornerstone for late 20th century technology and progress. In most cases, the synthetic chemist is actually making a compound discovered biologically. The work of Nobel laureate, E.J. Corey, in retrosynthesis has provided a rational process for developing novel synthetic routes, and is responsible for many if not most of the organic syntheses today. Such technology has provided our modern world with its materials and enabled advances in nearly every field of human endeavor. However, it is inherently inefficient and has also generated an unfathomable amount of organic, toxic waste. I will introduce retrobiosynthesis as the biotech equivalent of retrosynthesis and claim that it will have a similar impact, enabling the future of more efficient, and more environmentally sound production of the materials of the future.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 29

Author(s): To be announced

### ORAL PRESENTATION 30

ZHIYONG LU, UNIVERSITY OF COLORADO SCHOOL OF MEDICINE

Author(s): Zhiyong Lu, Larry Hunter

#### *Mining Protein Transport Data from GeneRIFs*

Protein transport, by which proteins synthesized in the cytoplasm are inserted into or moved across membranes, is essential to all living organisms. Understanding the mechanism of protein transport has been a central theme in cell biology and has been studied for several decades. Thousands of individual research results have been published in the scientific literature, but have not been captured systematically. We therefore propose to build a central repository of protein transport data by extracting statements from the biomedical literature using natural language processing technology. We will use this task to demonstrate our work in developing openDMAP (Direct Memory Access Parser)—a new kind of concept recognition system. In this application, we aim to automatically extract protein transport relevant concepts (including the transport origin, destination, cargo and driver) from GeneRIFs (Gene Function Into Reference)—concise phrases describing a gene function or functions. In this presentation, we will discuss several aspects of the system including the annotating training data, concept recognition via patterns and some preliminary results produced by the parser on a dozen protein transport GeneRIFs.

### ORAL PRESENTATION 31

RAZVAN LAPADAT, UNIVERSITY OF COLORADO AT DENVER AND HEALTH SCIENCES CENTER

Author(s): Razvan Lapadat, Sanjiv Bhawe, Lawrence Hunter, Paula Hoffman, Boris Tabakoff

#### *Transcriptional Control and Behavioral Changes in Selectively Bred Mice*

Cells must continually adapt to changing conditions by modifying their gene expression profiles. One of the core components involves transcriptional regulatory interactions. We focus transcriptional regulatory networks both because the increasing number of transcriptional profiling data sets, with “whole transcriptome” chips becoming the norm in the field and the fact that the process of gene expression can be regarded as the origin and effector of a response. In our strategy for the analysis of coregulated genes we use promoter sequence similarity searches in differentially expressed genes, cross species transcription factor mapping and in silico signaling pathway and literature mining. We have added to the analytical repertoire of our analytical techniques siRNA and miRNA binding prediction tools. Transcriptional profiles of whole brain extracts from mice selectively bred

## PRESENTATION ABSTRACTS

for ethanol preference (HIGH/LOW) or acute functional tolerance were measured using Affymetrix microarrays. Differentially expressed genes were analyzed using a positional scoring matrix algorithm for the first 2 kb upstream regions to identify conserved sequence patterns. Independently, the same regions were analyzed for the most conserved transcription factor binding sites based on a cross-species model. Resulting predictions were used together with the differentially expressed genes as input for signal transduction pathways and literature mining in order to establish a model of the interaction network modulating the gene expression events. The identified processes synthesize the integrated neuronal response of cells bearing different genotypic fingerprints, including signal transduction, transcriptional regulation, ion channel activity and neuronal activity modulation. Our work strongly suggests that combining transcription control module discovery with interaction network data mining represent a powerful approach for cis-regulation of gene expression and the involvement of signal transduction mechanisms using high-throughput techniques.

---

### ORAL PRESENTATION 32

**THOMAS MARR, UNIVERSITY OF ALASKA**

Author(s): Tom Marr

*The Survival Value of Dormancy: Human Hematopoietic Stem Cells and Hibernating Ground Squirrels*

I will present the results of a literature review on the survival value of the dormancy phenotype as it exists from bacteria to insects to whole mammals, focusing on examples of extreme expression of the phenotype in highly stressful environments. I will also present the results of a ~48k transcript gene expression screen of dormant human hematopoietic stem cells in chronic myelogenous leukemia negative individuals vs. chronic myelogenous leukemia positive patients.

---

### ORAL PRESENTATION 33

**JULIUS GOTH, NORTH CAROLINA STATE UNIVERSITY**

Author(s): Julius Goth, ClarLynda Williams-DeVane, Jihye Kim

*Gene Classification by Decision Tree Data Mining Methods: Identification of Genes Likely to be Involved in Human Genetic Disease*

Due to the copious volume of annotated biological information and sequence data, it has become increasingly practical to analyze and extract common patterns at the sequence level. Several databases provide human disease gene information, and recent studies have shown distinct characteristics among them. Therefore, data mining algorithms have proven to be an invaluable method in detecting interesting patterns in these datasets. Clustering algorithms, a popular choice within the bioinformatics community, may not be an appropriate choice in situations where prior information is readily available; conversely, decision tree algorithms

## PRESENTATION ABSTRACTS

provide a supervised method capable of learning patterns from known disease features. Furthermore, decision tree algorithms produce stable models whose results are easy to read and analyze, reproducible, and fairly accurate given sufficient training data and a measured amount of attributes. We propose a study on the effectiveness of decision tree classifiers by applying probability scores of conserved sequences against previously identified disease and healthy genes. A major benefit of discerning disease genes from non-disease genes is to provide further direction for the annotation of newly discovered genes. All models resulting from our analyses will be learned via the Weka framework.

