



POSTER ABSTRACTS – *updated 11/13/08*

Title: *Shannon's Uncertainty and Kullback-Leibler Divergence in Microbial Genome and Metagenome Sequences*

Presenter: Sajia Akhter, San Diego State University

Authors: Sajia Akhter, Robert Edwards

Abstract: All genome sequence data contains inherent information in it. Shannon's uncertainty theory can be used to measure how much information a sequence has. Here we show that the amount of information in sequences from metagenomes correlates with the number of similar sequences that will be found by comparison to databases of known sequences. Hence, a sequence with more information (higher uncertainty) has a higher probability of being significantly similar to other sequences in the database. Measuring uncertainty maybe a rapid way to screen for sequences likely to be similar to things in the database, to prioritize assignment of computational resources, and to show which sequences with no known similarities are likely to be false negatives. To predict which sequences could be coding based purely on the information content in them, we compared the uncertainty of intergenic and protein coding regions for complete bacterial genomes. The intergenic region was more likely to have higher uncertainty, but was not predictive of the coding potential of short sequences. Since uncertainty could predict useful short sequences; we could divide the long sequences in small fragments (100 bp) and measure the uncertainty; then compare the consecutive uncertainty to predict the useful portion of sequences. Amino acid content in the genome may reflect lifestyle restrictions of an organism, and may also be predictive of coding potential. To compare the amino acid composition for each of the complete bacterial genome sequences we calculated the Kullback-Leibler divergence from the mean amino acid content. We demonstrate that (i) there is a significant difference between amino acid utilization in different phylogenetic groups of bacteria; (ii) that the bacteria with the most skewed amino acid utilization profile are endosymbionts or intracellular pathogens; (iii) the skews are not restricted to one or a few metabolic processes but are across all subsystems; (iv) amino acid utilization profile are strongly correlate with GC percent.

Title: *A Comparison of Next-state (dynamic) and Co-temporal (static) Modeling of Time Course Data: Multiple Approaches Provide Complementary Information*

Presenter: Edward E. Allen, Wake Forest University

Authors: Edward E. Allen, David J. John, William H. Turkett, Stan J. Thomas, Richard F. Loeser, Leslie B. Poole, Elizabeth M. Hiltbold, Jacquelyn S. Fetrow

Abstract: Determining networks, or relationships between biological entities (genes or proteins), that underlie experimental data is a major, unsolved problem in modern biology. An important issue is a consideration of how the data are used in the modeling process and how the models are ultimately interpreted. Often co-temporal (i.e., correlative, static or invariant) modeling techniques are applied to the data, but the models themselves are interpreted in a next-state (i.e., dynamic) context. In this poster, the mathematics of next-state and co-temporal approaches are compared and contrasted. Both techniques were implemented in a computational algebra modeling approach and the two algorithms applied to several time course experimental data sets, including protein modification (Western blot) and gene expression (microarray) data. Comparison of the models shows that the information extracted by the two algorithms (i.e., the relationships between genes or proteins, as specified by model edges) is different yet complementary. Comparison of the models to previously published data and established networks suggests the need for different biological interpretation of these next-state and co-temporal relationships. Overall, analysis of the results suggests that the application of both modeling approaches may be useful in identifying the full set of relationships or networks that underlie experimental time course data.

Title: *Computer-based Determination of ETEC Pathotype-specific Genes to Discover Targets for Molecular Diagnosis and Reverse Vaccinology*

Presenter: Ramy Aziz, Cairo University

Authors: Heba M. Amin, Abel-Gawad M. Hashem, Ramy K. Aziz

Abstract: Members of the versatile enterobacterial species, *Escherichia coli*, normally colonize the mammalian colon but readily cause a wide range of intestinal and extraintestinal diseases. These diseases range from mild intestinal disturbance to urinary tract infections and may also occur in outbreaks of shigellosis-like dysentery and cholera-like watery diarrhea. In particular, enterotoxigenic *E. coli* (ETEC) claims 380,000 lives every year, mostly children, and is on the top of the World Health Organization's list of deadly infectious agents in the world. In Egypt, ETEC poses a serious public health threat, especially in rural areas, where vaccination is not available to most people at risk. Developing affordable molecular diagnostic tools as well as ETEC-specific vaccines can be greatly accelerated by pathogenomic analysis of all *E. coli* strains, i.e., by the determination of genomic regions associated with the different pathotypes. *E. coli* genomes are known for their high plasticity to the extent that two *E. coli* strains may differ by as much as 25% of their genomes. This inter-pathotypic plasticity is exhibited as gene gain or loss and is mainly driven by pathoadaptive mutations, genetic

rearrangements, and horizontal gene transfer. E. coli strains are often polylysogenic, with multiple complete or rudimentary prophages scattered within their genomes. For the reasons mentioned above, we undertook the current study aiming to identify horizontally transferred, pathotype-associated E. coli genes, with emphasis on those encoding virulence proteins or exotoxins. To estimate the extent of horizontal gene transfer in ETEC, we used a combination of computational tools, including comparative genomic analysis (comparing dinucleotide skew patterns in microbial genome), GC% analysis, and web-based IslandPath analysis (<http://www.pathogenomics.sfu.ca/islandpath>). To determine ETEC pathotype-specific genes or signature genes, we used comparative genomic tools available in the National Microbial Pathogen Data Resource (NMPDR) platform (<http://www.nmpdr.org>), including the Signature Genes Tool and the Homolog Spreadsheet Tool. We identified 451 genes that differentiate this pathotype from other E. coli strains, based on bidirectional-best-hit signature analysis. We also identified 181 genes that are characteristic to two closely related genomes (24377A and 2348/69). Most of the ETEC-specific genes were mapped to prophages, prophage-like elements, and other pathogenicity islands. However, some of the signature genes, e.g., ORFs 2760-2765, seem to be rather lost in other E. coli strains (due to their conservation among Enterobacteriaceae, e.g., Shigella and Salmonella). Our ongoing studies are testing some of these ETEC-specific genes as targets for multiplex PCR amplification to develop a rapid diagnostic typing method. Future studies will analyze the surface-association and antigenicity of these signature genes products as a first step in a reverse vaccinology strategy to develop novel ETEC vaccines.

Title: *Extended Analysis and Prediction of Gain and Loss of Phosphorylation Sites in Cancer*

Presenter: Peter H. Baenziger, Indiana University School of Medicine

Authors: Peter H. Baenziger, Maricel G. Kann, Matthew Hahn, Matthew E. Mort, Predrag Radivojac, Sean D. Mooney

Abstract: Coding region mutations in human genes are responsible for a diverse spectrum of diseases and phenotypes. Among lesions that have been studied extensively, we have insights into several of the biochemical functions disease-causing mutations disrupt. Here we extend our investigation of the role of phosphorylation in somatic cancer mutations and inherited diseases. Previously we showed that somatic cancer mutation datasets had a significant enrichment for mutations that cause gain or loss of phosphorylation when compared to our control datasets (putatively neutral nsSNPs and random nsSNPs). Of the somatic cancer mutations, those in kinase genes represent the most enriched set of mutations that disrupt phosphorylation sites, suggesting phosphorylation target site mutation is an active cause of phosphorylation deregulation. We extend our work to identify the most likely candidates for gain or loss of a phosphorylation site in inherited disease, and we apply Ingenuity Pathway Analysis (IPA) and ScanSite to identify pathways and specific kinase targets likely to be disrupted.

Title: *Analyses of Features for Prediction of Protein-protein Interactions in Protein Hetero-complexes and Their Impact on Human Disease*

Presenter: Angshuman Bagchi, Indiana University School of Medicine

Authors: Angshuman Bagchi, Eunseog Youn, Sean D. Mooney

Abstract: Analyses of features for prediction of protein-protein interactions in protein hetero-complexes and their impact on human disease Angshuman Bagchi¹, Eunseog Youn², Sean D. Mooney^{1*} ¹Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine, 410 W. 10th Street, Suite 5000, Indianapolis, IN 46202 ²Department of Computer Science, Texas Tech University, Lubbock, Texas 79409-3104 Email: angshuman@compbio.iupui.edu Email: eun.youn@ttu.edu Email: mooney@compbio.iupui.edu Many proteins function via inter-protein interactions. Proteins may interact for over many time-scales from nascent to obligate. Understanding these interactions is crucial in the elucidation of biological pathways and cellular functions, and these processes have therapeutic applications. In the present study we have extracted features from protein sequences, three dimensional structures and structural environments using the non-redundant chains of protein hetero-complexes from protein data bank (PDB) and evaluated those features to discriminate between interacting and non-interacting protein residues. The features have been ranked based on their area under the receiver operating characteristics (ROC) curve i.e, the AUC values. As has been found in other studies, the most important features are based on sequence conservation. The discriminatory features have been used to train and evaluate a support vector machine (SVM) using 10 fold cross-validation. The accuracy of the prediction is found be 77%. We are using this approach to analyse disease mutations. It is well known that disease mutations affect protein-protein interaction interfaces. So, the features that are used to discriminate between protein binding and non-binding residues may also serve as good indicators of disease mutations that disrupt the binding abilities of the proteins in disease situations. *Author for correspondence

Title: *Discovery of HuR Binding Sites Using a Machine Learning Approach*

Presenter: Shweta Bhandare, University of Colorado Boulder

Authors: Jon Miller, Shweta Bhandare, Paul Johnson, Nick Farina, Debra Goldberg

Abstract: HuR is a ubiquitous mRNA binding protein that leads to mRNA stabilization and translation. Large scale assays combined with sequence and secondary structure analysis has yielded a probabilistic binding motif for HuR. While the binding motif has been identified, it is unclear what effect several other variables have on influencing HuR binding. These variables include the presence of multiple HuR motifs in a single mRNA, the location of the motif in the mRNA and the U-richness of the flanking regions. Additionally, motif searching alone is not a particularly effective method for identifying binding targets because the motif is short and allows for extensive variation at many

nucleotide positions. We have developed a machine learning model combining motif searching and machine learning that greatly improves the accuracy of predicting HuR binding targets. We trained our machine-learning model using the Lal et al. immunoprecipitation (IP) assay data. IP is an antibody technique to determine what a protein of interest binds to. This data-set provides a list of candidate transcripts that are positive and negative HuR binders. Existing tools were used to search the transcripts from this data-set, and the results of the search were used to generate a set of independent variables for each transcript. These variables were fed into a machine learning algorithm to generate a model for predicting HuR binding in future transcripts. To generate the model, we began by separating our data into two sets, a training set and a test set, each containing one-half the HuR binding transcripts and one-half the non-HuR binding transcripts. The training set was used to build a machine learning model for classification. Next, we used the model to classify the transcripts in the test-set as binders or non-binders. We assessed the accuracy of the model's predictions and compared them to the accuracy of the motif-searching process, if even one motif was found in the section of the transcript under consideration (whole sequence or 3' UTR), then we labeled that transcript an HuR binder. Otherwise, we labeled it a non-binder. Our model gives a total accuracy of 78% while motif searching with the whole sequence gives 22% accuracy and with the 3' UTR sequence gives 50% accuracy. The test-set contains many transcripts that were not HuR-binders but did contain the HuR binding motif. Our model accurately predicts that many of these transcripts are not HuR binders, and has a false positive rate of 10% compared to 93% for motif searching on the whole sequence and 41% for motif searching on the 3' UTR sequence. Machine learning has been used for binding prediction in a variety of ways; however, the attempt made in this project to use it as a secondary step to motif searching appears to be novel. Our focus on HuR, which is not an extensively studied RNA-binding protein, also makes this a unique project.

