MOLECULAR BIOINFORMATICS FOR DISEASE

YVES A. LUSSIER*, YOUNGHEE LEE
Center for Biomedical Informatics and Section of Genetic Medicine, Dept. of Medicine, Computational Institute,[4]Institute for Genomics and Systems Biology, and UC Cancer Research Center; The University of Chicago, IL, 60637, U.S.A.

PREDRAG RADIVOJAC*
School of Informatics, Indiana University Bloomington, IN 47408, U.S.A.

YANAY OFRAN*
Faculty of Life Sciences, Bar Ilan University, 52900 Israel

ATUL BUTTE*
Depts. of Pediatrics and of Medicine, Stanford University, Sanford, CA 94305, U.S.A.

MARICEL KANN*
Dept. of Biological Sciences, University of Maryland, Baltimore County, MA, USA

This session focuses on the emerging field of molecular bioinformatics techniques applied to understand disease, and establish better prognosis or treatments. With the advent of whole genome association studies, there is a clear need to amass and think about and relate phenotypes in a computational way. For example, Brunner and van Driel showed how syndromes sharing clinical features could be related to the same protein complex (Nature Reviews Genetics 2004 5:545). While the application of methods from the domain of systems biology to the phenotypes of model organisms has enabled the prediction of molecular pathways and gene functions, it is not yet obvious how such methods will scale to phenotypes and diseases of higher organisms. Investigating methods to improve either genetics or phenomics methods independently carries enormous scientific and societal merit in this quest for disease genes. Yet taken together, powerful joint methods are likely to emerge such as phenome-anchored whole genome association studies, where prior probabilities stemming from independent studies can be used to increase, *ab initio*, the sensitivity of association studies or interpret the associations a posteriori through meta-analyses.

* Session co-chairs

The accepted papers for the session on Molecular Bioinformatics for Diseases comprises original research projects that span human genome variation, host-pathogen interactions, phenotypic relationship, and proteomics.

The first two papers address methodological improvements in human genome variation studies and provide insight in the molecular mechanisms human diseases. **Daemen et al.** proposed a new methodology in which copy number variations (CNV) resulting from array Comparative Genomic Hybridization (CGH) are transformed into features for classification purpose using a special variant of hidden Markov Models (HMM) able to capture recurrent copy number alterations by coupling the HMMs of individual samples. They demonstrate that the combination of HMMs for the detection of copy number alterations with Weighted Least Squares Support Vector Machines (LS-SVM) classifiers offers a novel methodological approach for classification based on copy number alterations and also suggest that copy number variations are not sufficient in survival-related studies and that other high-throughput data sources should be investigated. Conversely, **Zhang et al.** were interested in the effects of genetic variation across multiple loci. They developed an algorithm, FastChi, that utilizes an upper bound of the two-locus chi-square test in that an upper bound enables the algorithm to perform the two-locus chi-square test on a smaller prioritized number of candidate SNP pairs. Moreover, they contribute to solve the computational challenge of controlling for multiplicity of tests in genome-wide association study. The computational aspect of the chi-square test was examined on a combined dataset consisting of SNPs from the Broad/MIT mouse database and of the GNF mouse dataset. They demonstrate their method is orders of magnitudes faster than the classical - brute force - alternative.

The second group consists of three papers that explore associations between host-pathogen with particular focus on intracellular host proteins that relied on extensive data integration. The pathogen species ranged from viruses (e.g. HIV, Influenza) to prokaryotes (e.g. *Vibrio cholera*). As protein-protein interactions (PPI) between human pathogens and their host are vital at every step of the pathogen life cycle, pathogens are and particularly viruses are dependent on human host proteins. However, the novelty of some studies is to introduce a systems biology approach to the problem by focusing on the human PPIs or pathways. Two research groups, **Tastan et al.** and **Kanabar et al.,** both focused on uncovering host-pathogen PPIs and associated protein networks. **Tastan et al.** propose a supervised learning framework by utilizing multiple information data sources such as co-occurrence of functional motif, interaction domains, ontology, posttranslational modifications and class, gene expression profiles, topological properties, and the similarity of HIB-1 proteins to human proteins' known binding partners. They trained and tested a Random Forest (RF) classifier with