---

### ORAL PRESENTATION 34

Author(s): To be announced

---

### ORAL PRESENTATION 35

**MIHAIL POPESCU, UNIVERSITY OF MISSOURI**

Author(s): Mihail Popescu, Gerald Arthur, Charles Caldwell

#### *Challenges in Analyzing Methylation cDNA Microarray*

We used a DNA methylation microarray-based method, differential methylation hybridization (DMH), to examine patterns of CpG island methylation in small B-cell lymphoma (SBCL). Methylation microarrays raise specific analysis problems. Some of the problems are due to the current technique, DMH, such as: detection only of the fully methylated genes if the CpG island region contains several restriction enzyme cut sites, different amount of DNA hybridization material used in each chip, and low signal to noise ratio. Other problems are intrinsic to the gene methylation process such as: a hypermethylated gene does not result in other genes being hypermethylated, the hypomethylation of the gene from a cancer patient compared to a normal person is ignored. We investigated three solutions to the above problems: a goal driven approach to normalization, a simultaneous processing of methylation and expression microarrays and a pathway centered analysis based on an ad hoc pathway methylation model. We applied the above methods to a set of methylation and expression cDNA microarrays obtained from 43 patients with three types of SBCL: MCL, FL and CLL. The analysis resulted in identifying sets of genes that were later confirmed to be hypermethylated in one of the lymphoma types studied.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 36

**JAMES LYONS-WEILER, UNIVERSITY OF PITTSBURGH CANCER INSTITUTE**

Author(s): James Lyons-Weiler

#### *Simple Disease Prediction Models for Cancer Screens?*

Because cancer is an intrinsically heterogeneous disease, with numerous and diverse genetic, genomic and proteomic etiologies, we expect diagnostic and prognostic information in biomarkers to be heterogeneously distributed among patients with the same cancer type. We have recently developed tests for finding potential biomarkers that may be relevant for a significant subset of patients. Further extending these concepts, we have designed a GA-optimized prediction model that anticipates the heterogeneity of cancer. For numerous novel biomarker data sets, we demonstrate that our prediction model outperforms the popular fixed-marker intensity input additive prediction models on independent test sets. Our results transform the notion that markers with low sensitivity and specificity for detection cancer are failures into a perspective of great opportunity of cancer biomarker panel development.

---

### ORAL PRESENTATION 37

**DANIEL RHODES, UNIVERSITY OF MICHIGAN**

Author(s): Daniel R. Rhodes, Shanker Kalyana-Sundaram, Vasudeva Mahavisno, Nicole Kasper, Terrence R. Barrette, Debashis Ghosh and Arul M. Chinnaiyan

#### *A Molecular Concept Map for Human Biology and Disease*

The proliferation of genome-scale experimentation, data collection, and computational analysis has led to a mass of invaluable molecular data scattered across disparate databases, publications, and websites. In order to unify disparate genomic data, we define a 'molecular concept' as any set of genes or proteins that are related in some way (e.g. genes repressed by Interleukin-10, genes activated in prostate cancer, genes with OCT-4 binding sites, etc.). Here we present the compilation and integrative analysis of a diverse collection of molecular concepts ranging from protein complexes, to cancer signatures, to microRNA target genes. This analysis generated thousands of molecular concept links suggesting common regulatory and functional pathways in human biology and disease. Many of the links are validated by the literature, while others suggest novel pathways deregulated in human disease. For example, the 'IL-10 repression' concept largely overlaps with the 'Burkitt's lymphoma' concept, consistent with the observed anti-tumorigenic effects of IL-10 in Burkitt's Lymphoma. Also, the 'Low grade prostate cancer' concept strongly overlaps with the 'activated by androgen' concept, whereas the 'High grade' concept overlaps with the 'repressed by androgen' concept, suggesting that differential androgen pathway activity may be responsible for the observed heterogeneity of prostate cancer.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 38

DONG CHEN, UTAH STATE UNIVERSITY

Author(s): Dong Chen, Jon L. Pearson, Desai Prerak, Jake Michaelson, Bart C. Weimer

#### *Development of a Proteome Database from a Genome Sequence or ESTs for Protein Identification and in silico Predictions*

Production of plant and microbial genome sequences presents the opportunity to have a searchable in silico proteome database. This approach provides a yard stick by which the methods and data quality can be evaluated from 2D gel analysis to a putative protein identification. We have created four proteome databases (Alfalfa, *Lactobacillus acidophilus* NCFM, *Lactobacillus plantarum* WCFS1 and *Lactobacillus johnsonii* NCC 533) based on the genome sequences available in the public domain. The protein sequences from each ORF were used to calculate protein molecular weight, pI, and create a theoretical Trypsin digested database. For the organisms without ORF information or a genome sequence, ESTs were converted to protein sequences by FLIP. These databases have been linked to the Pathway Tools database to enable visualization of the metabolic relevance of the proteins in a graphical presentation format. The created proteome databases at the Center for Integrated BioSystems were used to generate theoretical 2D gels and for protein identification through peptide fingerprint mapping. Specific examples of the success for this approach will be presented.