Title: *An Empirical Evaluation of Relevance Vector Machine for Gene-expression-based Cancer Diagnostics*

Presenter: Ljubomir Buturovic, Pathwork Diagnostics, Inc.

Authors: Ljubomir Buturovic, Shibin Qiu

Abstract: An empirical evaluation of Relevance Vector Machine for gene-expression-based cancer diagnostics Ljubomir J. Buturovic and Shibin Qiu Pathwork Diagnostics Inc., Sunnyvale, CA, USA Cancer diagnostics using gene expression faces several challenges including feature selection from high dimensional data, high accuracy multiclass classification, and the generation of predictive probabilities. Although Support Vector Machine algorithm (SVM) has achieved fairly high accuracy, it has shortcomings in other aspects of the requirements. Especially, its' simulated probability scores become less discriminative as the number of classes get larger. The Relevance Vector Machine (RVM) has several attractive properties: sparse solutions (implying good control of overfitting), reliable estimates of posterior probabilities of class membership, automatic determination of the regularization parameter, relaxed kernel function requirements, and straightforward generalization to multiclass cases [1]. It has achieved promising results on practical data sets [1], [2], [3]. We therefore applied RVM to clinically relevant

classification problems in distinguishing colorectal, gastric, and pancreatic cancer types, as well as predicting breast cancer recurrence. Our results indicate that RVM exhibits inferior performance compared with more established methods in the field, such as neural networks, SVM, and LASSO. We hypothesize that this may be a result of the evidence approximation of the predictive distribution, in which model hyperparameters are fixed to the values maximizing the marginal likelihood function, or else due to convergence issues in optimization. In conclusion, our empirical findings thus far do not support the use of RVM in the present form in diagnostic tests based on analysis of gene expression profiles. [1] M. E. Tipping, Sparse Bayesian Learning and the Relevance Vector Machine, *Journal of Machine Learning Research*, vol. 1, pp. 211-244, 2001. [2] G. C. Cawley and N. L. C. Talbot, Gene selection in cancer classification using sparse logistic regression with Bayesian regularization, *Bioinformatics*, vol. 22, No. 19, pp. 2348-2355, 2006. [3] C. M. Bishop, *Pattern Recognition and Machine Learning: 7.2 Relevance Vector Machines*. Springer, New York, 2006.

Title: *Sequence Cooccurrence and Covariation Suggest Specific Physical Interactions Between Type VI Secretion System Components*

Presenter: J. Gregory Caporaso, University of Colorado Denver

Authors: J. Gregory Caporaso, Amy Dear, Larry Hunter, Marcelo Sousa, and Rob Knight

Abstract: The Type VI Secretion System (T6SS) is a recently discovered protein complex used by pathogenic bacteria to transfer effector proteins from the pathogen's cytosol to the host's cytosol. Twenty-four *Salmonella typhimurium* proteins are potentially involved in the T6SS apparatus, but the specific protein-protein interactions important for complex formation are currently unknown. A brute-force approach to identify the interactions biochemically would involve testing 276 pairs of proteins for physical interactions: an expensive and time-consuming task. We have identified proteins that cooccur in bacterial gene clusters, and applied joint-entropy-normalized mutual information (NMI) to detect covarying positions between multiple sequence alignments of cooccurring proteins. Covariation between proteins is taken as evidence of physical interaction between those proteins, and covariation results are therefore applied to guide the biochemical analysis of the T6SS apparatus by predicting the pairs of proteins most likely to interact. Of the twenty-four *Salmonella typhimurium* proteins, thirteen were found to frequently cooccur in bacterial gene clusters. Our covariation analysis focused on these thirteen proteins, and predicted sixteen pairs of proteins (involving ten of the thirteen cooccurring proteins) to physically interact. Of these sixteen predictions, one has been biochemically confirmed; an additional six appear to be correct based on current models of the T6SS; eight involved predicted cytoplasmic proteins that have not been previously characterized; and one prediction is inconsistent with the current model. This work has shed light on the organization of the T6SS complex, and all sixteen predictions are currently being evaluated biochemically. Our approach to identifying protein-protein interactions on the basis of cooccurrence and covariation is easily expandable to other sets of proteins, and therefore has potential to greatly reduce the cost and labor involved in protein-protein interaction studies.

Title: *Natbox: A Network Analysis Toolbox in R*

Presenter: Shweta S. Chavan UALR / UAMS Joint Bioinformatics program

Authors: Shweta S. Chavan, Radhakrishnan Nagarajan

Abstract: Classical biological paradigms have focused on understanding changes in single genes across distinct biological states. Such analysis while useful may not provide sufficient insight into their interactions or functional relationships (FRs). Understanding FRs is crucial, since genes work in concert as a system as opposed to independent entities. On a related note, phenotype formation is mediated by pathways consisting of complex interactions between several genes and not a single gene. Commercial packages (Ingenuity Pathway Analysis, Ingenuity Systems, CA) and Pathway Studio (Ariadne Genomics, Rockville, MD) provide menu-driven graphical user interface (GUI) for retrieving FRs from documented literature on all possible functional relationships on a given set of genes. However, recent studies have provided compelling evidence of non-canonical signaling mechanism that demand inferring network structure from the given data as opposed to direct inference from classical documented pathways. Bayesian structure learning (BSL) (Pearl, 2000, Friedman, 2000) techniques infer the interactions between the given set of genes in the form of acyclic graphs have been successfully used for inferring transcriptional and protein networks. The inherent probabilistic nature of gene expression and access to high-throughput assays that facilitate simultaneous measurement of transcriptional, translational and post-translational activities are some of the reasons for wide-spread use of BSL. Probabilistic nature of gene expression can be attributed to inherent noisiness at transcriptional, translation levels and heterogeneity within/between cell population(s). We have developed an integrative environment Network Analysis Toolbox (NATbox) that integrates functions across existing packages in the open-source language R and provides a convenient GUI for (i) modeling FRs and network structure from gene expression profiles using correlation metrics and a battery of BSL techniques from the R-package bnlearn. Subsequently, identify the robust dependencies using bootstrap procedure and highlight the same in the resulting acyclic graph for visualization. (menu: Bayesian Network) (ii) investigate topological properties of the resulting network using social network analysis metrics from the R-package (sna) (menu: SNA) (iii) retrieve published literature from PUBMED corresponding to functional relationships of interest. (menu: Mining FRs) (iv) impute missing values in incomplete data using the R-package impute.knn (menu: Impute). Each menu functions independently. This flexibility allows NATbox to be used across distinct paradigms and permit addition of new functionalities. NATbox was funded by the grant R03LM008853 to RN. References: Pearl, J (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press. R Development Core Team (2008) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Scutari, M. (2008) bnlearn: Bayesian network structure learning, Version 0.8. Friedman, N. et al. (2000) Using Bayesian Network to Analyze Expression Data. *J. Computational Biology*. 7,601-620. Butts, C.T. (2007) sna: Tools for Social Network Analysis, Version 1.5.

Title: *Statistical Data Mining and Its Applications to Microarray Analysis*

Presenter: Argon Chen, National Taiwan University

Authors: Argon Chen

Abstract: In this talk, the similarities and differences between Statistical Inference and Data Mining will be first discussed. A term “statistical data mining” is then defined. Typical statistical analysis techniques, such as regression analysis/tree, PCA/FA, CCA/PLS, Clustering, and Discriminant Analysis, will be manifested using plain language and examples of applications to microarray analysis. The applications will include microarray analyses for diagnoses and prognoses of liver cancers and breast cancers. Finally, new advances in nonlinear multivariate analysis techniques will be briefly introduced in this talk.

Title: *BIOGRAM: Algorithms for Identifying "Similar" Proteins by Functional Annotation*

Presenter: Judith Cohn, Los Alamos National Laboratory

Authors: Judith D. Cohn, Susan M. Mniszewski, Jennifer F. Harris, Hong Cai, and Ruy M. Rubeiro

Abstract: There are a wide range of methods for identifying "similar" proteins, the most common being BLAST, which is based on sequence homology. In many cases, however, it is useful to find proteins which exhibit a similar functional signature within a cell or organism but do not necessarily have homologous sequences or share any a single biological feature. We present BIOGRAM (BIOlogical Graphical Measurement), a software system under development, which uses the mathematical structure underlying the Gene Ontology (GO) to measure protein similarity based on annotations assigned to nodes in the three branches of the GO: Biological Process, Molecular Function, and Cellular Component. This system is built upon the POSOLE (POset Ontology Laboratory Environment) framework, a toolbox for working with partially ordered sets, which previously was the basis for POSOC, a categorization tool for automating functional annotation (Verspoor et al 2006). In particular, BIOGRAM incorporates the measures of hierarchical recall and precision which were used in POSOC to evaluate annotation performance. BIOGRAM is being developed within the context of an LDRD-DR project focussed on understanding host-pathogen Interactions in avian influenza, which combines experimental and computational approaches and allows us to test our algorithms and software within an ongoing experimental setting. Verspoor K, Cohn J, Mniszewski S, and Joslyn C (2006). A categorization approach to automated ontological function annotation. *Protein Science* 15:1544-1549.

Title: *Co-evolutionary Analysis of Mediator: a Multi-subunit Protein Complex*

Presenter: Elizabeth Eskow, University of Colorado Boulder

Authors: Elizabeth Eskow, Greg Caporaso, Debra Goldberg, Robyn Knight, Dylan Taatjex

Abstract: An analysis of the intermolecular co-evolution (correlated mutations) of proteins that comprise Mediator, a multi-subunit protein complex that is conserved throughout eukaryotes and is required for expression of all protein-coding genes, is in progress. Because the subunits of the human Mediator complex have diverged more significantly throughout evolutionary time than other general transcription factors, the study of the intermolecular interactions of human Mediator (i.e. beyond what can be gained from studying yeast Mediator interactions) is especially important and challenging. The size and complexity of the human Mediator subunits makes the identification of these intermolecular interactions using only biochemistry difficult without some guidance regarding which proteins interact and which key areas (residues) of the proteins to examine. A co-evolutionary analysis of all possible intermolecular interactions between pairs of proteins in Mediator subunits will provide this guidance by measuring which protein pairs exhibit relatively high numbers of coevolved residue pairs, and providing the locations of the significantly co-evolved pairs in their respective sequences. The biochemistry, in turn, provides validation (in the best case) for the predicted results or valuable feedback that can be applied toward method improvements. Our collaboration provides a unique symbiotic relationship between computational biologists and biochemists that may ultimately result in both greater understanding of the biochemistry of an extremely important protein complex (Mediator) and a computational method for identifying protein-protein interactions that could be applied to problems of a much greater (genome-wide) scale. The design of a procedure for co-evolutionary analysis of intermolecular interactions involves making algorithmic choices at each step to account for both cost and accuracy. An outline of the procedural steps and challenges follows: (1) Select sequences such that each protein in the potential interaction pair is represented once per species. The challenge is to select the 2 sequences per species that are most likely to interact out of many potential candidates. (2) Perform a multiple sequence alignment over the selected sequences for each protein. MSA algorithms generally maximize conserved alignments, and may not provide the best alignment for examining co-evolution. (3) Analyze the co-evolution of each aligned residue column with each column from the other protein. The choice of which method of analysis and whether to include phylogenetic information must be made, taking both accuracy and cost into account. (4.) An alternative alphabet may be used in the previous step instead of the 20-letter alphabet of amino acids. Reduced alphabets that symbolize specific biochemical properties such as charge or hydrophobicity may provide increased biochemical insights about the interactions, but the risk is that the more restrictive alphabet size will not represent the correlated mutations accurately. The overarching goal in the algorithm design is to make choices that provide defensible results for the biochemists to test within a reasonable timeframe, so that both groups can benefit from the feedback loop described above.

