



POSTER ABSTRACTS (Updated 12/6/05)

Presenter: Aimale, Valerio

Affiliation: SeiraD, Inc

Title: SCANS: System for Compression and Analysis of Nucleotide Sequences

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

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

Presenter: Caporaso, Gregory

Affiliation: University of Colorado Health Sciences Center

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

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

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

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

Presenter: Doderer, Mark

Affiliation: University of Texas at San Antonio

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

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

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

Presenter: Gabow, Aaron

Affiliation: University of Colorado at Denver and Health Sciences Center

Title: Visualizing Large, Multiscale Biological Systems

Author(s): Aaron Gabow, Lawrence Hunter

Abstract: Biologists often visualize protein-protein interaction data by mapping proteins to nodes in a graph, interactions to edges, and use an automated graph layout algorithm to position the nodes. While this method can create picture that meets basic aesthetic concerns, it is not good at conveying information. The graph can have thousands of nodes -- too many data points shown to effectively search for a particular node, and the graphic does not lend itself to an immediate summary or meaning. We have integrated several ideas from graph theory and graph visualization to create an application capable in presenting large-scale biological interaction data.

Many biological data sets can be modeled as small-world networks, and often these ones are multi-scale when clustered. We use this hierarchical decomposition to assist in layout and to provide a representation that gives an effective overview of a graph while allowing the user to pick particular areas of interest.

Presenter: Hart, Reece

Affiliation: Genentech, Inc.

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

Author(s): Reece Hart

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

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

Presenter: Joslyn, Cliff

Affiliation: Los Alamos National Laboratory

Title: Mathematical Techniques for Predicting and Analyzing Ontological Protein Function Annotations

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

Abstract: The Protein Function Inference Group (PFIG) at the Los Alamos National Laboratory (LANL) has developed an approach to automatically produce novel Gene Ontology (GO) functional annotations of proteins based on categorizing the regions of the GO to which similar (in some sense) proteins are annotated. Current input spaces include proteins which are near neighbors in BLAST space, or which are described by similar terms occurring close together in documents. Validation of this or similar methods depends on the development of evaluation

metrics which are appropriate to the mathematical structure of the space of annotations, in this case, the directed acyclic graph (DAG) structure of the GO. In this talk, we first outline the LANL architecture for ontological function annotation, implemented within our POSet Ontology Laboratory Environment (POSOLE), and for both the CASP and BioCreative evaluations. We present some test results from the BLAST-based effort. We then discuss our novel evaluation metrics, and conclude with a consideration of their applicability to the general problem of measuring the consistency of annotation sets.

Presenter: Lapadat, Razvan

Affiliation: UCDHSC- Department of Pharmacology

Title: Transcriptional Control and Behavioral Changes in Selectively Bred Mice

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

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

Presenter: Lovett, Lionel

Affiliation: Jackson State University

Title: RobustMap: A Fast and Robust Algorithm for Dimension Reduction and Clustering

Author(s): Lionel F. Lovett, II

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

Presenter: Philip, Gayle

Affiliation: National University of Ireland

Title: Bacterial Genes in a Eukaryotes

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

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

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

Presenter: Rachidi, Thami

Affiliation: 1Department of Chemistry and Biochemistry, University of Colorado, Boulder

Title: Semi-Automation of Hydrogen Exchange Data Analysis

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

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

Presenter: Schlueter, Shannon

Affiliation: Iowa State University

Title: Dynamic Gene Annotation and Analysis at PlantGDB

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

Abstract: The intricacies of gene annotation are among the most complex problems being addressed by modern genomics. Albeit the current methods of gene structure annotation are vastly more accurate and provide more complete coverage than those employed less than five years ago, the support of annotations on a per-gene basis differs extraordinarily. The level of confidence for individual gene annotations can be directly attributed to the verifying presence of expressed sequence alignments. The dependence of gene structure annotation on available EST and cDNA sequences makes all static estimations of gene structure problematic. For this reason, methods of maintaining gene annotation must acknowledge confidence in a predicted gene structure. These methods must also incorporate dynamic reporting of the evidence supporting each annotated property. To provide confidence estimators and evidence reporting for current gene annotations we developed GAEVAL, the Genome Annotation Evaluation Algorithm. This system has been integrated with the existing xGDB genome data browser and tools for community curation of gene annotations at www.PlantGDB.org.