---

### ORAL PRESENTATION 39

STEPHEN BILLUPS, UNIVERSITY OF COLORADO, DENVER AND HEALTH SCIENCES CENTER

Author(s): Stephen Billups

#### *Identifying Temporally Dependent Variation in Noisy Gene Expression Data Without Replicates*

Gene expression microarray technology has made it possible to measure mRNA expression levels of thousands of genes simultaneously. Measuring these expression levels at different times over the course of a biological process offers a powerful tool for deciphering the roles various genes play in these processes. Of primary interest in such studies is the variation of expression levels due to temporal dependencies (relative to the biological process). However, such temporally dependent variation can become obscured by other sources of variation. This presentation addresses the question of how to identify genes with significant temporal variations. The method exploits the fact that temporally related variation is of lower frequency than other sources of variation. This enables the calculation of a test statistic based on smoothing. This statistic is then used in a false discovery rate method to identify genes with significant temporally dependent variation.

## PRESENTATION ABSTRACTS

### ORAL PRESENTATION 40

AMITAVA KARMAKER, UNIVERSITY OF TEXAS AT SAN ANTONIO

Author(s): Amitava Karmaker, Kihoon Yoon, Stephen Kwek

#### *A Study to Identify Interactions Between Transcription Factors and Non-housekeeping Genes to Determine Expression Regulation of Cancer Genes*

Discovering genes primarily responsible for any particular cancer is a challenging issue in cancer research. The expression of non-housekeeping genes are predominantly regulated by transcription factors. To study gene expressions in normal and cancer cell types, microarray technology has been widely used. In our research, we perform a systematic study to determine which genes are regulated by which transcription factors and whether an over/under-expressed gene is caused by changes in expression levels of its regulatory transcription factors. The dataset we used consists of 26,260 unique genes for 115 normal tissue samples from 19 different organs. We construct a family of subsets of genes that are co-expressed in a specific tissue or a group of tissues that exhibit similar function. Within each of the subsets of genes, we identify the transcription factors that may regulate gene expression levels. We also found several groups of transcription factors that may act collectively as co-factors in gene regulation. This is modeled by using Pearson's Correlation Co-efficient and linear regression. In other words, we have constructed a partial gene regulatory network. More importantly, we have found evidence of some transcription factors responsible for causing abnormal genes expressions in various types of cancerous cells.

## POSTERS

### *SCANS: System for Compression and Analysis of Nucleotide Sequences*

VALERIO AIMALE, SEIRAD, INC

Author(s): Valerio Aimale, Callum Bell, Joe Gatewood

SeiraD has developed a storage and analysis system (SCANS (TM)) for very large collections of complete human genomes. The software is based on suffix trees, delta compression and dictionary-based compression. The redundancy among human genome sequences means that a genome can be efficiently stored using delta compression as a set of differences versus a reference genome. The differences are identified by comparing sequences using a suffix tree algorithm. We have developed a distributed genome alignment server in which multiple suffix trees can persist in memory. This overcomes the need to build a suffix tree for each alignment, allows multiple comparisons to be done simultaneously, and takes advantage of a computational cluster environment. Deltas are extracted from the alignment output and stored in a relational database or in a custom-designed storage system. SCANS (TM) was able to store, in a test case, 5,000 simulated chromosomes (approximately 1 Terabyte of total data) using less than 4 Gigabytes of physical space. The achieved compression ratio amounts to 0.034 bits/base - two orders of magnitude better than currently available nucleotide sequence compression systems. Our technology offers a space and time efficient alternative to traditional database and sequence alignment approaches.

---

### *A Sentence Recognizer for Mutant Protein Structure Studies: Toward Intelligent Systems for the Management of Structural Biology Data*

GREGORY CAPORASO, UNIVERSITY OF COLORADO HEALTH SCIENCES CENTER

Author(s): J. Gregory Caporaso, K. Bretonnel Cohen, Lawrence Hunter

We are developing an information extraction system to identify and compile results of site-specific mutagenesis structural biology studies. Structural analyses of site-specific mutants are often performed to determine the significance of individual amino acid residues in the higher order structures of proteins. The quantity of literature documenting these mutant protein structures is already overwhelming and rapidly increasing. An automated literature search for mutation/structural change (M/SC) mappings could populate and keep current a database of the results of these studies which would constitute a valuable asset in many areas of structural biology research. We are experimenting with various machine learning approaches to build a sentence classifier which identifies discussion of M/SC findings in PUBMED abstracts, and have received promising preliminary results using support vector machines. We present data on the performance of M/SC sentence classifiers constructed using varied feature types and classification algorithms, in addition to information on the growth of the structural biology literature base and the role of sentence classification in our larger goal.