Title: *Molecular Characterization of Insertional Mutagenesis Sites in Gene Therapy Studies*

Presenter: Jessica Dantzer, Indiana University School of Medicine Center

Authors: Jessica Dantzer, Brandon Peters, Sara Dirscherl, , Jonathan Nowacki, Scott Cross, Xiaoman Li, Kenneth Cornetta, Mary C. Dinauer & Sean D. Mooney

Abstract: Research in gene therapy involving genome integrating vectors now often includes analysis of insertional mutagenesis sites across the genome using methods such as ligation mediated (LM)-PCR or linear amplification-mediated (LAM)-PCR. We have previously published SeqMap (<http://seqmap.compbio.iupui.edu/>), a secure, web-based, comprehensive integration site management tool that automatically analyzes and annotates large numbers of integration sites in mouse and human upon a common genomic framework. Recent updates to SeqMap include a comprehensive user guide, a submission wizard to guide users through the sequence submission process, and the newest versions of the human and mouse genomes for use in the analysis of submitted sequences. In this abstract we will continue our work to show how these data are being used for functional characterization including gene ontology enrichment analysis, pathway analysis with IPA, analysis of integration sites with respect to the TSS and comparisons with specific functional genomic datasets.

Title: *Using Protein Functional Features to Predict Deleterious Mutations Using Saturation Mutagenesis Data*

Presenter: Uday S. Evani, Indiana University

Authors: Uday S. Evani, Vidhya Krishnan, Sean D. Mooney

Abstract: Approximately half of the known disease-causing mutations result from amino acid substitutions (AAS). Hence much effort is going into building tools which can distinguish between disease mutations and harmless polymorphisms. In the last ten years many tools have been developed to optimally predict complex disease causing mutations. These tools include SIFT, PolyPhen, PMUT, PANTHER, LS-SNP, TopoSNP, SNPs3D and others. Most of these tools use sequence conservation and structure based features like solvent accessibility and secondary structure for building the classification model, because of which they are good at picking up mutations causing classical monogenic inherited diseases, but they do not perform as well when it comes to catching somatic mutations associated with more complex diseases like cancer. However, none of these tools actually predict the underlying cause of the disease. In this study we hypothesize that features based on protein function will perform similarly to sequence and structure based approaches, but will also give insights into molecular mechanism. These are novel features such as catalytic residue prediction, protein stability prediction, post translational modification prediction and protein interface prediction. We are aiming to use these features to achieve better sensitivity and specificity in separating functional/deleterious and neutral mutations. To evaluate this approach, we decided to use saturated mutagenesis data obtained from Lac Repressor, HIV 1 protease and bacteriophage T4 lysozyme for building features and then training and testing the model. SIFT (Sorting Intolerant From Tolerant) is used as a benchmark for evaluating our model. SIFT is a good choice for comparison, because it, too, is trained on these data sets. In this presentation will present our results and compare protein functional features to sequence and structural features.

Title: *Graph Theoretic Properties of Known Complexes*

Presenter: Suzanne Gallagher, University of Colorado

Authors: Suzanne Gallagher and Debra S. Goldberg

Abstract: One of the goals in studying the protein-protein interaction network (PIN) is to find $\text{emph}\{\text{protein complexes}\}$, groups of proteins that bind together to perform a specific task. Several different graph theoretic properties have been hypothesized to correlate closely with protein complexes, and there have been many attempts to find protein complexes in the interaction data by using those properties, often without a principled method to determine the values of parameters used by the algorithms or even whether the property used is a good indicator for protein complexes. Here we perform a survey of the topological properties of known complexes, both in isolation and as they appear in high-throughput data sets, to get a better idea of how an unknown complex is expected to appear in the data. In particular, we have computed degrees, edge density,

clustering coefficients, betweenness centrality, vertex and edge connectivity, and subgraph properties. The vertex (edge) connectivity is the number of proteins (interactions) that must be removed in order to disconnect the subgraph corresponding to a protein complex. This is the first usage of the property of connectivity for analyzing PINs.

Title: *Optimizing Proteomic Biomarker Discovery – Incorporating Prior Experimental Data*

Presenter: David Gnabasik, University of Colorado Boulder

Authors: Mark Duncan, David Gnabasik

Abstract: The promise of proteomics for biomarker discovery and biological pathways elucidation remains unfulfilled, despite significant advances in instrumentation, enhanced algorithmic sophistication and substantial investment. In large part, the problem relates to the fact that studies often do not deliver adequate statistical power given the quantitative precision of the analytical methods, limited sample numbers and constraints, available time and cost. Also, study design is often subject to the influence of chance and bias. We propose an interactive software system that optimizes the proteomic biomarker discovery process. While based on the approach and equipment used specifically for 2D-PAGE gel electrophoresis data, it is more generally applicable. The key objective is to develop a system to organize, store and represent proteomic data in a manner so that it can be used as prior information in subsequent experiments, thereby minimizing the need to perform proteomic experiments de novo each time – a system that incorporates previous findings and maximizes the benefits derived from each new study. The primary characteristics of this proteomic database are that: * prior data is organized and managed as measurements constrained by experimental context; and * prior data is represented and manipulated as joint probability distributions. The relevant experimental context includes the hypotheses, study design and variables, data variance and covariance structures, measurements and errors, statistical tests and assumptions, data transformations and normalization procedures used to generate the prior dataset. The specific output of the system is a Bayesian representation of a dataset as a probability distribution, along with a relevance graph, representing its similarity to all other comparable datasets in the database. For example, a previous study analyzing cancerous mouse lung tissue can be used as a comparable prior dataset to a similar, newly planned experiment with a quantifiable degree of confidence and relevance. The first priority in clinical proteomics is to establish reproducible quantitative differences between groups (e.g., cancer and non-cancer) that can guide clinical decision-making. Once a biomarker candidate passes this criterion, targeted protein identification is then performed. Otherwise, protein identification consumes enormous resources and distracts, in terms of time, sensitivity and precision, from efforts to discover new biomarkers. By directly addressing these study design issues, the proteomic database will enhance and accelerate proteomic research by providing an experimental focus in the face of overwhelming biological and computational complexity. This project provides a set of tools and methods for optimal performance of proteomic discovery studies. The tools allow the reliable comparison of data derived from two or more groups by extracting relevant similarities from the data.

They utilize prior, quantitative experimental data and knowledge, thereby permitting the possibility of aggregated and integrated proteomic experiments.

Title: *PHYLOCLUST – A Web Based Service for Clustering Human Genes Using Evolutionary Distances as Clustering Features*

Presenter: Chirayu Goswami, Indiana University

Authors: Chirayu Goswami, Jianfi Hu, Xiaoman S. Li

Abstract: Phyloclust is a new web service for clustering Human genes using phylogenetic distance measures as clustering criterion (features). In comparative evolutionary analysis, often times it may be required to cluster the genes, having similar evolutionary path or topology of trees, together. In this way we may have genes evolved in similar fashion clustered in one cluster. Orthologous genes in multiple species can be used to retrieve the pair wise evolutionary distances between the genes. These pair wise distances can then be used to cluster these genes. Pair wise distances between genes in one phylogenetic tree can give clear-cut idea about the tree topology & hence the arrangement of genes in the tree. Genes clustered this way in one cluster will show similar pattern of evolution. Phyloclust is a web based application which does this task online. It clusters user inputted genes hierarchically as well as by k-means measure. It uses Euclidean distance as similarity measure on the pair wise evolutionary distances of the genes under study, with other species. Currently Phyloclust uses Human & 3 other species Chicken, Mouse & Dog to retrieve pair wise evolutionary distances between them & uses these distances to cluster Human genes. The clusters formed have trees showing similar topology & also similar arrangement of genes within groups falling in same cluster. Phyloclust is available at :

<http://evolution.compbio.iupui.edu/phyloclust/home.php>

Title: *Laboratree: Extensible Basic Research Application and Data Management Using the OpenSocial Platform*

Presenter: Jamison R. Hemmert, Indiana University

Authors: Jamison R. Hemmert, Brandon Peters, Joshua D. Waymire, Kishore K. Kamati, Peter H. Baenziger, Craig A. Sanders, Joy A. Nellis, Justin J. Dantzer, Peter J. Serguta, Jessica L. Dantzer, Sean D. Mooney

Abstract: Solutions for enabling collaboration continue to be a challenge for biomedical informatics service providers. With the recent rise of so-called Science 2.0 web based tools, many solutions have been proposed using Internet technology including wiki's, document management systems, and custom websites. Here we propose utilizing an open platform, Google's OpenSocial platform, to enable scientific collaboration. To this end we have developed the Laboratree research management system (<http://laboratree.org>) to enable collaborative document and data management creation, group and project messaging, centralized authentication using OpenID and application exchange using OpenSocial. While there are many other social networking tools available, they are

largely reproducing Facebook, MySpace or LinkedIn. By focusing less on social networking and adding features that enable secure project management and supporting open source and open standards, we believe our approach has a promising future for distributed development.

Title: *An Efficient Approach to Protein Folding Trajectory Analysis*

Presenter: Shubham Jain, Jaypee Institute of Information Technology

Authors: Shubham Jain, Saurabh Dhupar, Dhawal Verma, Sanjeev Kumar Sharma

Abstract: In our quest to understand how proteins fold we need to understand the molecular mechanisms behind folding that occur at nano-second time scale. Today with immense computational power, huge amount of simulation data for proteins folding have been generated [1]. But there is a need to interpret and extract features from this data in order to understand the folding mechanisms. This involves comparison of multiple protein structures in folding trajectories. In past progressive methods like RMSD [1] and distance matrices [2] have been used to generate scores for snapshots of folding trajectory analogous to BLOSUM scores in sequence alignments. These scores are then used to generate overall alignment scores for the trajectories. Both these techniques have their own advantages and disadvantages. In the present work, a 2 layer supervised learning model is proposed where first layer is based on RMSD approach and second layer is based on distance matrix approach. An algorithm is developed to refine the individual scores and develop a combined score based on both the techniques. This has been shown to be a robust technique when applied on Trp-Cage simulation dataset. 1. Ota M, Ikeguchi M, Kidera A: Phylogeny of protein-folding trajectories reveals a unique pathway to native structure. PNAS 2004, 101(51):17658–17663. 2. Sun, Ferhatosmanoglu, Ota, Wang An enhanced partial order curve comparison algorithm and its application to analyzing protein folding trajectories, BMC Bioinformatics 2008 3. Yang H, Parthasarathy S and Ucar D: A spatio-temporal mining approach towards summarizing and analyzing protein folding trajectories. Algorithms for Molecular Biology 2007, 2:3 doi:10.1186/1748-7188-2-3

Title: *Reconstructing the Metabolic Network: a Probabilistic Approach to Improve Enzyme Prediction*

Presenter: Stacy Hung, Hospital for Sick Children

Authors: Stacy Hung, James Wasmuth, and John Parkinson

Abstract: We are interested in studying the metabolic network of apicomplexan parasites, which include Plasmodium falciparum, the causative agent for the most severe forms of human malaria. Since the reconstruction of this network is determined by the set of enzymatic reactions, our early efforts have focused on accurately identifying the proteins with enzymatic functions. For P. falciparum, current metabolism databases use similar methods to predict enzymes engage in cross-talk, yet lack the degree of confidence for each assignment. Furthermore, their methods do not consider the sequence diversity present within a class of enzymes. Including this data into the search is likely to

identify enzymes that have undergone large scale sequence evolution, a property common in parasites. Here, we describe a new probabilistic method for enzyme prediction based on the global and local sequence alignments of 117,000 enzymatic proteins. By comparing the alignment scores of an unknown protein to those of all known enzymes, a confidence score is calculated, ranking the reaction classes relevant for that protein. This approach has been benchmarked against 82 experimentally verified enzymes from *P. falciparum*. Based on a 5-fold cross-validation sensitivity/specificity analysis, our method predicts the correct Enzyme Commission number with an average accuracy of over 80%. We have re-annotated the metabolomes for apicomplexan species and a number of model organisms including human, yeast, and *E. coli*. This method will be useful for reconstructing metabolic networks for newly sequenced organisms, and has the added advantage over existing enzyme classifiers by providing multiple predictions with an associated score.

