9th ANNUAL INSTITUTE FOR GENOMICS AND BIOINFORMATICS (IGB) BIOMEDICAL INFORMATICS TRAINING (BIT) PROGRAM SYMPOSIUM

Donald Bren Hall - Room 6011

Thursday, May 19 2011

Session 1: High-Throughput Sequencing and Applications

AREM: aligning short reads from ChIP-Sequencing by expectation maximization. Daniel Newkirk, **Jacob Biesinger**, Alvin Chon, Kyoko Yokomori and Xiaohui Xie

High-throughput sequencing coupled to chromatin immuno- precipitation (ChIP-Seq) is widely used in characterizing genome-wide binding patterns of transcription factors, cofactors, chromatin modifiers, and other DNA binding proteins. A key step in ChIP-Seq data analysis is to map short reads from high-throughput sequencing to a reference genome and identify peak regions enriched with short reads. Although several methods have been proposed for ChIP-Seq analysis, most existing methods only consider reads that can be uniquely placed in the reference genome, and therefore have low power for detecting peaks located within repeat sequences. Here we introduce a probabilistic approach for ChIP-Seq data analysis which utilizes all reads, providing a truly genome-wide view of binding patterns. Reads are modeled using a mixture model corresponding to K enriched regions and a null genomic background. We use maximum likelihood to estimate the locations of the enriched regions, and implement an expectation-maximization (E-M) algorithm, called AREM, to update the alignment probabilities of each read to different genomic locations.

Detecting Transposable Elements Insertions in *Drosophila melanogaster.* **Julie M. Cridland** and Kevin R. Thornton

In *D. melanogaster* transposable element (TE) insertions can have phenotypic consequences for an individual and individual *Drosophila* can differ substantially in the composition and location of TEs in their own genome. This means that TEs can be an important source of phenotypic differences between individuals. High-throughput, paired-end sequencing methods have provided a way to detect TE insertions on a whole genome scale.

We have developed a pipeline for using paired-end sequencing data to detect TEs and determined the TE profile for 162 *Drosophila melanogaster*. Fifteen of these lines have been kept in a laboratory environment for > 50 years and 147 of these lines are from the DGRP (Drosophila Genetic Reference Panel) population which were collected from a natural population collected in Raleigh, North Carolina. We have examined these genomes for which elements are inserting in the genome and the genomic regions in which these elements insert. We have found many more TEs present in the laboratory lines than the DGRP lines. We have also found an excess of elements that are at a particular genomic location in only one individual indicating that TE insertions are usually deleterious.

RNA-seq analyses of blood-induced changes in gene expression in *Aedes aegypti*. Bonizzoni, Mariangela, **Dunn, W Augustine**, Campbell, Corey L, Olson, Ken E, Dimon, Michelle T, Marinotti, Osvaldo, James, Anthony A

Hematophagy is a common trait of insect vectors of disease. Extensive genome-wide transcriptional changes occur in mosquitoes after blood meals, and these are related to digestive and reproductive processes, among others. Studies of these changes are expected to reveal molecular targets for novel vector control and pathogen transmission-blocking strategies. The mosquito *Aedes aegypti*, a vector of Dengue viruses and other pathogens, is examined for genome-wide changes in gene expression following a blood meal.

Transcriptional changes following a blood meal in *Ae. aegypti* females were explored using RNA-seq technology. Over 30% of more than 18,000 investigated transcripts accumulate differentially in mosquitoes five hours after a blood meal when compared to those fed only on sugar. Forty transcripts accumulate only in blood-fed mosquitoes. The list of regulated transcripts correlates with an enhancement of digestive activity and a suppression of environmental stimuli perception and innate immunity. The alignment of more than 65 million high-quality short reads to the *Ae. aegypti* reference genome permitted the documentation of errors in the current annotation of transcript boundaries, as well as the discovery of novel transcripts, exons and splicing variants. *Cis*-regulatory elements and *cis*-regulatory modules enriched significantly at the 5'-flanking sequences of blood meal-regulated genes were identified.