*Layered Classification using Homologous Similarity and Error-Correcting Output Coding for Predicting Protein Subcellular Localization*

MARK DODERER, UNIVERSITY OF TEXAS AT SAN ANTONIO

Author(s): Mark Doderer, Stephen Kwek, John Salinas, Kihoon Yoon

Automated predicting of subcellular localization given the amino acid sequence is important, but made difficult due to the multiple destinations and the lack of strong sorting signals in the sequence. This paper presents a new subcellular localization prediction method that builds upon methods that have been previously studied. In previous methods, the localization of an unknown protein is very accurately deduced from known proteins based on homology analysis. However, this approach may not perform well for a newly discovered protein whose sequence is distinct from the known ones. Another approach is to use traditional machine learning techniques to predict a destination label based on the amino acid composition of the protein. Although it is normally not as accurate as finding a homologous match, this will provide accurate prediction for sequences with no homologous matches. In this paper we introduce a layered combination of these two different techniques. We are able to achieve very good accuracy for protein subcellular localization label assignment for 12 locations using a dataset with a combination of homologous and non-homologous proteins. Especially, Error Correcting Output Code (ECOC) is shown here as a better approach to deal with a high number of classes in a prediction problem.

---

*Visualizing Large, Multiscale Biological Systems*

AARON GABOW, UNIVERSITY OF COLORADO AT DENVER AND HEALTH SCIENCES CENTER

Author(s): Aaron Gabow, Lawrence Hunter

Biologists often visualize protein-protein interaction data by mapping proteins to nodes in a graph, interactions to edges, and use an automated graph layout algorithm to position the nodes. While this method can create picture that meets basic aesthetic concerns, it is not good at conveying information. The graph can have thousands of nodes—too many data points shown to effectively search for a particular node, and the graphic does not lend itself to an immediate summary or meaning. We have integrated several ideas from graph theory and graph visualization to create an application capable in presenting large-scale biological interaction data. Many biological data sets can be modeled as small-world networks, and often these ones are multi-scale when clustered. We use this hierarchical decomposition to assist in layout and to provide a representation that gives an effective overview of a graph while allowing the user to pick particular areas of interest.

## POSTERS

### *Unison: Integrated Feature-Based Mining for Target Discovery*

REECE HART, GENENTECH, INC.

Author(s): Reece Hart

Unison is a database of, and web interface to, precomputed sequence and structure predictions for a comprehensive set of protein sequences. The integration of these data enables the mining of sequences based on holistic protein feature criteria, the synthesis of predictions for individual sequence analysis, and the refinement of hypotheses regarding the composition of protein families. Unison includes prediction results for signal peptides, transmembrane domains, GPI anchoring, subcellular localization, secondary structure, sequence motifs, HMM, PSSM, and threading alignments, and genomic localization. SCOP, GO, HomoloGene, patent, and other auxiliary information permit richer queries and interpretations of prediction results. The PDB schema enables reliable structural localization of sequence features. Unison was designed to be kept up-to-date easily and the build process is automated. Sequence “run histories” enable incremental updates of precomputed results. The Unison schema, code, web interface, and non-proprietary data are released under the Academic Free License. They are available online and for local installation at <http://unison-db.org>.

We have used Unison for mining projects involving TNF ligands, helical cytokines, death domains, and other protein families. I will demonstrate Unison's utility by describing our search for proteins containing Immunoreceptor Tyrosine Inhibitory, Activation. and Switch Motifs (ITIMs, ITAMs, ITSMs).

---

### *Mathematical Techniques for Predicting and Analyzing Ontological Protein Function Annotations*

Cliff Joslyn, Los Alamos National Laboratory

Author(s): Cliff Joslyn, Karin Verspoor, Judith Cohn and Sue Mniszewski

The Protein Function Inference Group (PFIG) at the Los Alamos National Laboratory (LANL) has developed an approach to automatically produce novel Gene Ontology (GO) functional annotations of proteins based on categorizing the regions of the GO to which similar (in some sense) proteins are annotated. Current input spaces include proteins which are near neighbors in BLAST space, or which are described by similar terms occurring close together in documents. Validation of this or similar methods depends on the development of evaluation metrics which are appropriate to the mathematical structure of the space of annotations, in this case, the directed acyclic graph (DAG) structure of the GO. In this talk, we first outline the LANL architecture for ontological function annotation, implemented within our POSet Ontology Laboratory Environment (POSOLE), and for both the CASP and BioCreative evaluations. We present some test results from the BLAST-based effort. We then discuss our novel evaluation metrics, and

conclude with a consideration of their applicability to the general problem of measuring the consistency of annotation sets.

*Transcriptional Control and Behavioral Changes in Selectively Bred Mice*  
RAZVAN LAPADAT, UNIVERSITY OF COLORADO AT DENVER AND HEALTH SCIENCES CENTER

Author(s): Razvan Lapadat, Sanjiv Bhawe, Lawrence Hunter, Paula Hoffman, Boris Tabakoff

