Pacific Symposium on Biocomputing 13:137-140(2008)

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

M. F. OCHS*

The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University,
Baltimore, MD 21205, USA
**E-mail: mfo@jhu.edu*
http://www.cancerbiostats.onc.jhmi.edu/ochs.cfm

J. QUACKENBUSH*

Dana Farber Cancer Institute, Harvard Medical School,
Boston, MA , USA
**E-mail: johnq@jimmy.harvard.edu*
http://www.hsph.harvard.edu/faculty/john-quackenbush/

R. DAVULURI*

Comprehensive Cancer Center, Ohio State University,
Columbus OH 43210, USA
E-mail: Ramana.Davuluri@osumc.edu
http://www.cancergenetics.med.ohio-state.edu/2732.cfm

Keywords: Bayesian analysis, Statistical models, Statistical data analysis, Controlled vocabulary

Introduction

New high-throughput methods have led to immense data sets in biological research. However, these data sets create significant problems for data analysis. For instance, the identification of interacting genetic variants placing individuals at risk for or providing protection from the development of polygenic diseases requires identification of sets of interacting genes. The standard statistical approaches were developed for cases of many samples and few loci of interest, and they cannot achieve power in the face of the enormous growth in our knowledge of genomics. Simple calculations show that as the number of typed loci and the number of potential interactions between genes increase, it will become impossible to design a study with sufficient statistical power using present analysis methods (e.g., with 1000 loci and 4-gene interactions there are more than 40 billion potential combinations). Similar problems exist for both microarray and proteomic data, as well as for any other high-throughput data type where a comparison is made to biological samples, which tend to be limited in number.

One potentially fruitful approach to overcoming this curse of dimensionality is to guide inference on these large data sets by inclusion of prior knowledge generated over many decades by biologists and geneticists. The knowledge can be used to develop models against which experimental data can be tested, or it can be used in the design of statistical distributions for sampling techniques. One example of such treatment is the growing use of Bayesian statistical methods coupled to Markov chain Monte Carlo techniques. However, there are now many groups working on highly diverse methods to address these problems.

A second area of active research is the integration of diverse data types, which can also be considered a use of biological information to guide analysis. For instance, in the case of polygenic diseases, it would be logical to limit the loci to be analyzed to those that link in some way to differences in gene expression between cases and controls, or to genes which encode proteins in pathways of biological interest to the disease. Thus, results from analyzing one form of data, microarrays or biological pathways, serve as prior knowledge in the analysis of a second form of data, genotypes. In addition, it is often highly desirable to use data from well studied organisms, such as fruit fly or nematode, to guide inferences in higher organisms. The need to link data across data domain, such as from a single nucleotide polymorphism (SNP) to protein interaction, requires establishing ontologies or at minimum controlled vocabularies that allow automatic linking of data elements. Linking data between species requires identification of orthologs and orthologous pathways and interactions.

This session focuses on these two broad issues in integrated data analysis: the development of analysis methods that utilize multiple types of data and the establishment of methods to integrate diverse data. The papers reflect the continuum from the SGDI tool for integrating diverse data in R to analyses of the gylcan proteome and a large genotypic data set.

Papers

The first two papers in the session provide tools for data integration during analysis. SGDI, the System for Genomic Data Integration, is built on top of the widely used R/Bioconductor framework.[1] It uses the concept of assays for high-throughput data (e.g., microarrays, SNPchips) and tightly binds phenotype data to these assays (e.g., tumor stage, patient data) to track phenotypic information throughtout the analysis workflow. An important feature of the system is the inclusion of an extended version of the Sequence Ontology providing semantic integration of the data. The NARADA system leverages molecular interactions and other annotations during the analysis of networks, such as genetic regulatory networks.[2] The primary network is projected onto the annotation space, and functional networks are deduced using annotations from this primary network created by, for instance, gene expression analysis. This converts the inference of relationships between genes or proteins to inference of relationships between biological processes.

The middle three papers of the session focus on three different issues in the analysis of diverse data. The first paper addresses the problem of building classifiers from multiple data types, here microarray and proteomics data.[3] Unlike some approaches that look at mRNA species and encoded proteins, the work presented here looks to find the best discriminatory mRNA species and independently the best protein species. A subset of these are combined within a single classifier built using least squares support vector machines (LS-SVM), providing better sensitivity and specificity with fewer overall features. The second paper provides a solution to an important problem in the analysis of large genomic data sets – lists of genes carried forward during analysis rely on thresholds, leading to loss of information and questionable use of statistical tests later in the chain of analysis.[4] This work provides an integrated probabilistic framework for appropriate inference on gene sets or pathways, including gene ontology. Statistically the approach provides a full joint probability distribution from both data and annotations for estimation of biological parameters (e.g., upregulation of a pathway). The authors apply this method in a mouse model of prostate cancer. The third paper introduces a method to look at the multiscale correlation structure between different types of data, which is important since we generally do not know the length scales of interest in genomic processes.[5] In this case, correlations between histone modifications and DNase activity and between repressing and activating histone modifications are studied. The methodology relies on wavelets to calculate correlations, Kolmogorov-Smirnof statistics to test significance between different comparisons, and permutation tests to compare the results to randomized sequences.

The final two papers introduce new analysis approaches that have the potential to include significant prior information. In the first paper, a methodology for handling the large number of potential interactions between genetic variants in genome wide association studies is presented.[6] The method relies on random draws of variants at loci and the comparison of the set of variants to phenotype. A variant that is associated with phenotype should, over many random draws, obtain a higher posterior probability of association. The distribution of variants for the random draws can reflect prior knowledge of the phenotype (e.g., pathways associated with cancer) or other independent knowledge. In the final paper, a new approach for identification of biomarkers in proteomic data is presented.[7] An ongoing issue in the field is the identification of peaks that associate with a covariate of disease (e.g., age), rather than with the disease itself. The method described here first eliminates peaks from mass spectra that correlate with noninformative parameters provided by prior information, and then isolates peaks that distinguish phenotype. The technique is demonstrated by isolating a proteomic signature distinguishing hepatocellular carcinoma and chronic liver disease.

References

1. V. J. Carey, J. Gentry, R. Gentleman and S. Ramaswamy, *Pac Symp Biocomput* (2008).
2. J. Pandy, M. Koyuturk, W. Szpankowsi and A. Grama, *Pac Symp Biocomput* (2008).
3. A. Daemen, O. Gevaert, T. De Bie, A. Debucquoy, B. De Moor and K. Haustermans, *Pac Symp Biocomput* (2008).
4. M. Bhattacharjee, C. Pritchard and P. Nelson, *Pac Symp Biocomput* (2008).
5. R. E. Thurman, J. A. Stamatoyannopoulos and W. S. Noble, *Pac Symp Biocomput* (2008).
6. M. A. Province and I. B. Borecki, *Pac Symp Biocomput* (2008).
7. H. W. Ressom, R. S. Varghese, L. Goldman, C. A. Loffredo, M. Abdel-Hamid, Z. Kyselova, Y. Mechref, M. Novotny and R. Goldman, *Pac Symp Biocomput* (2008).