Session 2: Gene Regulation and Disregulation

High-throughput analysis of cohesin-mediated gene regulation in the CdLS mouse model. Daniel Newkirk, Richard Chien, Aniello Infante, Kyoko Yokomori, Xiaohui Xie

Cohesin is an essential complex required for sister chromatid cohesion and chromosome segregation in mitosis. However, mutations of the cohesin loading factor Nipbl and cohesin subunits were found to cause the human developmental disorder Cornelia de Lange Syndrome (CdLS), strongly suggesting a critical role of cohesin in developmental gene regulation. In order to understand the transcriptional role of cohesin in CdLS pathogenesis, we performed chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-Seq) on mouse embryonic fibroblasts (MEFs) derived from wild type and Nipbl (+/-) mice that closely model human CdLS. Since cohesin also binds to repeat sequences, which may be important for its transcriptional activity, we developed a new tool called AREM to analyze repetitive regions traditionally excluded in ChIP-Seq analysis. We mapped genome-wide cohesin binding sites in the wild type and mutant MEFs. By correlating cohesin binding with MEF expression array data, we showed that cohesin sites are strongly associated with genes that are significantly dysregulated in the mutant, and that cohesin binding was indeed decreased at some of the most affected genes, such as the Pcdhb cluster and adipocyte-related genes. Our results indicate that the reduction of genome-wide cohesin binding causes expression abnormalities and underlies the pathogenesis of CdLS.

Transcriptional regulation of mammary gland development as a model for breast cancer. Michael L Salmans, Padhraic Smyth, Bogi Andersen

Mammary gland branching morphogenesis is driven by terminal end buds (TEBs), stem cell-enriched spherical structures at the ends of the growing ducts whose proliferative and invasive nature makes them an excellent model for oncomechanisms. To study the role of transcriptional regulation during mammary gland development we generated a mouse model expressing a dominant negative Co-factor of LIM (CLIM) protein, a transcriptional co-regulator required for branching morphogenesis. We performed a timecourse microarray analysis to characterize gene expression profiles from TEB and duct cells during four developmental stages of puberty. Through this analysis we have gained insights into (a) the transcriptional networks involved in normal mammary gland development; (b) the gene expression profiles that characterize TEB and duct cells; (c) the correlation between breast cancer gene signatures and TEB and duct gene signatures. We found a high correlation of the TEB gene signature with aggressive breast cancers, suggesting that while the TEB proliferates and invades in a controlled manner, it has cancer-like properties. We also identified the genes regulated by CLIM that are required for branching morphogenesis. Interestingly, CLIM is a direct transcriptional regulator of *Her2, Her3,* and *Fgfr2,* which are essential signaling proteins in mammary gland development and carcinogenesis.

Learning Epigenetic Markers of Cell-Specific Gene Expression. Corey Schaninger, Padhraic Smyth, Bogi Andersen

Embryonic stem cells generate multiple cell types during development that constitute different body tissues capable of carrying out diverse functions. This cellular specification is achieved by differential gene expression across the multiple cell types. The state of the chromatin is thought to affect transcription factor binding affinity and the ability of transcription factors to interact, which in turn affects gene expression. The presence of epigenetic markers, such as histone modifications, can signify the chromatin's state; these features vary amongst different cell types.

Some current studies use supervised learning to generate chromatin modification patterns for known promoter regions, while others focus on finding global chromatin signatures using unsupervised methods similar to motif finding. In our study, we will focus on cell-specific differences in expression to generate chromatin modification patterns for known promoter regions. We will use genome-wide histone modification data obtained by the BROAD Institute for 14 different cell types. The goal of this research is to produce interpretable rules for epigenetic markers by isolating genes expressed exclusively in a specific cell type and learning rules in a probabilistic manner. Our approach will employ supervised learning techniques, such as probabilistic decision trees, with limitations on the number of subtrees per rule.