Title: *MutDB: Enhanced Biochemical Analysis of Structural and Functional Features of Genetic Variations*

Presenter: Kishore Kamati, Indiana University

Authors: Kishore K. Kamati, Rakesh Sathyesh, Matthew Mort, Arti Singh, Adebayo Olowoyeye, Jessica Dantzer, Charles Moad, Vidhya G. Krishnan, Peter H. Baenziger, Maricel G. Kann, Predrag Radivojac, Randy Heiland, and Sean D. Mooney

Abstract: Understanding how genetic variation affects the molecular function of gene products is an emerging area of biomedical research. Single Nucleotide Polymorphisms (SNPs) are the most common form of genetic variation. This research work aims to identify the SNPs and disease-associated mutations that are most likely to mediate a phenotypic change. The two steps involved in this project are presented here. Firstly, we collected and annotated many of the known coding disease-associated mutations with both the protein structural and functional information that are already available. Working with datasets of SNPs and mutations from various resources is very challenging. Therefore, data were collected from different sources such as mutations from OMIM and Swiss-Prot, non-synonymous SNPs from dbSNP, cancer mutations and a model of neutral mutations from alignments of orthologous proteins. To view and browse the huge amount of data collected, we developed MutDB, a useful tool for understanding the molecular basis of disease-associated mutations. MutDB integrates annotated SNPs from dbSNP and amino acid substitutions from Swiss-Prot with protein structural and functional features, links to scores that predict functional disruption and other useful annotations. In addition, a new functionality that facilitates KEGG pathway visualization of genes where SNPs are located and a SNP query tool for visualizing and exporting sets of SNPs that share selected features based on certain filters were also developed. This enables the researcher to view the systems context of both a mutation and its associated phenotype. We have constructed an AJAX (Asynchronous JavaScript and XML) based SNP browsing tool that allows users to save searches, select subsets of SNPs and view Haploview-like haplotype maps. The public web interface to this dataset is available at MutDB (<http://mutdb.org/>). Secondly, we are currently designing a web interface to the

pipeline that predicts various structural and functional features that are not available in any of the public databases for a given amino acid substitutions. These features include catalytic residue prediction, post-translational modifications etc. These features would be added to the annotation of MutDB. In the future, we hope to extend this work to provide automated methods to predict likely functional mutations on a genome-wide scale.

Title: *A Motif-based Phylogenetic Mixture Model for Discriminating Transcription Factor Binding Sites*

Presenter: Hyunmin Kim, University of Colorado

Authors: Hyunmin Kim, Wanjun Gu, Todd Castoe, David D. Pollock

Abstract: Transcription factors (TFs) play a key role in transcription, the initial step of gene expression and regulation. By binding to non-coding DNA sequences (transcription factor binding sites, or TFBSs), these proteins directly control the level of gene expression in a cell. Evolution of transcription factor binding sites (TFBSs) across species is constrained by the functional requirements of their interaction with transcription factors. Therefore, TFBSs are generally more conserved than surrounding sequences. The analysis of TFBS conservation among genomes from different species is becoming more feasible as more genomes become available, although the gain or loss of functionality is difficult to predict. Understanding of TFBS-oriented regulation is also becoming an important part of evolutionary biology. To infer an evolutionary model for the TFBSs of a specific TF, we need to acquire a set of known TFBSs studied in different species, and an alignment of background sequences to establish the rate of neutral nucleotide substitution in the genome. Since there are few complete datasets of multi-species TFBSs, we extract potential TFBSs of other species by studying orthologous promoters controlled by the target TF, and then align conserved motifs. Although, there are a variety of multiple alignment tools available, and their performances have been benchmarked in a particular condition, few of them are designed for discovering short motifs of potential TFBSs. It might be helpful for the alignment tools to consider realistic null models. However, the level of conservation is dependent on the location and context in mammalian sequences, and uniform null model is hardly applicable to all the cases. Many previous TFBS models are based on Halpern and Bruno (HB) model, which generates position-specific substitution model under a background noncoding model. TFBS position-specific profiles can be integrated into evolutionary models, but TFBSs may be incorrectly predicted, and true (functional) TFBSs may be mixed with artifacts. Our main goal is to evaluate the compatibility of different models with the existing data (alignments of TFBSs and twice as many alignments for decoys), to see how much different parameterizations helped with identifying TFBSs, and to see if the binding-site specific models could be used to discriminate among similar binding sites. To test this idea, we implemented motif-based mixture models and tested them on known Ap1 and Sp1 binding sites and flanking motifs. By using phylogenetic mixture models, TFBSs can be modeled along with neutral and conserved background evolutionary processes, in the face of predictive uncertainty. We found that the trend of slower evolution of TFBSs than that of artifacts is reflected in the mixture models, for which the true TFBSs are always more enriched in the more conserved class than the decoys. For example, 49% of Ap1

sites overall (190/387) were in the conserved class in the two-component symmetric rate-variable model, but 73% of known functional AP1 sites (37/51) were in this class, whereas only 46% of decoy AP1 sites (153/183) were in it. The corresponding true and decoy percentages for SP1 sites were similar (73% and 37%).

Title: *An Algebraic Topology Approach to the Classification of Protein Domain Structures*

Presenter: Michael Knudsen, Bioinformatics Research Center

Authors: Michael Knudsen, Robert Penner, Jørgen Andersen, Carsten Wiuf

Abstract: With today's standards in protein structure solving, new protein structures are added to the public databases at an ever increasing rate. At time of writing the PDB database contains 53,521 unique protein structures. Protein domains have been extracted from the PDB entries and classified in the CATH database. The classification is hierarchical with four nested main levels called class, architecture, topology, and homology, respectively. Classification on the class level (alpha, beta, and alpha/beta domains) is relatively easy, but already at the architecture level, manual work is still needed. We present a novel method for describing domain structures based on concepts from algebraic topology. Using the backbone atoms and the hydrogen bonds of a protein domain we create a combinatorial object, a so-called fatgraph, which is then transformed into a topological object. Compared to other mathematical descriptions of proteins, the topological object of our method does not depend on any particular embedding in an Euclidean space, and we define intrinsic quantities such as genus and boundary components of a protein domain. These are just a few examples of so-called topological invariants that can be associated with a protein domain. We have implemented algorithms which calculate the quantities above from PDB entries, and even at the lowest level, the homology level, simple classification schemes such as KNN perform very well on the output of our implementation. By using our method, the whole algebraic topology toolbox is readily applicable. Furthermore, the combinatorial object -- the fatgraph -- also deserves scrutiny in itself. Based on the initial results we believe that the ability to combine well established theory from combinatorics and algebraic topology may lead to better automated classification methods, and, perhaps even better, novel insight and understanding of protein structures.

Title: *A Functional Analysis of Novel Non-synonymous SNPs in the Thailand SNPs Discovery Project*

Presenter: Vidhya Krishnan, Indiana University

Authors: Somying Promso, Vidhya Krishnan, Eunseog Youn, Chintana Tocharoentanphol, Wasun Chantratita, Sean Mooney

Abstract: There are now an enormous number of single nucleotide polymorphisms (SNPs) available. Many of these SNPs have disease associations from past studies. Another useful problem is prediction of the molecular effect of SNPs in or near disease-associated genes, however a large number of these substitutions are unlikely to have any

detectable effects. Recently, the Thai SNP Discovery Project identified 58 non-synonymous SNPs (nsSNPs) in important proteins related to human health. To characterize nsSNPs that are likely to affect function in the Thai population and prioritize them for further functional studies, we analyzed two non-synonymous SNP (nsSNP) datasets using five in silico functional SNP prediction tools. Our data sets include non-synonymous SNPs in eight cardiovascular disease related genes and sixteen drug related genes. This analysis was performed using the publicly available web-based services; SIFT, PolyPhen, SNPs3D, PANTHER and PMUT. The prediction results were then compiled and the different methodological approaches compared to known disease-causing mutations, SNPs, and functional studies in these genes. Concordance between each of the five prediction tools was determined. These tools were then used to predict the SNPs in Thai project that are likely to affect the protein function. In brief, we find that nsSNPs; th515 (R511L), th2314 (K258N), th197 (I63R), th1752 (R193W), th1846 (F125I), th1650 (P720A), th247 (R54C), th1926 (R35W), th1952 (H139Y) are most likely to affect protein function. These nsSNPs are located in CYP1A1, CYP1A2, SULT1A2, CCL1, ITGAX, LIPG, CYP4B1, EPHX1 and TGFBR2 genes, respectively.

Title: *Beyond Dictionary Match for Disease Named Entity Recognition*

Presenter: Robert Leaman, Arizona State University

Authors: Robert Leaman, Christopher Miller, Graciela Gonzalez

Abstract: While there has been significant research into named entity recognition in the biomedical domain, the focus has primarily been on genes and proteins. There are many other entity types of interest to biologists, and some recent work has made progress towards the recognition of diseases in biomedical natural language text [Jimeno et. al., 2008]. Previous attempts to recognize disease entities have concentrated, however, on dictionary matching or other baseline approaches instead of the highly effective sequence tagging forms of machine learning, such as conditional random fields, which have become popular for the recognition of genes and proteins. In this work we evaluate the ability of the existing BANNER named entity recognition system, which is based on conditional random fields, to recognize disease entities [Leaman and Gonzalez, 2008]. We created the necessary training data from the gold standard dataset made publicly available by Jimeno et. al. by manually augmenting the annotations of each sentence in the corpus with the minimum supporting substring for the annotated entity. The results show that the machine learning approach significantly improves performance even with a small corpus. BANNER is a biomedical named entity recognition system implemented in Java and intended to be portable among many types of biomedical entities. It uses the MALLET implementation of 1st or 2nd-order conditional random fields (CRF) as the machine learning component, with a feature set consisting primarily of orthography, morphology and shallow syntax, including lemmatization and shallow parsing. BANNER was augmented with a dictionary match feature, employing a list of disease names obtained from the UMLS MetaThesaurus. BANNER is open source software and is freely available at <http://banner.sourceforge.net>. Manually augmenting the corpus created by Jimeno et. al. to create training data was necessary since while the original corpus included UMLS identifier of all diseases in the sentence, the exact location of each

mention within the sentence was not specified. The corpus was also further expanded by annotating sentences contained in the original corpus but not annotated previously. With 880 sentences annotated so far, the corpus is still relatively small compared, for instance, to the 20,000 sentences total provided by the BioCreative II corpus for gene mentions. BANNER was evaluated on the disease corpus using 5x2 cross-validation and exact boundary match, producing a precision of 75.79% and recall of 68.67%, for an f-measure of 72.06. The best baseline technique used by Jimeno et. al., dictionary matching, resulted in 54.94% precision and 65.86% recall for an f-measure of 60.88. MetaMap resulted in 31.27% precision and 30.07% recall for an f-measure of 30.66, and the statistical approach resulted in 26.07% precision and 30.18% recall for an f-measure of 27.97. The machine learning approach used by BANNER thus provides over 10% improvement in F-measure, primarily due to a dramatic improvement in precision. We anticipate that performance on disease entities will improve significantly as we continue the annotation effort and add specialized features geared towards improving the performance of BANNER to equal or possibly exceed the range offered for genes and proteins.