Cells must continually adapt to changing conditions by modifying their gene expression profiles. One of the core components involves transcriptional regulatory interactions. We focus transcriptional regulatory networks both because the increasing number of transcriptional profiling data sets, with “whole transcriptome” chips becoming the norm in the field and the fact that the process of gene expression can be regarded as the origin and effector of a response. In our strategy for the analysis of coregulated genes we use promoter sequence similarity searches in differentially expressed genes, cross species transcription factor mapping and in silico signaling pathway and literature mining. We have added to the analytical repertoire of our analytical techniques siRNA and miRNA binding prediction tools. Transcriptional profiles of whole brain extracts from mice selectively bred for ethanol preference (HIGH/LOW) or acute functional tolerance were measured using Affymetrix microarrays. Differentially expressed genes were analyzed using a positional scoring matrix algorithm for the first 2 kb upstream regions to identify conserved sequence patterns. Independently, the same regions were analyzed for the most conserved transcription factor binding sites based on a cross-species model. Resulting predictions were used together with the differentially expressed genes as input for signal transduction pathways and literature mining in order to establish a model of the interaction network modulating the gene expression events. The identified processes synthesize the integrated neuronal response of cells bearing different genotypic fingerprints, including signal transduction, transcriptional regulation, ion channel activity and neuronal activity modulation. Our work strongly suggests that combining transcription control module discovery with interaction network data mining represent a powerful approach for cis-regulation of gene expression and the involvement of signal transduction mechanisms using high-throughput techniques.

## POSTERS

### *RobustMap: A Fast and Robust Algorithm for Dimension Reduction and Clustering*

LIONEL LOVETT, JACKSON STATE UNIVERSITY

Author(s): Lionel F. Lovett, II

Medical databases, where 1-d objects (eg., ECGs), 2-d images (eg., X-rays) and 3-d images (eg., MRI brain scans) are stored can be very large due to the number of items and due to the number of attributes (high-dimensionality) associated with each item. Clustering reduces the number of items to their representative clusters and dimension reduction reduces the number of attributes. In addition, visualization of high-dimensional data requires reduction to lower-dimensional views that are often displayed as two or three dimensional plots. Traditional dimension reduction algorithms such as the singular value decomposition based principal components are computationally demanding and can be very slow. As the size of databases continues to grow, so does the demand for faster methods to visualize the data. RobustMap is a new, fast and robust dimension reduction method for high-dimensional datasets, which can separate outlying clusters from the main body of the data while computing a low-dimensional representation. It relies on stochastic concepts and on statistical distance distributions. The algorithm considers distance distributions from random and from extreme points to determine projection axes and clusters for dimension reduction. In determining the clusters, RobustMap focuses on the largest cluster, excluding outlying clusters. Along with visualization applications of this algorithm, the ability to quickly retrieve past cases with similar symptoms would be valuable for diagnosis, as well as for medical teaching and research purposes. Given the records of patients (with attributes like gender, age, blood-pressure etc.), RobustMap will detect any clusters, or correlations among symptoms, demographic data and diseases.

---

### *Bacterial Genes in a Eukaryotes*

Gayle Philip, NATIONAL UNIVERSITY OF IRELAND

Author(s): Gayle K. Philip and Dr. James O. McInerney

The malaria parasite (genus *Plasmodium*) is a unicellular eukaryote, which invades the erythrocytes of its vertebrate host through the course of a complex life cycle. The disease is estimated to give rise to 270-515 million clinical cases each year with 1-2.7 million deaths, mainly attributable to *Plasmodium falciparum*. It was our objective to identify the origin of each of the 5,295 *P. falciparum* protein-coding genes and in particular, cases where the nearest neighbour to the *P. falciparum* protein was a prokaryotic sequence.

Homologues were identified by performing a FASTA search of a sequence library set of 867,899 proteins made up from 20 Archaeal, 179 Bacterial and 25 Eukaryotic completed genomes. Phylogenetic trees were then reconstructed for each of the *P. falciparum* genes and were manually examined. A number of genes were found to

be candidates for having undergone a lateral gene transfer event. In particular, 35 genes were identified where a query *P. falciparum* gene was found to have homologues from bacterial genomes, but not from any archaeal or other eukaryotic genomes. 16 of these Plasmodium-bacterial specific genes have a function that is unknown, while the remaining genes are involved in different pathways including pyrimidine metabolism and fatty acid biosynthesis.

---

### *Semi-Automation of Hydrogen Exchange Data Analysis*

**THAMI RACHIDI, UNIVERSITY OF COLORADO, BOULDER**

Author(s): Thami Rachidi, Thomas Lee, Katheryn A. Resing, Natalie G. Ahn, Krzysztof J. Cios

Hydrogen exchange mass spectrometry (HX-MS) is a method that can monitor exchange between protons in protein backbone amides and deuterons from solvent. Proteolysis with pepsin enables measurement of HX within localized regions of proteins. A limiting step in this experiment is the time needed for data analysis, where  $m/z$  and intensity information are extracted from MS data and used to calculate the weighted average mass (WAM) for each ion. Contaminating peaks often interfere with accurate estimates of WAM, thus requiring significant manual analysis. A new software tool, named HX-Analyzer, is being developed to semi-automate steps in data analysis. The tool inputs a listing of MS files and a spreadsheet summarizing information about peptides of interest. It automatically produces extract ion chromatograms, then displays the corresponding spectra and allows the user to choose peaks for extraction of  $m/z$  and intensities. WAM values are calculated and saved to a spreadsheet format. HX-Analyzer allows the user to exclude contaminant peaks from the peak list to increase accuracy for WAM calculations. Data collected with an ABI QStar Pulsar QqTOF instrument was examined using HX-Analyzer. Preliminary results show a significant improvement in accuracy as well as ~20-fold reduction in time needed for data analysis.