Requirement of chromatin modifications for epigenetic switching in *Candida albicans*. Zhiyun Guan, Su Zhao, Qing Nie, Haoping Liu

Candida albicans, a major human fungal pathogen, switches stochastically between the distinct and epigenetically heritable white and opaque phases with different advantages in adapting to host niches. The master regulator *WOR1* is specifically expressed in opaque state and essential for opaque establishment and maintenance. It has been proposed that white-to-opaque switching occurs when *WOR1* expression surpasses a threshold, and a network of positive feedback loops of *WOR1, EFG1,* and other transcriptional regulators sustains opaque state. By tracking population-wide *WOR1* expression and cell fate commitment, we find cells transition to high *WOR1* expression twice during induced white-opaque switching, yet only those in the second transition commit to opaque, which is not predicted by the threshold model. The *WOR1* expression pattern and time of opaque commitment are not altered in the *efg1* mutant, yet impaired in some histone modification mutants. We propose that epigenetic switching from the white to opaque chromatin state of the *WOR1* promoter go through two transitional chromatin states. We are developing a mathematical model, in which modification to neighboring nucleosomes through cooperative interactions drives transitions between chromatin states, to simulate the observed temporal dynamics of *WOR1* expression and delayed fate commitment.

Additionally, we are experimentally determining the transitional chromatin states at the *WOR1* promoter.

Session 3: Protein-Protein Interactions and Target Identification

Computational Prediction and Experimental Verification of New MAP Kinase Docking Sites and Substrates Including Gli Transcription Factors and Smoothelin-like 2. Elizabeth A. Gordon, Vishal R. Patel, Thomas C. Whisenant,' Robyn M Kaake, Lan Huang, Pierre Baldi and Lee Bardwell

To understand signaling networks, new methods are needed to identify novel kinase substrates. Spatial regulation of MAP kinase signaling occurs at multiple levels: in addition to subcellular compartmentalization, MAP kinases make extensive use of docking and scaffolding interactions to bind their regulators and substrates. We have developed a hybrid computational search algorithm that combines machine learning and expert knowledge to identify novel MAP kinase docking sites (D-sites), and used this algorithm to search the human genome. Predictions were tested by peptide array followed by rigorous biochemical verification with in vitro binding and kinase assays. We identified several new D-site-dependent MAPK substrates, including the hedgehog-regulated transcription factors Gli1 and Gli3, suggesting there may be a direct connection between MAP kinase and hedgehog signaling. This finding has potential translational relevance to pancreatic cancer, gastric cancer, melanoma, and several other tumor types. Another novel substrate we discovered is a relatively poorly characterized member of the Smoothelin family, SMTNL2. In the case of SMTNL2 we identified phosphorylation on residues in close proximity to the docking site and showed that they were MAPK dependent in cell culture. In humans, SMTNL2 expression correlates with aerobic capacity, and is downregulated in Duchennes muscular dystrophy (DMD). These and other new substrates are being further characterized in vivo using cell-based assays and fluorescent imaging methods.

Developing an integrated cross-linking mass spectrometry approach to analyze the structure and topology of the dynamic yeast 26S proteasome. Athit Kao, Scott Rychnovsky, Pierre Baldi, Lan Haung

The 26S proteasome is a vital multisubunit protein complex in eukaryotes responsible for routine protein degradation. Proteasomal dysfunction has been implicated in many diseases such as neural degeneration, making it an ideal drug target. A complex's structure and interactome are crucial in understanding function, especially during perturbation of cellular homeostasis. To glean structural information from complexes, the maturing field of cross-linking mass spectrometry aims to address inherent limitations in legacy structural methods. Chemically cross-linking residues and identifying spatially neighboring cross-linked peptides allows for a practical workflow to generate distance constraints for modeling. Current methods are time-consuming and labor intensive, but we have designed and synthesized new reagents with improved functionality along with developing new computational tools to improve the efficiency of the traditionally complicated task of cross-link identification. Using yeast 26S proteasome, we have identified tens of new cross-linked peptides unambiguously, which has not been feasible as of yet within utmost certainty for this complex. Thus, unlike most commercially available reagents and publically available computational tools, our protocols allow for a more precise workflow to unambiguously determine cross-linked sites. Identification of novel protein interaction interfaces will provide a new molecular basis for designing mechanism-driven therapeutic strategies targeting the proteasome.