Title: *CADEG, a Set of Software Tools for High Resolution Identification of Potential Chromosomal Abnormalities*

Presenter: Chang Liu, The University of Hong Kong

Authors: Chang Liu

Abstract: Our previous study using a large set of microarray-based gene expression profiling data has identified many clusters of adjacent and similarly expressed genes (CASEGs), and we believe that these clusters of genes can be used to discover unique biological phenomena such as chromosomal domains or chromosomal aberrations. The current study focuses on cluster of adjacent and differentially expressed genes (CADEGs). Several methods and software applications have been developed along similar lines. These include ArrayFusion, which integrates multi-dimensional analysis of CGH, SNP and microarray data; ChroCoLoc, MACAT and REEF and LAP, which are to estimate the probability of co-localization of microarray gene expression; and LS-CAP, which is to predict the cytogenetic aberrations directly. However, these methods usually predict a large region containing many genes and do not produce a definite list of co-localized and differentially expressed genes, which is usually needed to generate specific hypotheses for further testing. In this study, we describe a simple algorithm that can identify CADEGs by combining well-developed statistical methods such as ANOVA, general metrics such as Hypergeometric probability and the chromosomal locations of individual genes to identify cluster of genes that shows correlated expression regulations. And the algorithm can identify the chromosomal abnormalities to the individual gene level. The methods have been implemented using perl programming language. And three different versions of the software tools, a command line-based, a web-based and a desktop version with GUI have been implemented. This application will satisfy various needs for easy installation, maintenance, performance and security. An experimental data set has been used to demonstrate the usefulness of this method and the potential applications of CADEGs are discussed.

Title: *Predicting Molecular Functions Disrupted by Mutations Using In Silico Functional Profiling*

Presenter: Sean Mooney, Indiana University

Authors: Matthew Mortl, Uday S. Evani, Peter H. Baenziger, Brandon Peters, Rakesh Sayeth, Yanan Sun, Bin Xue, Vidhya Krishnan, Yaoqi Zhou, David N. Cooper, Eunseog Youn, Matthew Hahn, Maricel Kann, Predrag Radivojac, Sean D. Mooney

Abstract: An important challenge in translational bioinformatics is understanding how disease-causing mutations give rise to the molecular changes that lead to the disease phenotype. Currently amino acid substitutions (AAS) account for ~57% of all monogenic disease causing mutations and therefore much of the research in this area has focused on AAS; AAS have also been studied because of their great potential to alter protein function and the availability of experimental data linking disease to mutation. There are now many resources available with annotations describing biochemical features potentially useful in identifying disease-causing AAS. These include SNPs3D (<http://snps3d.org>), the SNP Function Portal (<http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx>), PolyDoms (<http://polydoms.cchmc.org/polydoms/>) and MutDB (<http://mutdb.org>) among others. However, these resources typically identify features present at or near a mutation site but do not quantify whether a specific site or annotation is actually disrupted by the mutation. Many tools have been developed to predict so-called functional or disease-causing AAS. These tools include SIFT, PolyPhen, LS-SNP, SNAP, the SVM at SNPs3D, and others. They all operate using approximately the same principles: they are all supervised and use features based on protein sequence, sequence conservation, evolutionary information or protein structure. One difference between these tools is in the choice of training data which can be data sets of human alleles, evolutionary mutations, or experimental mutations. However, these tools perform similarly and only perform binary classification (disease/non-disease or function/non-functional). Because the features are based only on protein sequence and structure (and occasionally on previously known sites), these tools are unable to fully identify molecular causes of disease beyond disruptions of protein structure or sequence conservation. We seek to extend this area of research to predict actual molecular disruptions that cause disease. To this end, we are evaluating novel features based on predicted molecular functions. Our approach is relatively simple: we utilize statistical methods to predict amino acid functions and then measure how the predictions change after mutation. Machine learning methods that predict structural and functional sites in amino acid sequences are well established and facilitate the prediction of secondary structure, solvent accessibility, post-translational modification, protein domain, and protein interaction participating residues. Here, we use these methods to evaluate whether we observe statistically significant differences between disease-causing mutation data sets and data sets of mutations that are unlikely to have any effect on molecular structure and/or function. In this study, we compare several structural feature prediction methods, post-translational modification prediction methods and a sequence-based predictor of catalytic residues against three types of data sets, mutations causing inherited disease, somatic cancer-associated mutations and

polymorphic sites observed in gene resequencing projects. We find that mutations causing inherited disease are likely to be predicted to be deleterious using SIFT, and tend to be enriched for structural disruption by comparison with polymorphisms. In addition, we find that cancer-associated mutations are predicted to be significantly enriched in post-translational modification sites, protein disorder and are less likely to be predicted to be deleterious by SIFT than inherited mutations.

Title: *The Human Gene Mutation Database (HGMD) and its Biomedical Exploitation*

Presenter: Matthew Mort, Cardiff University

Authors: Matthew Mort, Peter Stenson, Edward Ball, Katy Howells, Andrew Phillips, Nick Thomas, David Cooper

Abstract: The Human Gene Mutation Database (HGMD) is a comprehensive core collection of data on germ-line functional variation in nuclear genes often underlying or associated with human inherited disease (<http://www.hgmd.org>). Data catalogued include single base-pair substitutions (in coding, regulatory, and splicing-relevant regions), micro-deletions, micro-insertions, indels, repeat expansions, gross gene deletions and insertions/duplications (including copy number variations) and complex rearrangements. By September 2008, the database contained in excess of 80,000 different lesions detected in 3,064 different nuclear genes, with new entries accumulating at a rate in excess of 5000 per annum. Variants in HGMD have recently been mapped to the human genome assembly to facilitate a greater understanding of the mechanism by which genetic variation contributes to inherited disease.

Title: *Enabling Basic Biomedical Core Services Using Informatics: Experiences Developing a Fee-for-service Bioinformatics Core*

Presenter: Joy A. Nellis, Indiana University

Authors: Joy A. Nellis, Peter H. Baenziger, Jamison R. Hemmert, Ganesh Shankar, Brandon J. Peters, Lang Li, Sean D. Mooney

Abstract: Enabling basic biomedical core services using informatics: Experiences developing a fee-for-service bioinformatics core Authors: Bioinformaticians are often asked to develop and lead informatic services ranging from analysis of high throughput datasets to cyber-infrastructure development. Now, since many medical centers have or are competing for funding specifically for clinical and translational research, developing translational, biomedical informatic services becomes a requirement. At Indiana University School of Medicine, we have developed a fee-for-service model for providing both short-term and long-term, informatic solutions to research projects, academic groups, and the campus as a whole. The Bioinformatics Core at Indiana University School of Medicine has been operating for two years serving basic and clinical researchers through application development, data analysis, database development, web and graphic design, intranet hosting, and more. Through our diverse resources and personnel, a mature development environment, and our dynamic project scope we are able to serve a diverse set of researchers with rapid development and quality solutions to

many biological and informatic problems. We present example projects, a simple explanation of core operations, and lessons learned in collaboration, project development, advertising, and establishing new projects. We discuss practices we have adopted in development, analysis, and technology including project tracking, code repository use, and our project management tool: Laboratree. We also describe challenges we have faced deploying wikis (such as MediaWiki or Twiki), document management tools (such as Laboratree), project tracking tools (such as dotProject or TRAC), and various messaging platforms. We hope this presentation enables other researchers to initiate core services efficiently and in an intellectually and fiscally sustainable manner.

Title: *A Semantic Architecture for Biologics Research at Merck*

Presenter: Eric Neumann, Science Commons, MIT

Authors: Eric Neumann, Jaime Melendez, Mike Bevil

Abstract: Discovery Research is more often than not an inexact process since it involves novel techniques and new modes of biological reasoning. Case in point is the ability to create millions of protein variants by combinatoric sequence synthesis, which then also requires large-scale screening and assay tracking systems. This has a direct impact on how informaticists need to efficiently structure information produced and utilized by discovery researchers. It is often the case that traditional IT approaches such as formal database design are inadequate due to the loose definitions and changing nature of discovery methodologies. More often than not, researchers are compelled to use flexible applications such as Excel® for handling their irregular, shifting information. This leads to secondary problems involving de-centralized data storage, multiple non-standard formats, and unclear data semantics. An information system that supports evolving data models, enterprise data relations, and company-wide annotations would be of great use to these researchers and their respective companies. We present here a functioning proof-of-concept showing how Semantic Web technologies and standards can be to applied data management and information mining for an enterprise biologics discovery group. The analysis of the outcomes regarding both researcher satisfaction and system development advantages are also described here.

Title: *The Phylogeny of the MHPs and Beyond*

Presenter: Mary J. O'Connell Dublin City University

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

Abstract: What leads to the development of specificity in an enzyme? We present our computational approach to determining the amino acids responsible for diversity of function in the mammalian heme peroxidases. The evolutionary relationships of these enzymes was until now unknown and despite their relevance in the fields of asthma, Alzheimers disease and inflammatory vascular disease research little was known of their evolution. We demonstrate, using a root mean squared deviation statistic, how the removal of the fastest evolving sites aids in the minimisation of the effect of long branch attraction and the generation of a highly supported phylogeny. Based on this phylogeny

we have pinpointed the amino acid positions that have most likely contributed to the diverse functions of these enzymes. Many of these residues are in close proximity to sites implicated in protein misfolding, loss of function or disease. Our study has (i) fully resolved the phylogeny of the MHPs and the subsequent pattern of gene duplication, and (ii), we have detected amino acids under positive selection that have most likely contributed to the observed functional shifts in each type of MHP. We are currently investigating the effects on function of mutating these positively selected sites to ancestral forms.

Title: *COBRA: A Web-based Tool for Genotype/phenotype Data Management.*

Presenter: Brandon Peters, IU School of Medicine

Authors: Brandon Peters, Rebecca Fletcher, Kirthi Krishnamraju, Anne Nguyen, Jason Robarge, Lang Li, Mark Crowder, Todd Skaar, David Flockhart, Sean Mooney

Abstract: The examination of inherited variations in genes that dictate drug response is making personalized medicine a reality. High throughput DNA sequencing technologies present us with unprecedented opportunities for studying these inherited variations and allow for the discovery of clinically useful pharmacogenomic applications. However, these advances in high throughput technologies have made management of the resulting deluge of genotype and sequence data increasingly more difficult. To efficiently manage large-scale SNP genotype and sequencing data there is a strong need for a software system that can help regulate the acquisition, organization, storage, retrieval, analysis and exportation of the data. We have created a pharmacogenetic information management system to facilitate these needs. Our goal is three-fold: [1] create a well-structured software tool/database to manage genotype data and annotate genetic polymorphisms from resequencing data for variant discovery, [2] apply the tool to facilitate analysis of the data by conducting meta studies to elucidate genotype/phenotype associations, and [3] create automated tools for submitting data to national public repositories such as the Pharmacogenetic KnowledgeBase (PharmGKB). In this project, we use the database we created to test the hypothesis that TGF-beta response is altered by genotype. Additionally, we developed a feature to streamline genotype and sequencing data submissions to a public resource, The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB), whose mission is to catalyze pharmacogenomic research. Currently, we have developed an infrastructure to manage genetic data for the Consortium on Breast Cancer Pharmacogenetics (COBRA) project and we have developed initial tools for querying, downloading and exporting the data to PharmGKB using XML. In summary, a genetic data information management system has been created to efficiently handle genotype and sequencing data and to facilitate pharmacogenomic association studies.

