Pacific Symposium on Biocomputing 14:353-355 (2009)

KNOWLEDGE-DRIVEN ANALYSIS AND DATA INTEGRATION FOR HIGH-THROUGHPUT BIOLOGICAL DATA

M. F. OCHS[1], J. QUACKENBUSH[2], R. DAVULURI[3], H. RESSOM[4]

[1] *The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205, USA*, E-mail: mfo@jhu.edu,

[2] *Dana Farber Cancer Institute, Harvard Medical School, Boston, MA, USA*, E-mail: johnq@jimmy.harvard.edu

[3] *The Wistar Institute, Philadelphia, PA, USA*, E-mail: rdavuluri@wistar.org,

[4] *Lombardi Cancer Center, Georgetown University, Washington, DC, USA*, E-mail: hwr@georgetown.edu,

Keywords: Genomics, Controlled vocabulary, Database, Bayesian analysis

Introduction

There is a data explosion overtaking medicine and biology due to the development of high-throughput technologies for measuring most molecular constituents of cells. Novel sequencing approaches are now nearing the $1000 genome. SNPchips already measure individual genetic variation for the same cost, and microrarrays provide a global picture of the transcripts, with miRNAs and exon level analyses beginning. Proteomic measurements, both from mass spectrometry and from antibody arrays, are probing increasing amounts of the proteome, and NMR techniques are doing the same for metabolic components. The goal is a deeper view into the complexity of biological systems, with the ultimate prize being personalized medicine and finely targeted therapies.

The deluge of data has created new problems, some held in common with older fields dealing with high-throughput data (e.g., particle physics) and others unique to the field (i.e., context dependent data from cell types, environment, etc.). Already it is impossible for any individual or research group to review the available data, bring it together, integrate it across the various molecular domains either directly, such as DNA exons to specific mRNA variant, or in context, such as by pathway, and analyze it. It is necessary to develop methods to encode data in forms usable by automated systems, to encapsulate our knowledge of the biology and medicine, and to leverage this knowledge during analysis.

Data integration can be done at a number of levels.[1] Encoding of data will ideally use a formal ontology that encodes our knowledge,[2] and at least will utilize standards as provided by controlled vocabularies[3] or mediated schemas.[4] Future data systems are likely to comprise both data warehouses, for subsets of data within a single institution,[5] and data federation, for automated querying of data from multiple sources.[6]

This session focuses on steps along the path to the integration of genomic data, the automated capture of biological information, the development of analyses that leverage biological knowledge, and the creation of integrated and federated data resources.

Papers

The first three papers in this session focus on integrating data from multiple data sources. Carey and Gentleman present a new data structure within the R/Bioconductor system that integrates SNP, expression, annotation, and phenotypic data and which links to the many existing Bioconductor tools for analysis, as well as to the UCSC genome browser for visualization. Hart and Mukhyala describe the UNISON database system and web interface, which integrates proteomics data from many sources, precomputes structures and other values, and provides a rich set of query tools. Shen-Orr and colleagues introduce a knowledgebase of the adaptive immune system and its cytokine mediated cell-cell interactions. This knowledgebase was generated through text mining using an approach focused on a minimal number of terms and their synonyms.

The next two papers integrate data across multiple species to gain insight into cellular processes. Bidaut and Stoeckert integrate expression data from mouse and human tissues using homologene and cell hierarchy. They then utilize an artificial neural network classifier to deduce transcriptional signatures for different points along the differentiation of stem cells. Fox, Taylor and Slonim link protein interaction data across four species and identify novel features of hub proteins, showing that the interaction partners of hub proteins are more conserved across evolution than other proteins.

The final four papers utilize prior information from biological studies to refine biomarker discovery and genome wide association studies. Phan and

Pacific Symposium on Biocomputing 14:353-355 (2009)

colleagues refine support vector machine learning of biomarkers by modifying the distance metric comparing expression profiles based on reliable biomarkers reported in the literature. Webb-Robertson and colleagues use a Bayesian framework to integrate proteomic and metabolomic data for improved detection of pathogens. Bush, Dudek, and Ritchie address the combinatoric explosion in GWAS when looking for combinations of SNPs linked to a disease by focusing analysis on SNPs related to pathways or genes linked to the disease under study. Pan and colleagues use local phylogenic alignments in GWAS to guide the choice of regions under study.

These papers demonstrate all aspects of the focus of this PSB session. While it is very early in the era of knowledge-driven analysis in the biomedical sciences, these papers show that difficult problems are beginning to be solved by new approaches that rely on data integration. These approaches include both tools for linking diverse data across molecular domains and species and methods that leverage prior knowledge to provide new insights into biology. As our data resources continue to grow exponentially, such methods will grow in importance and should provide ever greater insight.

References

1. B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy and P. Tarczy-Hornoch, *J Biomed Inform* **40**, 5 (2007 Feb).
2. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat Genet* **25**, 25 (2000 May).
3. O. Bodenreider, *Nucleic Acids Res* **32**, D267 (2004 Jan 1).
4. R. Shaker, P. Mork, M. Barclay and P. Tarczy-Hornoch, *Proc AMIA Symp* , 692 (2002).
5. R. Dhaval, J. Buskirk, J. Backer, C. K. Sen, G. Gordillo and J. Kamal, *AMIA Annu Symp Proc* **11**, p. 939 (2007).
6. N. Stolba, T. M. Nguyen and A. M. Tjoa, *Conf Proc IEEE Eng Med Biol Soc* , 4355 (2007).