Session 4: Chemoinformatics and Drug Discovery

Learning to predict chemical reactions. Matthew A. Kayala, Chloe-Agathe Azencott, Jonathan H. Chen, and Pierre Baldi

Predicting the course of chemical reactions is essential to the practice of organic chemistry, with applications ranging from improving drug synthesis to understanding the origin of life. Previous computational approaches are not high-throughput, are not generalizable or scalable, or lack sufficient data. Here, we describe a new approach to reaction prediction. Using a physically motivated conceptualization, we describe mechanistic reactions as interactions between coarse molecular orbitals (MOs). Using an existing rule-based system, we derive a restricted dataset of 2989 productive and 6.15 million unproductive mechanistic steps. And from machine learning, we pose identifying productive reactions as a ranking problem: given input reactants and conditions, learn a ranking model over potential MO interactions such that the top-ranked yield the major products. Our artificial neural network based implementation follows a two-stage approach. We first train atom-level reactivity classifiers to filter the vast majority of non-productive reactions. Then, we train ranking models on pairs of interacting MOs to learn a relative productivity function over mechanistic steps. Our trained models exhibit close to perfect recovery of the rule-based labels. Furthermore, the ranking system correctly predicts multi-step reactions and shows promising generalizability, making reasonable predictions cases not handled by the rule-based expert system.

When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. **Ramzi Nasr** and Pierre Baldi.

As repositories of chemical molecules continue to expand and become more open, it becomes increasingly important to develop tools to search them efficiently and assess the statistical significance of chemical similarity scores. Here, we develop a framework for modeling, predicting, and approximating the distributions of chemical similarity scores and their extreme values in large databases. From the distributions of the scores and their analytical forms, *Z*-scores, *E*-values, and *p*-values are derived to assess the significance of similarity scores. In addition, the framework also allows one to predict the value of standard chemical retrieval metrics, such as sensitivity and specificity at fixed thresholds, or receiver operating characteristic (ROC) curves at multiple thresholds, and to detect outliers in the form of atypical molecules. Numerous and diverse experiments that have been performed, in part with large sets of molecules from the ChemDB, show remarkable agreement between theory and empirical results.

Title Forthcoming **Paul Rigor**

Although there are several open-source and commercially available computational tools for virtual drug screening -- including Dock, Autodock and Schroedinger's Maestro; there is still a lack of a more general, tool agnostic and scalable framework that is able to leverage the advantages offered by readily available docking and molecular dynamics programs in a high-performance computing (HPC) environment. We have developed a framework built on top of an HPC pipeline and existing proteomics and chemical informatics tools -- such as ChemDB, COSMOS, SCRATCH -- to support an iterative virtual screening methodology. We have applied our approach to two candidate therapeutic targets involved in cancer metastasis: 1) Annexin A2 and 2) S100A4/metastasin. Further, we await the preliminary experimental validation of our *in silico* predictions. Moreover, future extensions to the pipeline and related machine learning tools will be discussed.

High-Throughput 3D Structure Prediction of Small Molecules. Peter Sadowski, Arlo Randall, Pierre Baldi

The next generation of virtual screening systems will not be limited by current libraries of known molecule structures. Databases such as PubChem contain accurate 3D structures for millions of small molecules, but the next step is to computationally explore the much larger space of virtual molecules which have never been created. State of the art methods for predicting such structures use quantum mechanics methods which are accurate but slow. Faster prediction tools such as the commercial CORINA and the free Open Babel are less accurate and are unable to make predictions for most organometallic molecules. A system already developed at UC Irvine, named COSMOS, improves upon these methods by treating virtual molecules as assemblies of smaller molecular fragments. With a large set of pre-computed fragment structures, COSMOS can predict 3D structures for millions or billions of virtual molecules and potential drugs by intelligently combining these fragments. Here we are using highly accurate quantum mechanics methods to pre-compute a large open-source library of fragment structures, which will enable us to provide COSMOS as a free tool to the scientific community.