Title: *Insights on the “Oaks” of the Forest of Life*

Presenter: Pere Puigbo, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

Authors: Pere Puigbo, Yuri I. Wolf, Eugene V. Koonin

Abstract: The reconstruction of the “Forest of Life” is based on the idea that prokaryotes, effectively, share a common gene pool. This gene pool consists of genes with widely different ranges of phyletic spread, from universal to rare ones only present in few species. Phylogenetic trees of all these genes comprise the “Forest of Life”. Here, we analyzed 102 Clusters of Orthologous Genes that are represented in at least 90 of selected 100 prokaryotes (59 Bacteria and 41 Archaea), i.e., operationally universal orthologs. Thus, the trees obtained for these clusters are the biggest trees in the “Forest of Life”, the “oaks”. For 44% of the universal clusters, we detected evidence of horizontal gene transfer between Archaea and Bacteria. The rest of the trees do not show indications of such interdomain gene transfers but are not free of horizontal gene transfers inside the two domains. These gene transfers appear to be independent among the “oaks”, so the single strongest signal of similarity between them is based on vertical inheritance. Despite the presence of the vertical signal, tree topologies, on average, show limited similarity to each such that a single connected network of trees could be formed only with the 50% similarity cut-off. By analyzing this network, we identified 4 “oaks” with the strongest signal of vertical inheritance that are similar to the greatest number of other trees. This work is the first step to reconstruct the “Forest of Life” and decipher its intrinsic architecture. Although the “oaks” in the forest show a strong signal of vertical inheritance, there is also a moderate to strong signal of horizontal gene transfers even among these trees. Given the lack of a general topological congruence, these trees cannot be used as representatives of all trees in the forest and less so to approximate the “Tree of Life”.

Title: *Category Theory and the Notion of Natural in Bioinformatic Models*

Presenter: Peter Salamon, San Diego State University

Authors: Peter Salamon, James Nulton, David Aaby, Andrew Detzel and Barbara Bailey

Abstract: In the pursuit of natural patterns on the set of distances between phage proteins, a natural correspondence between two mathematical structures arises: distance graphs and random walks. The correspondence leads to evolutionary importance ranking on phage proteins but leaves various details unspecified. Such details are nicely decided by requiring the correspondence between distance graphs and reversible Markov chains to be functorial.

Title: *A HUB for Indiana's Clinical and Translational Sciences Institute*

Presenter: Craig Sanders, Indiana University

Authors: The IndianaCTSI HUB Team

Abstract: NIH has awarded several Clinical Translational Sciences Awards (CTSAs) in an effort to improve the way biomedical research is conducted across the country, reduce the time it takes for laboratory discoveries to become treatments for patients, engage communities in clinical research efforts, and train the next generation of clinical and translational researchers. Indiana University and Purdue University worked together to receive one of these awards in order to promote this cause over the entire state of Indiana. We are using the web based HUB community building platform developed at Purdue University to construct a centralized online portal to coordinate translational activities across the state of Indiana. HUBs have been previously used to power nanoHUB and other research networks. Each HUB is based on the Joomla! content management system, and we have extended it to give researchers access to clinical trial metadata, researcher profile data, internal proposal reviews, user submitted resources and other useful applications of translational research. We also will utilize metrics provided by HUB technology to help demonstrate progress towards the stated goals of IndianaCTSI.

Title: *Towards a Multi-Level Calculus for Cellular Modeling and Simulation*

Presenter: Trevor Sarratt, University of Tulsa

Authors: Stephen Tyree, Rayus Kuplicki, Trevor Sarratt, Scott Fujan, John Hale

Abstract: Exploration of cellular biology and biochemistry has unveiled a vast number of structures and processes yet to be fully understood. These are being cataloged and modeled, but there is no standard or complete mechanism for developing and interacting with these models. As a result, researchers have only a limited capacity to “connect the dots” among different theories and observations. Furthermore, due to the complexity of cellular environments, the description and simulation of holistic cell models remains difficult, a problem exacerbated by the lack of sufficient tools. We adopt the position that a multi-level, domain specific process calculus could fill this void. Process calculi are formal methods for specifying and examining systems of independent, concurrently-operating components. Hence, a process calculus is an ideal foundation for modeling the massively concurrent systems of cellular biology. Formal specification in a process calculus supports modularity, composition, and proof mechanisms in the description and examination of biological systems. Yet a general purpose calculus cannot efficiently express the primitive behavior of cells and their components. Special syntax and semantics are needed to capture cellular behavior. Several features critical to a workable calculus for cell biology have already been explored. For example, stochastic molecular interactions, structural assembly and decomposition, and robust abstraction have been successfully integrated into biologically-inspired process calculi. Existing execution models have shown these calculi to be sufficient for simulating many complex systems. However, additional concepts we deem vital include those of cell membrane behavior

and reaction locality. Existing biological calculi provide similar extensions, but model development is cumbersome and scalable execution models are lacking. As an alternative to supplementing an existing calculus with the requisite primitives, a multi-level calculus may produce a better modeling and simulation environment. The lower level of the model will adapt the Stochastic Pi-Calculus and the efficient Stochastic Pi-Machine (SPiM) for an execution model. Operating above this level will be another calculus designed to enforce containment and locality constraints on the lower level processes. The upper level will manage adjacencies of and transitions among membrane-bound containers and sub-container localities. A proof-of-concept model has been implemented wherein SPiM instances are organized within a static grid with process-specific diffusion rates set between adjacent grid spaces. The system executes by the iteratively simulating each grid space's SPiM instance for a short time interval followed by randomly shuffling processes in accordance with the diffusion rates. Further work is needed to define the operation of the upper level calculus and the interaction between the layers. Ultimately, a scalable execution model and intuitive specification system must follow to permit hypothesis-driven, model-based exploration. We believe these goals are achievable and necessary to promote systematic biological study at the cellular level.

Title: *OSIRiS: An Open Source PHP-Based Laboratory Management System*

Presenter: Peter Serguta, Indiana University

Authors: Jessica Dantzer, Peter Serguta, Aaron Baker, Christopher Hobbick, Rupa Chatterji, Kelley Faber, Cheryl A. Halter, Sean D. Mooney

Abstract: Laboratories often have need of software to enable easy cataloging and tracking of large numbers of samples. While there are many LIMS and sample tracking solutions available, they are usually expensive, difficult to set up or require special training to use. The Indiana University Bioinformatics Core, in collaboration with the Division of Hereditary Genomics and the IU DNA and Cell Repository (DNACR, NCRAD NIH 5U24AG021886-07), has developed a web-based open-source system as an easy to use solution for small groups and companies. The Open Source Inventory and Reporting System (OSIRiS) is written primarily in PHP and utilizes the open-source LAMP (Linux, Apache HTTP server, MySQL, PHP) framework. It also works with other server and database combinations including Windows and Microsoft SQL Server, and has been well tested in both Firefox and Microsoft Internet Explorer. OSIRiS allows a laboratory to track their inventory of different samples, and allocate these samples for shipment through request forms. Perl scripts allow users to import sets of data from Microsoft Excel spreadsheets and comma-delimited (.csv) text files. Users have the ability to add, edit, or delete nearly any aspect of a sample's data, provided their permissions allow such actions. Access controls are provided to assign a role and specific studies to users to allow multiple labs and study coordinators to use the same system. The software is currently being used by the Bioinformatics Core at IU School of Medicine, the National Gene Vector Biorepository (NGVB) and the IU DNA and Cell Repository. The software is available upon request from the authors.

Title: *Expression Profiles of Human PMS2-Related Genes: Bioinformatics and Experimental Studies*

Presenter: Elena K. Shematorova, Russian Academy of Sciences

Authors: Elena K. Shematorova, Dmitry G. Shpakovski, and George V. Shpakovski

Abstract: The expression patterns of human PMS2-like sequences found on chromosome 7 of the genome of *Homo sapiens* have been studied by bioinformatic and experimental means. Human PMS2 protein belongs to important components of the DNA mismatch repair (MMR) system [1]. In contrast to the majority of eukaryotes, not only PMS2 gene but also the fifteen different PMS2-like sequences have been found on chromosome 7 in the human genome [2-4]. Expression of the fifteen different PMS2-like sequences produce plural mRNAs of different types. Computational analysis of all hPMS2-related ESTs and cDNAs available in current databases has allowed us to classify them into three main classes covering exons 1-5 (class I), 9-11 (class II) and 11-15 (class III) of the original hPMS2 master gene, respectively. In addition, cDNAs belonging to the two first classes were further subdivided into two different subgroups each, producing five structurally distinguished subgroups of PMS2-like mRNAs in total. Because nucleotide sequences of the most complete members of the every subgroup have extensive coding capacities ranging at least from 190 to 240 aminoacids, we have isolated them from different sources and recloned in suitable vectors for yeast two-hybrid system and for heterologous overexpression in *Escherichia coli*. Complete PMS2-like cDNAs representing four out of five subgroups have been already cloned and structurally characterized. Particularly of interest, for the first time cDNAs were isolated containing the first coding, ATG-containing exon of two PMS2-related human genes belonging to the class I (see above). This first exon was missing in all previously reported mRNAs [2, 3]. This fact provides an important evidence in favor of existence of the hPMS2L proteins in human proteome [5]. The work was supported by the program "Molecular and Cell Biology" (direction 'Functional Genomics') of the Russian Academy of Sciences and by grant No. 07-04-01167 of the Russian Foundation for Basic Research (RFBR).
References 1. Modrich, P., and Lahue, R., Mismatch repair in replication fidelity, genetic recombination, and cancer biology, *Annu. Rev. Biochem.*, 65: 101–133, 1996. 2. Horii, A., Han, H.J., Sasaki, S., Shimada, M., and Nakamura, Y., Cloning, characterization and chromosomal assignment of the human genes homologous to yeast PMS1, a member of mismatch repair genes, *Biochem. Biophys. Res. Commun.*, 204: 1257–1264, 1994. 3. Kondo, E., Horii, A., and Fukushige, S., The human PMS2L proteins do not interact with hMLH1, a major DNA mismatch repair protein, *J. Biochem.*, 125: 818–825, 1999. 4. De Vos, M., Hayward, B.E., Picton, S., Sheridan, E., and Bonthron, D.T., Novel PMS2 pseudogenes can conceal recessive mutations causing a distinctive childhood cancer syndrome, *Am. J. Hum. Genet.* 74: 954–964, 2004. 5. Shpakovski, D.G., Shematorova, E.K., and Shpakovski, G.V., Human PMS2 gene family: origin, molecular evolution, and biological implications, *Dokl. Biochem. Biophys.*, 408:175–179, 2006.

Title: *Application of Next-generation Sequencing Technology for Comparative Transcriptome Analysis in the Nematodes C. Elegans and C. Briggsae*

Presenter: Heesun Shin, Simon Fraser University

Authors: Heesun Shin, Shu Yi Chua, Steven J.M. Jones, David L. Baillie

Abstract: We employed Illumina sequencing technology to sequence *C. elegans* L1 starved and fed, *C. briggsae* L1, and *C. briggsae* mixed population. With existing *C. elegans* L1 EST and SAGE data, we will be able to analyze transcriptional activities and discover novel transcripts specific to the first larval stage. We will compare *C. elegans* L1 data with *C. briggsae* L1 transcriptome data. Also, *C. briggsae* transcriptome data of mixed population will be a useful addition for comparative genomics studies. We have an average of approximately 15 million sequence tags that are 42 bases long for each of the four libraries (*C. elegans* L1 starved and fed, *C. briggsae* L1, and *C. briggsae* mixed population). I plan to compare gene expression profiles of starved L1 stage worms and fed L1 stage worms to investigate L1 stage specific gene expression changes involved in development in response to starvation or feeding. The stage specific transcriptome data from the *C. briggsae* L1 library will be used to perform an interspecies comparison with *C. elegans* L1 transcriptome data. By comparing and contrasting these two transcriptomes I hope to reveal species specific characteristics. Finally, I will analyze the transcriptome from mixed population *C. briggsae*. Transcriptome data of this depth is not currently available for this species and will provide a global picture of transcriptional activities in *C. briggsae*. This resource will provide the basis for an in-depth comparative genomic analysis between *C. elegans* and *C. briggsae*.