---

### *Dynamic Gene Annotation and Analysis at PlantGDB*

**SHANNON SCHLUETER, IOWA STATE UNIVERSITY**

Author(s): Shannon D Schlueter, Matthew D Wilkerson, Qunfeng Dong, Volker Brendel

The intricacies of gene annotation are among the most complex problems being addressed by modern genomics. Albeit the current methods of gene structure annotation are vastly more accurate and provide more complete coverage than those employed less than five years ago, the support of annotations on a per-gene basis differs extraordinarily. The level of confidence for individual gene annotations can be directly attributed to the verifying presence of expressed sequence alignments. The dependence of gene structure annotation on available EST and cDNA sequences makes all static estimations of gene structure problematic. For

## POSTERS

this reason, methods of maintaining gene annotation must acknowledge confidence in a predicted gene structure. These methods must also incorporate dynamic reporting of the evidence supporting each annotated property. To provide confidence estimators and evidence reporting for current gene annotations we developed GAEVAL, the Genome Annotation Evaluation Algorithm. This system has been integrated with the existing xGDB genome data browser and tools for community curation of gene annotations at [www.PlantGDB.org](http://www.PlantGDB.org).

---

### *Integration of Expression Data and Transcriptional Control Network: Significant Regulators Driving Expression Changes*

**ANDREY SIVACHENKO, ARIADNE GENOMICS, INC.**

Author(s): A. Y. Sivachenko, A. Yuryev, N. Daraselia, I. Mazo

Microarrays provide an invaluable insight into the biomolecular mechanisms, however raw results are disjoint genome-wide “one-gene-at a time” datasets with high levels of noise. Placing data into the biological context through integration with different data sources is critical both for noise reduction and for objectively quantifiable system-level hypothesis formulation. We analyze differential expression (DE) data in the context of large network of known transcription regulation events. DE data sample downstream of a regulator is compared to the sampling distribution derived from the network, with network connectivity taken into account. The analysis is aimed at elucidating regulators with statistically significant patterns of downstream expression changes and explaining DE data in terms of activated/suppressed regulatory cascades. The set of plausible regulatory events provides conceptual data reduction and a step towards elucidating/building extended pathways. We apply our analysis to a few disease datasets, demonstrate robustness and statistical significance of the results, and show that the sets of regulators suggested as putatively involved in the differential response are potentially interesting biologically and exhibit statistically significant overlap with sets of known disease associated genes. Assembling significant regulators into a putative signaling pathway and applications of our procedure to other networks (metabolic, binding) are also discussed.

*Bayesian Inference in Simple Visual Perception***MOURAD TADE SOUAIAIA, MARYMOUNT UNIVERSITY**

Author(s): Geoff Ghose , Mourad Tade Souaiaia

Perceptual Bayesian Inference states that probability distributions are built up and used as degrees of belief when making a perceptual decision. Thus past experiences influence the detection of stimuli by skewing perception toward stimuli values that have been observed most. Bayesian-like perception has been observed in 3-D shape perception and reaching tasks. If Bayesian inference applies to simple stimuli discrimination, distribution of observed stimuli will influence the performance of visual discrimination by skewing perception to the values where the distribution is greatest.

Here we test whether Bayesian processes might be involved in such low-level perception by asking subjects to perform an orientation discrimination task in which we bias the distribution of orientations that are presented. We build up a hypothesis by modeling predicted performance for two distributions that are different in shape, but identical in mean and variance. We build up the model using Bayes theorem, where the performance is predicted by multiplying the conditional probability with the likelihood function, or the normally distributed firing of neurons at a given stimuli multiplied by the distribution of stimuli. In Matlab we obtain performance curves for two different distributions.

Our data indicates that subject's performance is affected by probability distribution in a manner consistent with our models. When all the subjects' data is averaged we observe a performance curve similar to our model. Also every subject does perform better in discriminating difficult stimuli when using a two peaked distribution (Fisher test  $p=.10$ ) and there is no significant difference when the stimuli is not difficult. This is consistent with our model, because the two peaked distribution will skew difficult stimuli to a detectable range, while the normal distribution will skew otherwise detectable stimuli to non-detectable levels. These results indicate that subjects can use prior experience to make perceptual judgments of low level stimuli, and suggest that perceptual capabilities can be improved by adopting particular distributions.

## POSTERS

### *Whole Genome Transcript Analysis with Affymetrix Exon Microarrays*

CHUCK SUGNET, AFFYMETRIX

Author(s): Charles Sugnet, Alan Williams, Yaron Turpaz, Jim Veitch, Tyson Clark, Anthony Schweitzer, Melissa Cline, Hui Wang, Raymond Wheeler, John Blume

Ongoing technology developments have enabled the tiling of over 6 million probes on a single Affymetrix oligo microarray. This technology push has enabled the development of commercial whole genome exon arrays. Current exon microarray implementations consist of a single microarray covering roughly 1 million exons with 1.4 million probesets consisting of 4 perfect match probes per probeset for the human genome. Whole genome exon microarrays present researchers with additional insight into transcriptional complexity, such as alternative splicing, but also raise the need for improved and new analysis algorithms and improved computational efficiency. We have developed new algorithms for identifying alternative spliced exons based on ANOVA. We have applied these algorithms to a publicly available 11 tissue data set (3 replicates each) and 7 paired colon normal/cancer tissues data sets to identify exons whose splicing is differentially regulated.

