Pacific Symposium on Biocomputing 7:3-5 (2002)

HUMAN GENOME VARIATION: DISEASE, DRUG RESPONSE, AND CLINICAL PHENOTYPES

FRANCISCO M. DE LA VEGA
Applied Biosystems, 850 Lincoln Centre Drive,
Foster City, CA 94404, USA

ISAAC S. KOHANE
Children's Hospital Informatics Program & Harvard Medical School,
300 Longwood Avenue, Boston, MA 02115, USA

JULIE A. SCHNEIDER and J. CLAIBORNE STEPHENS
Genaissance Pharmaceuticals, Inc.,
Five Science Park, New Haven, CT 06511, USA

With the completion of a rough draft of the human genome sequence in sight, researchers are shifting to leverage this new information in the elucidation of the genetic basis of disease susceptibility and drug response. Massive genotyping and gene expression profiling studies are being planned and carried out by both academic/public institutions and industry. Researchers from different disciplines are all interested in the mining of the data coming from those studies; human geneticists, population geneticists, molecular biologists, computational biologists and even clinical practitioners. These communities have different immediate goals, but at the end of the day what is sought is analogous: the connection between variation in a group of genes or in their expression and observed phenotypes. There is an imminent need to link information across the huge data sets these groups are producing independently. However, there are tremendous challenges in the integration of polymorphism and gene expression databases and their clinical phenotypic annotation

This is the third session devoted to the computational challenges of human genome variation studies held at the Pacific Symposium on Biocomputing[1,2]. The focus of the session has been the presentation and discussion of new research that promises to facilitate the elucidation of the connections between genotypes and phenotypes using the data generated by high-throughput technologies. Nine accepted manuscripts comprise this year's original work presented at the conference.

A major incentive for collecting genetic variation data is to use this information to identify genomic regions that influence disease susceptibility or drug response. In this volume, Zhang *et al.* outline a new approach to identify clinically relevant genes that produce quantitative phenotypes. Although similar methods have been developed to measure the strength of association between haplotypes and binary (case-control) data, Zhang *et al.*'s method is particularly valuable because many

important clinical phenotypes display quantitative inheritance. On the other hand, the manuscript of Moore and Hahn introduce a novel computational approach using cellular automata (CA) and parallel genetic algorithms to identify combinations of SNPs associated with clinical outcomes. They use a simulated dataset of a discordant sib-pair study design to demonstrate that the CA approach has good power to identify high-order nonlinear interactions with few false-positives. Given the current uncertainties on the genetic architecture underlying complex disease[5], it is critical to develop new approaches, such as the CA advanced by the authors, that can test for association in the presence of allelic heterogeneity[6] and epistatic interactions between loci.

Large quantities of DNA sequence variation data is needed to better understand the contribution of genetics to human disease, drug response, and clinical phenotypes. In order to insure the quality of these data, fully automated genotyping processes are required: from assay design, assay validation, assay interpretation, quality control, to data management and release. One of the major challenges involved in developing a streamlined, high-throughput genotyping is creating appropriate software to support the system. In their conference paper, Heil *et al.* describe the components of a successful, ultra high-throughput genotyping process developed at Celera Genomics. Their approach could be an excellent starting point for those involved in developing similar infrastructures elsewhere.

How to properly store and combine complex biological data is an extremely important subject in the post-genome era. Among the challenges to develop an efficient data or knowledge base are the diversity of semantics, potential uses, and data sources. Ontologies have been successfully applied in the past to develop knowledge base systems to store complex data, such as the Gene Ontology for gene annotations[3], and RiboWeb[4] for capturing experimental results in scientific literature. The contributions of Rubin *et al.* and Oliver *et al.* to this conference present a successful application of ontologies on genotype-phenotype data in relation to clinical drug response. The approach used in "PharmGKB" presented by the authors address many of the complex problems arising when retrieving data from diverse genomics and clinical databases, and when updating links to external database domains. Their methodology may be very helpful for making the diverse genomics data better suited for scientific analysis.

Molecular profiling is a tool that is gaining acceptance to classify tissue samples and other clinical outcomes based on gene and potentially protein expression profiles. Its accuracy depends on the appropriate analysis of the resulting datasets, and typically involves multivariate statistics and other machine learning techniques. The paper of Ben-Hur *et al.* describes an algorithm to investigate the stability of the solutions of clustering algorithms. The authors apply their method to the hierarchical clustering of microarray and synthetic data. On the other hand, Ghosh applies a regression analysis to data that has been first

transformed by Singular Value Decomposition (SDV), for uncovering possible relations between microarray expression data of tumor samples and tumor diagnosis. The problem is a novel application for SVD, which has been recently applied to microarray data in a different but complementary approach. The paper of Potter and Draghici addresses a clinically important problem: classification of HIV protease's resistance to IC90 drug solely from protein sequences. Their contribution shows that improved accuracy can be achieved by combining SOFM classifiers.

As high-throughput genotyping and expression-measurement methodologies are applied to large populations, the opportunity now exists to use existing clinical phenotypic annotations (i.e., the extended medical record) in the analysis of the relationship between genotype/haplotype variation and phenotype. Typically, however, the forward link is sought, leading from genetic variation data to the inference of clinical phenotypes. The paper of Malin and Sweeney in this volume offers instead a reverse approach, allowing the inference of genetic variability data based on clinical phenotypes. In this unusual approach, clinical/hospital/claims data is brought together with phenotype/genotype through the use machine learning techniques to predict the underlying genotype.

Acknowledgments

We would like to acknowledge the generous help of the anonymous reviewers that supported the selection process for this session, as well as the panelists that joined us to discuss the challenges in this field.

References

1. F. M. De La Vega, and M. Kreitman. “Human genome variation” In: *Pacific Symposium on Biocomputing 2000*, R.B. Altman *et al*. (Eds.). World Scientific Press, Singapore (2000).
2. F.M. De La Vega, M. Kreitman, and I. S. Kohane. “Human genome variation: Linking genotypes to clinical phenotypes” In: *Pacific Symposium on Biocomputing 2001*, R.B. Altman *et al*. (Eds.). World Scientific Press, Singapore (2001).
3. The Gene Ontology Consortium. “Creating the gene ontology resource: design and implementation” *Genome Res.* **11(8)**, 1425-1433 (2001).
4. R.O. Chen, R. Feliciano, R.B. Altman. “RIBOWEB: linking structural computations to a knowledge base of published experimental data” In *Proc Int Conf Intell Syst Mol Biol* **5**, 84-87 (1997).
5. A.F. Wright and N.D. Hastie. “Complex genetic diseases: controversy over the Croesus code” Genome Biology **2(8)**, comment 2007.1–2007.8 (2001).
6. J.K. Pritchard. “Are Rare Variants Responsible for Susceptibility to Complex Diseases?” *Am. J. Hum. Genet.* **69**,124–137 (2001).