Title: *Hominoids-specific Molecular Evolution of PMS2 Gene Family and its Possible Biological Implications*

Presenter: George V. Shpakovski, Russian Academy of Sciences

Authors: George V. Shpakovski, Elena K. Shematorova, Dmitry G. Shpakovski

Abstract: Using bioinformatic (BLAST, BLAT, ClustalW, DS Gene and PHYLIP software packages) and experimental (PCR cloning and sequencing of cDNA and genomic fragments from various species) approaches we carried out a molecular characterization and phylogenetic analysis of the sixteen *PMS2*-like genes present at several loci on chromosome 7 of *Homo sapiens* and corresponding *PMS2* paralogues in other primate species (*Callithrix jacchus*, *Macaca mulatta*, *Nomascus leucogenys*, *Pongo abelii*, *Gorilla gorilla* and *Pan troglodytes*). Our results indicate that amplification of the *PMS2* genes is characteristic only for higher primates and was originated as duplication of the *PMS2CL* region (exons 9-15; also known as $\psi 0$ pseudogene) approximately 18 Mya [million years ago] in lineage leading to modern Lesser (gibbon) and Great Ape species after its separation from branch of Old World Monkeys. The evolution follows by multiple rounds of amplification of the region *PMS2NL* (exons 1-5; $\psi 1$ - $\psi 14$ pseudogenes) in Asian (gibbon and orangutan; only one *PMS2NL* gene subgroup present) and African (gorilla and chimpanzee; appearance of the second *PMS2NL* subgroup with

only two different members) Apes. *Homo sapiens* represents an extreme case of amplification of the second *PMS2NL* gene subgroup with eight for the most part clusterly arranged new members: $\psi 2-\psi 3$, $\psi 5$, $\psi 6-\psi 8$, $\psi 11-\psi 12$. In summary, we have shown that molecular evolution of PMS2 protein, the essential component of the eukaryotic mismatch repair and gene conversion systems, correlates well with the route of evolution of the ape lineage leading to hominids. Our data suggest that a novel tripartite polypeptide system of PMS2-like proteins, which could replace the original PMS2 protein in some of its novel or substantially diverged functions, have been established at least in *Homo sapiens* – with possible impact on such vitally important genetic processes as gene conversion, somatic hypermutation and class switch recombination [1]. The work was supported by the program "Molecular and Cell Biology" (direction 'Functional Genomics') of the Russian Academy of Sciences and by grant No. 07-04-01167 of the Russian Foundation for Basic Research (RFBR).

References

[1] Shpakovski, D.G., Shematorova E.K., and Shpakovski, G.V., Human *PMS2* gene family: origin, molecular evolution, and biological implications, *Dokl. Biochem. Biophys.*, Vol. 408, No. 5, pp. 175-179, 2006.

Title: *Abundance Of Universal Stress Protein Family In Finished Prokaryotic Genomes*

Presenter: Shaneka Simmons, Jackson State University

Authors: Shaneka Simmons, Hari Cohly, Rajendram Rajnarayanan, Dwayne Sutton, and Raphael Isokpehi

Abstract: Stress is an altered physiological condition caused by factors that tend to disturb an organism's homeostasis. Proteins with the Usp domain are known to provide cells with the ability to respond to environmental stresses such as nutrient starvation, drought, high salinity, extreme temperatures, and exposure to toxic chemical. We have exploited the availability of protein domain annotation in the Integrated Microbial Genome (IMG) Database (<http://img.jgi.doe.gov>) to scan for genes predicted to encode proteins that contain the universal stress protein domain in finished prokaryotic genomes. The objective was to compile a list of the completed prokaryotic genome sequences and determine the presence and number of genes that encode proteins with the universal stress protein domain. We used query tools on the IMG Database to scan over 600 finished prokaryotic genomes for genes annotated with Protein Family (Pfam) Domain PF00582. We identified 538 genomes with at least one gene annotated with the Usp domain. The number of genes with this domain per genome ranged from 1 to 36. The genomes of *Haloarcula marismortui* ATCC 43049 and *Natronomonas pharaonis* DSM 2160 had the highest number of genes annotated with the universal stress protein domain. The high number of Usp genes in these two archaea reflects their ability to survive in extreme conditions. *Saccharopolyspora erythraea* NRRL2338 was ranked among the genomes with at least 30 Usp genes. This mycelium-forming actinomycete produces erythromycin A, the clinically important macrolide antibiotic. The Usp genes may contribute to conferring resistance to a range of common antibiotic classes observed with

S. erythraea. Interestingly, except for *Mycobacterium leprae*, all the *Mycobacterium* (another group of actinomycetes) genomes analyzed contained at least 10 Usp genes. *Mycobacterium tuberculosis*, the causative agent of tuberculosis is known to be highly resistant to antibacterial drugs. Our future research will determine from multiple genomes, the functional coupling of proteins of the Universal Stress Protein family to other proteins. Acknowledgements: Mississippi NSF-EPSCoR “Innovations through Computational Sciences” Award (EPS-0556308); the Research Centers in Minority Institutions (RCMI) – Center for Environmental Health (NIH-NCRR G12RR13459-09); U.S. Department of Homeland Security under Grant Award Number 2007-ST-104-000007 (Bioinformatics in Biodefense Career Development Program) and Chemical Materials and Computational Modeling (W912HZ-04-2-0002). Disclaimer: “The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.

Title: *Modelling Elastin Self-Assembly*

Presenter: Hongyan Song, Hospital for Sick Children

Authors: Hongyan Song & John Parkinson

Abstract: Modelling Elastin Self-Assembly Hongyan Song & John Parkinson Program of Molecular Structure & Function Hospital for Sick Children 101 College St., Rm 15-704 Toronto, ON M5G 1L7 Abstract Elastin is a polymeric structural protein which plays an integral role in the extracellular matrix (ECM) of elastic tissues such as aorta, arteries and lung parenchyma. Due to its important role within these tissue, disorders in elastin production and assembly have been implicated in cardiovascular diseases such as aneurysms, atherosclerosis and hypertension. Consisting almost entirely of alternating hydrophobic and cross-linking domains, elastin is produced as a monomer, tropoelastin, which self-assembles (in a process termed coacervation) into fibers. Interactions between elastin fibers and other components of the ECM result in the formation of a complex three dimensional meshwork in which elastin is responsible for imparting properties of extensibility and elastic recoil. Tropoelastin is a highly hydrophobic protein composed largely of two types of domains that alternate along the polypeptide chain. Each domain is represented by an exon in the elastin gene. Hydrophobic domains are rich in glycine, proline, alanine, leucine and valine and are thought to be responsible for imparting the elastomeric properties. Cross-linking domains have at least partial alpha-helical character and contain lysine residues which are destined to form covalent cross-links that stabilize the polymeric form of elastin. The self-assembly properties of elastin appear to be determined by interactions between hydrophobic domains. This is thought to result in the alignment of the cross-linking domains allowing the stabilization of the polymeric matrix through the formation of lysine based cross-links generated through the action of lysyl oxidase. However what is not known is how the domain architecture of tropoelastin influences the self-assembly process and resultant morphology of the elastin fibres, and further, how alterations in its molecular structure may ultimately impact cardiovascular disease. To explore the influence of domain architecture on elastin self-assembly, we have developed a novel three dimensional simulation environment based on the use of

diffusion limited aggregation. In a typical model, elastin monomers – represented by individual rod-like particles – are introduced and allowed to freely diffuse until they encounter the growing aggregate whereupon they adhere with a probability determined by the proportion of hydrophobic contact. Through repeating this process several thousand times, large aggregates can be constructed which can then be analysed for their connective and mechanical properties. Using this tool, we have been performing a systematic scan of the impact of different hydrophobic and cross-linking domain configurations on fibril morphology. For example, increasing the number of domains while maintaining monomer length, reduced the overall length of the resultant aggregate while increasing the number of connections between individual monomers. These results may provide insights into the observed domain expansions associated the evolution of the elastin sequence. Here I present the design and construction of the simulation environment and discuss our initial findings.

Title: *Analysis of Workflows on Inexpensive, Massive Computational Grids in the Cloud*

Presenter: Jason A. Stowe, Cycle Computing

Authors: Jason A. Stowe, Ian D. Alderman, and Dr. Irene M. Ong

Abstract: It is increasingly possible for scientists who need to perform large amounts of computation, such as sequence alignments, to obtain inexpensive, short term access to a large number of computing resources. This model for obtaining access to resources, known as "Cloud Computing", provides scientists, even those with few resources, with the ability to obtain results quickly and inexpensively relative to purchasing and administering clusters in house. We investigate the trade-offs of using cloud computing by analyzing the costs of using Blast and Gromacs on Amazon's EC2 (Elastic Compute Cloud). In particular, we compare the response times and costs of executing typical BLAST queries on each of the machine types offered by EC2. We consider five different machine types ranging from a simple single-core machine with a 32-bit architecture to a high end eight-core 64-bit machine, each with a different cost. Pricing is per hour, and includes additional charges for data transfer. Our goal is to be able to answer questions such as: "If I want to execute Blast on a certain size of input with this database, how long will it take and how much will it cost?", "Which machine configuration should I use to minimize cost and how long will it take for my jobs to complete?", as well as "How much will it cost and how long will it take to get this work done as quickly as possible?" This last question is particularly relevant because Cloud Computing permits very large numbers of machines to be used for short amounts of time. In order to be able to answer these questions, we are running a variety of BLAST/Gromacs jobs in Amazon EC2, comparing the costs and run-times. We are managing the jobs with tools we have developed and with Condor, an open source distributed batch computing system. The tools we have developed include custom Amazon Machine Images (AMIs) which have BLAST tools and Condor installed, a web interface for managing EC2 sessions, capable of automatically starting more AMIs as they are needed and stopping them as jobs complete, and a management console for Condor. In our presentation, we will

demonstrate these tools as well as presenting our results.

