Pacific Symposium on Biocomputing 13:228-230(2008)

MOLECULAR BIOINFORMATICS FOR DISEASE: PROTEIN INTERACTIONS AND PHENOMICS

YVES A. LUSSIER*, YOUNGHEE LEE[1]
Center for Biomedical Informatics and Section of Genetic Medicine, Dept. of Medicine and UC Cancer Research Center; The University of Chicago, IL, 60637

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

YANAY OFRAN*
Dept. of Biochemistry & Molecular Biophysics, Columbia University New York, NY 10032, U.S.A.

MARCO PUNTA*
Biochemistry & Molecular Biophysics, New York, NY 10032, U.S.A.

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

MARICEL KANN *
National Center for Biotechnology Information, NIH Bethesda, MD 20894, U.S.A.

This session focuses on the emerging fields of protein interactions in diseases and phenomics: from protein-protein interactions to supracellular phenotypes. Experimental studies indicate that protein interactions play a key role in many diseases, even in some that are considered complex or multifactorial. While altered phenotypes are among the most reliable manifestations of altered gene functions, research focused on systematic analysis of phenotype relationships to study human biology is still in its infancy. In this summary, the words phenome and phenomics are used to describe the physical totality of all traits of an organism (Mahner, J Theor Biol 1997 186:55-63). The audiences targeted by this session bring together a broad audience: bioinformaticians, systems biologists, biomedical informaticians, physicians, pharmacologists, computer scientists, statisticians, members of the pharmaceutical industry and others to share their experience and scientific findings in this area.

The papers accepted for the session on Molecular Bioinformatics for Diseases comprise original research that pertain to biological scales ranging from phenomics, or

[1] co-author of the manuscript
[*] Session co-chairs and co-authors

relationship of whole organism's phenotypes, to proteomics. More specifically, they capitalize on novel computational methods and technological developments in bioinformatics that analyze disease or disease-associated phenotypes in a biology scale between the nanoscale where protein domains, reactions to a scale below the organism taken as a whole where disease phenotypes and phenomes are observed. .

The first two papers address human disease with phenomic datasets ranging from clinical conditions, laboratory biomarker, to the electronic medical record (EMR). **Alterovitz et al**. worked on an information theoretic framework (entropy) for discovery of novel biomarkers, integrating biofluid (e.g. blood, urine) and tissue information. This study uses 26 proteomes from 45 sources to identify candidate biofluids and biomarkers responsible for functional information transfer in the tissue domain. Among results are significant associations between tissues (e.g. cerebrospinal fluid, saliva, urine) and biomarkers (e.g. EGFR, BRCA1). A novel multipartite network of tissue-biomarker-biofluid is proposed as well as candidate biomarkers of biofluid/tissues that may have clinical applications. **Chen et al**. combined gene expression measurements and patient data from the hospital electronic medical records to examine clinical proxies for maturation and associated genes. The method was used to compare trends among different clinical laboratory test in response to an increase in age. They propose the lymphocyte count as a proxy measure for aging and infer that correlated expression of genes in the EMR are also implicated in the process of aging.

The next four papers explored associations between microarrays and an original set of approaches to the systems biology of phenotypes or literature-based statistical approaches to provide phenotypic meaning to gene expression. Since microarray data analysis often predicts many false positive genes, **Hu et al**. propose a general framework to discover the relation between two or more disease conditions in human. They applied networking pathways to association study of non-insulin dependent diabetes mellitus and obesity. Their methods involved the integration of numerous databases including microarrays with KEGG pathways. Diabetes mellitus and Obesity-associated pathways are presented. **Badea et al.** propose a clustering algorithm which is capable of simultaneously factorizing two distinct gene expression datasets. The aim of the algorithm is to uncover gene regulatory programs that are common to the two phenotypes. The methods were evaluated with gene expression profiles that are common between the more homogeneous pancreatic ductal adenocarcinoma (PDAC) and the more heterogeneous colon adenocarcinoma. The approach identified that the PDAC signature is active in a large fraction of colon adenocarcinoma. **Gevaert et al.** present an approach to integrate information from literature abstracts into probabilistic models of gene expression data in order to improve model building in gene selection. They investigated if a Bayesian network model with a text prior can be used to predict the prognosis in cancer.

Finally, these three papers consist of two proteomic studies and one DNA methylation study applied to phenomic datasets. **Sridhar et al**. developed a branch-and

bound algorithm to formulate the optimal enzyme combination identification problem as an optimization problem on metabolic networks. They demonstrate that the algorithm can accurately identify the target enzymes for known successful drugs in the literature and reduce the total search time by several orders of magnitude as compared to the exhaustive search. In contrast, **Singh et al.** were interested in global alignment of multiple protein-protein interaction (PPI) networks. They developed an algorithm that maximizes the overall match across all "input" networks. It was applied to the global alignment of protein-protein interaction networks from five species: yeast, fly, worm, mouse and human. The authors propose an original way to unveil functional orthologs cover multiple (5) species. **Kim et al.** use GpG flanquing sequence composition to predict methylation susceptibility and identify susceptible methylation sites in disease-related tissues (e.g. primary leukemia, lymphoma cells and normal blood lymphocytes).

Acknowledgements

The session co-chairs would like to thank numerous reviewers for their help in selecting the best papers among many excellent submissions and Dr. Yong Huang for his suggestions.