Presenter: Sivachenko, Andrey

Affiliation: Ariadne Genomics, Inc.

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

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

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

Presenter: Souaiaia, Mourad Tade

Affiliation: Loyola Marymount University

Title: Bayesian Inference in Simple Visual Perception

Author(s): Geoff Ghose , Mourad Tade Souaiaia

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

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

stimuli multiplied by the distribution of stimuli. In Matlab we obtain performance curves for two different distributions.

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

Presenter: Sugnet, Chuck

Affiliation: Affymetrix

Title: Whole Genome Transcript Analysis with Affymetrix Exon Microarrays

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

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

Presenter: Tenku, Kemeni

Affiliation: University of Colorado, Denver

Title: Determination and Analysis of Genes Involved in the Cleft Lip/Palate Defect

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

Abstract: Birth defects affect approximately 5% of all infants in the USA. The cleft lip and/or palate (CL/P) is one of these defects and it affects roughly 1/1100 and 1/1600 births, respectively. The "non-syndromatic" CL/P (nsCL/P) accounts for ~70% of all cases. This is a non-Mendelian multifactorial disorder due to interaction of multiple genes and environmental factors during fetal craniofacial development. Herein we present data-mining results of microarray data from a mouse model. Three tissues—frontonasal, lateral nasal and maxillary

prominences—during critical periods of orofacial developments (ED 10.0 to ED 12.5) were selectively examined, since this is the period relevant to the development of the orofacial cleft. By normalizing, filtering and using different statistical multiple comparison correction test methods on the data, the differences in gene expressions were evaluated and gene lists of statistically significant genes, potentially contributing to the defect, were produced. Through programming tools, the most up- and down-regulated genes at the different developmental stages and locations were selected. These genes were analyzed via Onto-Express, a Gene Ontology analysis tool, to determine their putative functions. Preliminary results demonstrated that DNA binding, protein binding, transcription factor activity and structural molecule activity are involved in the process of craniofacial development.

Presenter: Wagner, Chad

Affiliation: San Diego State University

Title: Protodist: An Analysis of Error

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

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

Presenter: Weimer, Bart

Affiliation: Utah State University

Title: Integrating Statistical Analysis of Gene expression Data onto Metabolic Pathways Facilitates Understanding of Gene Expression in the Metabolic Context

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

Abstract: The majority of the genes within a microbial genome are linked to metabolic reactions important in intermediary metabolism and survival. The understanding of gene expression profiles and regulation is greatly facilitated by appropriate statistical analysis in concordance with methods to display the data. The display of gene expression data over time from a time series experiment is critical for the biologist to quickly visualize the directions in which whole pathways progress in response to time and other physiological factors such as metabolite concentration or stress. Physical mapping of expression data to pathways is non-trivial and is

extremely time consuming. Here we describe the analytical tools built from resources available in the public domain to meaningfully depict pathways and gene expression data in concordance with the proper statistical analysis. Tools for statistical analysis of gene expression using repeated measures were developed using Bioconductor. Interfaces were created that are accessible through the Apple Bioinformatics cluster server at the Center for Integrated BioSystems. Bioconductor was also used to draw gene expression maps that were overlaid to pathways from Pathway Tools using Perl scripts to integrate the pathways and heat maps in a single visualization file. Differential color display of the gene labels was used to show genes that significantly changed over time. We used this tool to greatly facilitate analyzing, displaying, manipulating, and understanding microarray data more conveniently for the biologist to make metabolic conclusions quickly.

Presenter: Yoon, Kihoon

Affiliation: The University of Texas at San Antonio

Title: Analysis of human promoters and gene expressions by an integrative approach: constructing an index toward gene expression patterns.

Author(s): Kihoon Yoon, Stephen Kwek

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

Presenter: Zhang, Jiexin

Affiliation: The University of Texas M.D. Anderson Cancer Center

Title: Inferring Three-way Gene Interactions from Microarray Data Sets

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

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