---

### *Determination and Analysis of Genes Involved in the Cleft Lip/Palate Defect*

KEMENI TENKU, UNIVERSITY OF COLORADO, DENVER

Author(s): Kementi Tenku, Tzu Pahng, Susan Trapp, Trevor Williams, Lawrence Hunter

Birth defects affect approximately 5% of all infants in the USA. The cleft lip and/or palate (CL/P) is one of these defects and it affects roughly 1/1100 and 1/1600 births, respectively. The “non-syndromatic” CL/P (nsCL/P) accounts for ~70% of all cases. This is a non-Mendelian multifactorial disorder due to interaction of multiple genes and environmental factors during fetal craniofacial development. Herein we present data-mining results of microarray data from a mouse model. Three tissues—frontonasal, lateral nasal and maxillary prominences—during critical periods of orofacial developments (ED 10.0 to ED 12.5) were selectively examined, since this is the period relevant to the development of the orofacial cleft. By normalizing, filtering and using different statistical multiple comparison correction test methods on the data, the differences in gene expressions were evaluated and gene lists of statistically significant genes, potentially contributing to the defect, were produced. Through programming tools, the most up- and down-regulated genes at the different developmental stages and locations were selected. These genes were analyzed via Onto-Express, a Gene Ontology analysis tool, to determine their putative functions. Preliminary results demonstrated that DNA binding, protein binding, transcription factor activity and structural molecule activity are involved in the process of craniofacial development.

*Protdist: An Analysis of Error***CHAD WAGNER, SAN DIEGO STATE UNIVERSITY**

Author(s): Chad Wagner, Anna Salamon, Pat McNairnie, Rob Edwards, Peter Salamon

The present study proposes new data-based methods predicated on the assumption of approximate ultrametricity for estimating the accuracy of phylogenetic distance measures for particular sets of proteins. Using a database of over 19,000 phage proteins, we find good validity for approximate ultrametricity up to a PROTDIST value of about 1.5 where the behavior makes a clear transition. The structure is more evident using pairwise alignments than multiple alignments and over 640,000 pairs of proteins were aligned in a pairwise manner as part of this study. We present several ways of seeing the ultrametricity and the transition around the PROTDIST score of 1.5. Our findings demonstrate the utility of these methods for estimating the accuracy of PROTDIST for phage proteins at different distances. The implications for phylogenetic inference are considered.

*Integrating Statistical Analysis of Gene Expression Data onto Metabolic Pathways Facilitates Understanding of Gene Expression in the Metabolic Context***BART WEIMER, UTAH STATE UNIVERSITY**

Author(s): Jake Michaelson, Balasubramanian Ganesan, Jon L. Pearson, Dong Chen and Bart C. Weimer

The majority of the genes within a microbial genome are linked to metabolic reactions important in intermediary metabolism and survival. The understanding of gene expression profiles and regulation is greatly facilitated by appropriate statistical analysis in concordance with methods to display the data. The display of gene expression data over time from a time series experiment is critical for the biologist to quickly visualize the directions in which whole pathways progress in response to time and other physiological factors such as metabolite concentration or stress. Physical mapping of expression data to pathways is non-trivial and is extremely time consuming. Here we describe the analytical tools built from resources available in the public domain to meaningfully depict pathways and gene expression data in concordance with the proper statistical analysis. Tools for statistical analysis of gene expression using repeated measures were developed using Bioconductor. Interfaces were created that are accessible through the Apple Bioinformatics cluster server at the Center for Integrated BioSystems. Bioconductor was also used to draw gene expression maps that were overlaid to pathways from Pathway Tools using Perl scripts to integrate the pathways and heat maps in a single visualization file. Differential color display of the gene labels was used to show genes that significantly changed over time. We used this tool to

## POSTERS

greatly facilitate analyzing, displaying, manipulating, and understanding microarray data more conveniently for the biologist to make metabolic conclusions quickly.

---

*Analysis of Human Promoters and Gene Expressions by an Integrative Approach: Constructing an Index Toward Gene Expression Patterns*

**KIHOON YOON, THE UNIVERSITY OF TEXAS AT SAN ANTONIO**

Author(s): Kihoon Yoon, Stephen Kwek

Identification of gene controlling elements in human is fundamental to the understanding of the mechanisms of diseases. Here, we present an integrative analysis of promoter sequences and gene expressions of normal human tissues to create a promoter complexity index (PCI) as the primary indications of tissue specificities and expression levels. To achieve this goal, our approach must be sensitive enough to detect subtle differences in the controlling regions. We applied a new sequence signal detection algorithm to promoter and downstream regions of transcription start sites (TSSs). Our approach considers two cases that typical pattern finding algorithms may not be able to handle, (1) patterns reside on non-fixed positions relative to TSSs, but yet the regions are limited to  $\sim 30$  bp and (2) rare pattern signals which may be ignored easily by “over representation”-based methods. The patterns found were further refined by integrating the mRNA expression profiles of normal human tissues to minimize possible false positive pattern detections. We are also currently developing a better way of gene expression analysis schemes to draw more meaningful co-expression information. In summary, we have identified unique sequence patterns from the promoters of housekeeping and tissue-specific genes which may reflect different gene controlling mechanisms.