Prediction of molecular composition of organic aerosols using chemoinformatics approaches. **David R. Fooshee**, Tran B. Nguyen, David L. Bones, Sergey A. Nizkorodov, Patrick J. Roach, Julia Laskin, Alex Laskin, and Pierre Baldi

Organic aerosols (OA) play an important role in controlling the earth's climate, and can significantly affect public health in urbanized areas. Tools for studying OA, such as high resolution mass-spectrometry, reveal a surprisingly high level of molecular complexity in OA. A single organic precursor can produce OA containing hundreds of different compounds; ambient OA is even more complex due to additional starting materials and photochemical "aging" processes. Although starting conditions and endpoint products for OA formation may be empirically known, the models explaining the transition between these two points are often highly simplified, and omit photochemical aging entirely. In this paper we take a new approach. We explicitly specify the reactions by which a set of OA precursors may react with each other, including photochemical aging. We then use a computational chemical brewing system that we developed to iteratively apply the specified reactions to all starting molecules and pairs of molecules, to all the products of these reactions, and so on, generating tens of thousands of structures. Finally, we compare these *in silico* results with experimental results from high resolution mass-spectrometry, and suggest molecular structures and synthesis pathways for the experimentally observed molecular formulae.

Session 5: Modeling Systems and Diseases

Computational prediction of NOTCH targets in the C. elegans stem cell niche. Michael Zeller, Olivier Cinquin, Pierre Baldi

In NOTCH signaling in the C. elegans germline, membrane bound ligands of the Glp-1 family are triggered by contact with the somatic distal tip cell. Through cleavage of the NOTCH ligand, the transcription factor LAG-1 is activated in the nucleus, negatively regulating meiotic entry of the germ precursor cells Z2 and

Z3, through a number of downstream 3' UTR binding proteins: GLD-1, FBF-1/2, MEX-3, among others. In an attempt to identify all of the remaining NOTCH germline targets, we mapped all of the known 3' UTR binding protein consensus sequences to the WormBase WS220 release of the C. elegans genome. We used the latest 3' UTR annotation information on 17091 of the 47361 predicted and known genes to predict binding sites of the known 3' UTR binding proteins. LAG-1 binding sites were mapped to regions surrounding gene transcription start sites. Quantifying LBS alone, 5955 genes are identified as potential targets of NOTCH. In combination with predicted 3' UTR binding sites and filtered for germline expression, we predict ranked sets of genes, among which GLD-1 (both a direct and indirect NOTCH target) is included. Further experimental validation is being done using RNA interference.

Extensions to the Sigmoid Modeling System and kMech: An Enzyme Mechanism Modeler. **B. Compani**, T. Su, I. Chang, P. Baldi, E. Mjolsness

Motivation: Progress in systems biology critically depends on developing scalable informatics tools to model, visualize and simulate complex biological systems. Flexibly storing information about these systems, and models of these systems, is an essential tool for the facilitation of research. Simulatable models of these systems can allow researchers to make targeted decisions about their experimental designs, potentially reducing costly and unnecessary wet bench exploration. This can pave the way for a more enhanced and efficient research cycle in a broad spectrum of biological research fields, and be of particular utility in the next generation approach to the drug discovery process.

Here we describe a generalized version of the kMech enzyme mechanism modeling tool. With this utility approximately forty existing enzyme mechanism models expressed explicitly in the kMech/Sigmoid platform can be expressed implicitly by a single parameterized input notation. Subsequent sub reactions can be generated procedurally and all potential “new” kMech enzyme mechanisms that follow the previous motif do not have to be explicitly created.