this extensive feature set. This approach achieved 23% of average Mean Average Precision (MAP) score. This work is among the first to predict the global set of interactions between HIV-1 and human host cellular proteins. In contrast, **Kanabar et al** apply a probabilistic model, the Joint Intervention Network (JIN), to the relative gene expression changes of a set of indicator genes in regulatory network of *V. cholerae* to discover mechanisms critical to its environmental persistence. This method implements an epistasis analysis for quantitative, multivariate phenotypes. They analyzed the expression response of several indicator genes under various single and multiple knock-outs of three known biofilm regulators (e.g. hapR, upsR, and upsT). Another other described by **Kasson and Pande**, proposes the use of mutual information to examine the functional regulation of influenza hemagglutinin. The approach consist of employing mutual information over structure-based protein sequences from crystallographic data which allows to extract clusters of concerted mutation sites and comparison of influenza isolates from two distinct Influenza strains, 5N1 and human H3N2, which yield substantial differences in spatial localization of the clustered residues of proteins sequences abstracted from crystallographic data. They suggest that the residues identified on common between H5N1 and H3N2 may be important to the basic functional roles of hemagglutinin and also be useful targets for future mutagenesis experiments. Therefore, the combined sequence-information and structural analysis help generate novel functional hypotheses that would not occur via either method alone.

Finally, the last paper describes the regulatory network of transcriptional factors (TF) related to breast cancer metastasis classifier. **Lim and et al**. aim to develop a computational systems biology approach that infers robust prognostic markers by identifying upstream “Master Regulator Genes”, causally related to the presentation of the phenotype of interest. Recently the research group of **Dr. Califano** introduced ARACNe (Nature Genetics 2005 37:382), an algorithm for the dissection of transcriptional networks that can infer the targets of TF from microarray expression profile. ARACNe is interrogated for the unbiased inference of TFs that induce or suppress specific gene signatures associated with poor prognosis in breast cancer.

Acknowledgements

The session co-chairs would like to thank numerous reviewers for their help in select the best papers for the session Molecular Bioinformatics for Disease and Dr. Andrea Califano for his participation as a Session Keynote.

4

MOLECULAR INTERACTION NETWORK BASED DISCOVERY OF MASTER REGULATOR GENES

ANDREA CALIFANO
KEYNOTE SPEAKER FOR
"MOLECULAR BIOINFORMATICS FOR DISEASE"
Center for Computational Biology and Bioinformatics, Dept. of Biomedical Informatics, Columbia University, 1130 Saint Nicholas Avenue, New York, NY 10032, USA,

The identification of genes acting synergistically as master regulators of physiologic and pathologic cellular phenotypes is a key open problem in systems biology. Here we use a molecular interaction based approach to identify the repertoire of transcription factors (TFs) of a master regulatory module responsible for synergistic activation of a tumor-specific signature. Specifically, we used the ARACNe algorithm and other computational tools to infer regulatory interactions responsible for initiating and maintaining the mesenchymal phenotype of Glioblastoma Multiforme (GBM), previously associated with the poorest disease prognosis. Expression of mesenchymal genes is a hallmark of aggressiveness but the upstream regulators of the signature are unknown. Starting from the unbiased analysis of all TFs, we identify a highly interconnected module of six TFs jointly regulating >75% of the genes in the signature. Two TFs (Stat3 and C/EBPb), in partic ular, display features of initiators and master regulators of module activity. Biochemical validation confirms that the TFs in the module bind to the inferred promoters in vivo and ectopic expression of the master TFs activates expression of the mesenchymal signature. These effects are sufficient to trigger mesenchymal transformation of neural stem cells, which become highly tumorigenic in vivo, and promote migration and invasion. Conversely, silencing of Stat3 and C/EBPb in human glioma cells leads to collapse of the mesenchymal signature and reduction of tumor aggressiveness. Our results reveal that activation of a small transcriptional module is necessary and sufficient to induce a mesenchymal phenotype in malignant brain tumors.