Title: *Biological Sequence Simulation For Complex Evolutionary Hypotheses*

Presenter: Cory Strobe, University of Nebraska

Authors: Cory Strobe, Stephen Scott, Kevin Abel and Etsuko Moriyama

Abstract: The goal of sequence simulation is to realistically portray the evolutionary wrestling match between (1) processes that change a biological sequence through mutation, insertion/deletion (indel), as well as more dynamic chromosomal arrangement and (2) functional constraints that restrict such changes. Such processes are intertwined during the course of evolution, forming the patterns that we see in extant biological sequences. When sequence simulation was initiated in the 1990's, simulation methods mainly dealt with substitution processes, incorporating information about substitution patterns, relative substitution rates across sites (the Gamma distribution) and sites that are invariable throughout the evolutionary history of a group of sequences. Such methods were useful for testing hypotheses about phylogenetic relationships and substitution patterns during sequence evolution. These methods use a phylogenetic tree with branch lengths to determine the path that sequences are simulated along, where the branch lengths (the number of substitutions per site) determine the rates of evolution and a model of continuous substitution evolution processes (e.g., PAM or BLOSUM for proteins and HKY or GTR for nucleotides) accounts for multiple substitutions occurring at the same site. Substitution-based methods, however, did not incorporate the information on the insertions and deletions (indels). The pioneering indel sequence simulation application, Rose, introduced these events and extended the Gamma distribution to encompass indel constraints as well. In Rose, if a site is assigned with a low "Gamma" rate below a certain threshold, an indel is forbidden to occur. To further aid in sequence alignment analysis, the "true" multiple alignment that reflected the true evolutionary path of the sequences was provided by Rose. True multiple alignments can be used to test, e.g., the accuracy of multiple sequence alignment methods. A "general recipe" of sequence simulation is as follows: Input consists of a guide tree, a root sequence, and various substitution and indel parameters -- all of which are global over the simulation run. During the simulation, the sequence being simulated is treated homogeneously between sites (sites undergo similar evolutionary constraints). Current simulators each introduce disjoint piecewise improvements over the general model. Better sequence simulation methods should encompass the range of possible evolutionary changes needed to create realistic sequences, particularly site-specific constraints for both substitutions and indels and lineage-specific evolutionary constraints on subsequences and evolution parameterizations. Additionally, a flaw exists in many of the current simulation methods. Since little was known about the patterns of indel occurrence at the time of Rose's development, indels were simulated based only upon the probabilities and length distributions of insertions and deletions, with insertions and deletions following a continuous model, similar to substitutions. A continuous model, however, assumes that the length of the sequence will remain the same during the evolution time, which is clearly incompatible with insertion and deletion processes. To deal with this, iSGv2 adopts a discrete evolution model. We formalize the model of insertions and deletions

with respect to the sequence and functional constraints, leading into the rationale as to the flaw in representing indels similarly as substitutions.

Title: *Identification of Candidate Nuclear Receptors and Sterol Sensing Domain Proteins in the Eukaryotic Species using Multi-domain Information*

Presenter: Pooja Strobe, University of Nebraska

Authors: Pooja Strobe and Etsuko Moriyama

Abstract: *Check back for updates.*

Title: *Learning Causal Relationships between Genes from Steady State Data: Algorithms, Simulation and Application*

Presenter: Michael P. Verdicchio Arizona State University

Authors: Michael P. Verdicchio, Xin Zhang, Seungchan Kim, Chitta Baral

Abstract: Learning causal relationships between genes from steady state gene expression profiles is an important issue in bioinformatics and computational systems biology. Among the developed methods, the Inductive Causation (IC) algorithm has been proven to be effective for inferring causal relationships among variables. However, recent study in the context of gene regulatory networks shows that the IC algorithm results in low precision and recall rates. To improve the performance, we propose two algorithms, the modified IC (mIC) algorithm and the mIC_CoD algorithm, that utilize partial prior knowledge of gene topological ordering information for learning causal relationships among genes. We evaluate the performance of the algorithms on synthetic datasets and show that the precision and recall rates using the mIC and the mIC_CoD algorithms are significantly improved compared with those using the IC algorithm. We also evaluate the performance of mIC algorithm against more conventional Bayesian network (BN) inference method. The simulation study shows that the mIC algorithm outperforms the Bayesian network method in both precision and recall rates. We further apply the algorithms on a melanoma microarray dataset, and identify several important causal relationships within a network of genes. Among the discovered connections, the causal relationships associated with WNT5A, a gene playing an important role in melanoma, are supported by literature. Current efforts involve further comparisons of the four algorithms (IC, mIC, mIC_CoD, BN) across synthetic data sets of varying numbers of variables and samples, as well as further validation against current real-world data.

Title: *Using the OpenSocial Platform to Develop Open, Exchangeable Web Based Biomedical Research Applications*

Presenter: Joshua Waymire, Indiana University

Authors: Joshua Waymire, Brandon Peters, Jessica L. Dantzer, Sean D. Mooney

Abstract: Using the OpenSocial platform to develop open, exchangeable web based biomedical research applications Joshua Waymire, Brandon Peters, Jessica L. Dantzer, Sean D. Mooney Laboratree is an online research management system that promotes collaboration through social networking and utilizes Google's new OpenSocial platform to support the social networking culture. The OpenSocial platform employs XML files to house JavaScript and XHTML to build applications that can be embedded in OpenSocial compliant sites. The JavaScript makes sharing of information between sites and colleagues possible within an OpenSocial application. OpenSocial applications can be uploaded to any site that uses the OpenSocial platform (Laboratree, Orkut, etc). As a proof of concept, we have developed three scientific OpenSocial applications: a PubMed browser, MutDB MiniSearch, and an interface to the Open-Biomedical Annotator from the National Center for Biomedical Ontology at Stanford. The PubMed browser allows users to query PubMed's databases for relevant publications and to bookmark the results. These bookmarks can then be shared with the user's colleagues. The MutDB MiniSearch connects to the Mooney Lab's MutDB (<http://mutdb.org/>) to query its databases and provides links to relevant results. The Open-Biomedical Annotator application is a tool that utilizes the National Center for Biomedical Ontology's Open-Biomedical Annotator through web services. Text is submitted to the application to be formatted and finally processed by the annotator and relevant keywords/terms are returned. Results are displayed in a user friendly tabular format on an OpenSocial canvas. We believe the use of the OpenSocial platform will better enable scientists and researchers to manage their data and laboratories. All applications can be viewed at <http://laboratree.org/>.

Title: *Metagenomic Signatures of 86 microbial and Viral Metagenomes*

Presenter: Dana Willner-Hall, San Diego State University

Authors: Dana Willner-Hall, Rebecca Vega Thurber, Forest Rohwer

Abstract: Previous studies have shown that dinucleotide abundances capture the majority of variation in genome signatures and are useful for quantifying lateral gene transfer and building molecular phylogenies. Metagenomes contain a mixture of individual genomes, and might be expected to lack compositional signatures. In many metagenomic datasets the majority of sequences have no significant similarities to known sequences and are effectively excluded from subsequent analyses. To circumvent this limitation, di-, tri-, and tetranucleotide abundances of 86 microbial and viral metagenomes consisting of short pyrosequencing reads were analyzed to provide a method which includes all sequences that can be used in combination with other analysis to increase our knowledge about microbial and viral communities. Both principal components analysis and hierarchical clustering showed definitive groupings of

metagenomes drawn from similar environments. Together these analyses showed that dinucleotide composition, as opposed to tri- and tetranucleotides, defines a metagenomic signature which can explain up to 80% of the variance between biomes, which is comparable to that obtained by functional genomics. Metagenomes with anomalous content were also identified using dinucleotide abundances. Subsequent analyses determined that these metagenomes were contaminated with exogenous DNA, suggesting that this approach is a useful metric for quality control. The predictive strength of the dinucleotide composition opens the possibility of identifying invasive genes and/or genomes at the metagenome level and to assign ecological classifications to unknown fragments. Environmental selection may be responsible for this dinucleotide signature either through direct selection of specific compositional signals, or alternatively by promoting the increased abundance of a few dominant taxa.

Title: *Copy Number Probe Selection Using a MIP Assay*

Presenter: Brant Wong, Affymetrix, Inc.

Authors: Brant Wong, Chris Davies

Abstract: Designing an assay to detect copy number variations can be a challenging process. The assay must be responsive to variations in the concentration of target in the sample while avoiding cross hybridization to pseudogenes commonly associated with copy number regions. Using molecular inversion probes (MIPs) as our assay, we can successfully prioritize probe candidates in copy number regions and select the best probes to be used for the assay. Although MIP probes are normally used to assay polymorphisms, our proposal uses MIP probes to target non-polymorphic locations. Some rules were created to eliminate probe candidates that would not perform very well. Most rules were an attempt to reduce cross-hybridization effects. These included candidates with a repeat region close to the interrogation position or candidates with multiple genomic hits of shortmers within the probe. To maintain probe fidelity, any candidate was removed that had a SNP near the interrogation base. A score was also applied to each candidate based on the alignment of the probe to the entire genome. The score gave higher points for alignments with base matches immediately surrounding the interrogation base to reflect the dynamics of the MIP assay. Candidates with high scores (other than the hit to its expected location) were removed from the candidate pool. The remaining candidates were prioritized for selection. Priority was first given to probes with lower scores for its best hit aside from its expected location. Then priority was then given to probes with lower average scores. Lastly, priority was given to probes with fewer total number of hits. In order to spread probes across the length of the copy number region, the entire region was divided into partitions. For each partition, the highest priority probe was selected. This was repeated until we had at least the target number of probes for the copy number region. This resulted in a set of high quality probes that evenly spanned the entire copy number region. The probes minimized cross-hybridization and are designed to detect both deletions and duplications of copy number variants.

Title: *Gene Silencing Score*

Presenter: Abdullah al Zahrani, Fish Farming Centre

Authors: Zerubbel Ezekiel, Abdullah al Zahrani

Abstract: Gene silencing score refers to biological system based formula derivation for creating a computational gene silencing score if used in invivo studies. The algorithm is based on the formula that has an interface for selecting the target gene , target host and would show the user a score based on analysis with foothold on Blast e value ,EMBL Interpro protein protein interaction ,SCOP HMM e value and siRNA site detection. The selected formula works with the logic that the gene selection score is directly proportional to SCOP Evalue expressed as absolute dividnt of one and inversely proportional to blast e value expressed as absolute dividnt of one. The value is multiplied with EMBL interpro PPI as multiples of one and the target gene product protein function is given a score by the user judged based on its candidature as a target achiever.

Title: *Improving the Prediction of Regulatory SNPs Using Functional Information*

Presenter: Yiqiang Zhao, Indiana University

Authors: Yiqiang Zhao, Matt Mort, David Cooper, Sean Mooney

Abstract: Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation and can effect gene expression, transcript processing and protein function. SNPs that are located in promoter regions might be functional by affecting regulation of gene transcription. However, identification of expression affecting SNPs is challenging for many reasons including our lack of quantitative models of cis acting regulation, linkage disequilibrium and an abundance of neutral variants. In this work, we are attempting to identify functional SNPs from non-functional SNPs in the putative transcription regulatory region (defined as 2500bp upstream the TSS and 500 bp downstream TSS) using supervised machine learning methods. When using a number of experimentally validated SNP features on set of 338 annotated functional SNPs from the Human Gene Mutation Database (HGMD) as well as random controls, we found the distance to transcription start site was of great importance. To test whether regulatory SNPs tend to occur within functionally important genes, we extend our approach by incorporating features of gene expression, codon usage and functional complexity. With the incorporation of this information, we achieved a receiver operator characteristics plot AUC of 89%, which is better than using SNP features alone. Our results suggested functional SNPs in the putative transcription regulatory region would be better predicted when both SNP and gene information were considered.

Title: *Decision Tree and Neural Network Based Cancer Diagnosis Tool Canny Cancer Detector*

Presenter: Faiz ul haque Zeya, Bahria University

Authors: Faiz Zeya, Nimra Sikandar, Haris Vohra, Samad Bukhari

Abstract: This paper is about a machine learning application Canny Cancer detector which is a cancer diagnosis tool to distinguish cancerous from non cancerous tissue samples. Algorithms from area of machine learning and statistics are applied to the gene expression, which is the translation of information encoded in a gene into protein or RNA, acquired by using microarray technology to classify tissue sample. Two methods have been implemented in classification of the tissue samples namely the decision tree and the neural network. By comparing the result of both algorithms, it has been concluded that neural network classifier with back propagation learning technique and feed forward architecture produces better performance than decision tree learning algorithm.