---

*Inferring Three-way Gene Interactions from Microarray Data Sets*

**JIEXIN ZHANG, UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER**

Author(s): Jiexin Zhang, Yuan Ji, Li Zhang

It has been an important and challenging problem to infer the network of gene interactions from microarray data. Conventional methods use correlation of expression profiles between two genes to look for signs of co-expression. However, patterns of co-expression are often obscured because they change depending on biological conditions such as tissue types, or diseases. In this study, we used a large microarray dataset of various human cancers to survey for three-way gene interactions, in which co-expression of two genes depends on the expression level of a third gene. Such three-body interactions cannot be derived from two-body interactions based on pair-wise correlations. We used a model-based clustering algorithm to identify genes with bimodal expression profiles, and partitioned the samples accordingly. We then identified the gene pairs of which correlation of expression changed significantly between the two partitions of samples. To perform cross

validation, we randomly split our collection of 545 samples into a training-set of 360 and testing-set of 185. Our survey found ~83000 significant gene triplets (permutation test p-values  $< 10^{-9}$  in the training-set, of which 61% have p-values  $< 10^{-6}$  in the testing-set). Our results may prove valuable in the construction of complex gene networks.

## SPONSORS

### *Affymetrix*

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



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).

---

### *AMD*

90 Central Street; Boxboro, MA 01719  
url: [www.amd.com/lifesciences](http://www.amd.com/lifesciences); phone: 508-733-7391



AMD (NYSE: AMD) designs and produces innovative microprocessors and low-power processor solutions for the computer and communications industries. The company is the developer of AMD64 technology and AMD Opteron™ processor. AMD has conducted a number of computational biology performance studies that can be found at [www.amd.com/lifesciences](http://www.amd.com/lifesciences)

---

### *Apple Computer*

1 Infinite Loop; Cupertino, CA 95014  
url: [www.apple.com](http://www.apple.com); phone: 303-471-1575



Apple ignited the personal computer revolution in the 1970s with the Apple II and reinvented the personal computer in the 1980s with the Macintosh. Today, with an industrial-strength, UNIX-based operating system, the power and precision of 64-bit computing and seamless integration of hardware and software, the Apple platform is the ideal solution for science.

---

### *Ariadne Genomics, Inc.*

9700 Great Seneca Highway; Rockville MD 20850  
url: [www.ariadnegenomics.com](http://www.ariadnegenomics.com); phone: 240-353-5707



Ariadne Genomics develops user-friendly software tools for biologists in the areas of pathway analysis and systems biology. Ariadne Genomics products incorporate proprietary Natural Language Processing (NLP) and statistical algorithms designed to functionally interpret novel genetic information.

---

### *Cancer Informatics*

Libertas Academica Ltd; 2 Rimu Rise; Albany;  
Auckland, New Zealand; 1311  
url: [www.la-press.com/caninfo.htm](http://www.la-press.com/caninfo.htm); phone: 649-415-7704

**Cancer Informatics**

Libertas Academica Ltd brings the high standards of conventional journal publishing to Open Access electronic journals.

## SPONSORS

### *Dharmacon, Inc.*

2650 Crescent Drive, Suite 100; Lafayette, CO 80026  
url: [www.dharmacon.com](http://www.dharmacon.com); phone: 800-235-9880



**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.

### *Exagen Diagnostics, Inc.*

801 University Blvd SE, Suite 209; Albuquerque, NM 87106  
url: [www.exagendiagnosics.com](http://www.exagendiagnosics.com); phone: 505.272.7966



Exagen Diagnostics focuses on the rapid identification, validation and commercialization of genomic marker sets that advance the business/scientific objectives for pharmaceutical clinical trials or expand diagnostic testing to guide physicians in treating individual patients. Proprietary, in-silico discovery technology enables the identification of these small (3-4 gene) marker sets.

### *IBM*

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



IBM Deep Computing and IBM Healthcare and Life Sciences are delivering innovative and powerful breakthrough solutions to address the demands of intense computation, visualization, or manipulation and management of massive amounts of data for a variety of areas including health care and the life sciences.

### *PloS*

185 Berry Street, Suite 3100; San Francisco, CA 94107  
url: [www.ploscompbiol.org](http://www.ploscompbiol.org); phone: 415-624-1223



PLOS Computational Biology ([www.ploscompbiol.org](http://www.ploscompbiol.org)) features works of exceptional significance that further our understanding of living systems at all scales through the application of computational methods. A leader in the international open-access movement, PLoS publishes peer-reviewed science and medical journals that are available on the Internet at no charge to users.

## SPONSORS

*John Wiley & Sons, Inc.*

111 River Street; Hoboken, NJ 07030  
url: [www.wiley.com](http://www.wiley.com); phone: 877-762-2974



Founded in 1807, John Wiley & Sons, Inc., provides must-have content and services to customers worldwide. Its core businesses include scientific, technical, and medical journals, encyclopedias, books, and online products and services; professional and consumer books and subscription services; and educational materials for undergraduate and graduate students and lifelong learners.



**NOTES**

# NOTES



Printing of this program was sponsored by Ariadne Genomics, Inc.

---

Rocky'05 is a regional conference of the  
International Society for Computational Biology

