
Preface

The post-genomic revolution is witnessing the generation of petabytes of information annually, with deep implications ranging across evolutionary theory, developmental biology, agriculture, and disease processes. The great challenge during the coming decades is not so much in generating the data, for that will continue at an accelerating pace, but in converting it into the information and knowledge that will improve the human condition and deepen our understanding of the world around us. A first step in meeting that challenge is to structure data so that it is easily accessed, integrated, and assimilated. Data Mining in Systems Biology surveys and demonstrates the science and technology of this important initial step in the data-to-knowledge conversion. The volume is organized around two overlapping themes, network inference and functional inference.

Network Inference

Tsuda and Georgii (*Dense Module Enumeration in Biological Networks*) discuss a rigorous, robust, and inclusive approach to inferring a particular type of network; viz, networks defined by databases that record physical interactions between proteins. Willy, Sung, and Ng (*Discovering Interacting Domains and Motifs in Protein–Protein Interactions*) discuss a method for discovering interactions between protein domains and short linear sequences, which are fundamental to multiple cellular processes. In particular, they discuss and demonstrate how to exploit the surge in structural data to infer such interactions. Mongiovì and Sharan (*Global Alignment of Protein–Protein Interaction Networks*) describe a novel method for identifying proteins that are orthologous across species. Their method is based on alignment of protein–protein interaction networks. This paper and that of Tsuda and Georgii represent a good example of the knowledge amplification that can be achieved by research on different but potentially complementary projects carried out by different labs. These three papers illustrate important directions in the discovery and analysis of *protein–protein* interactions.

While protein–protein interactions define the repertoire of cellular processes, *protein–DNA interactions* regulate those processes. In general, gene/protein networks defined by such interactions can be inferred from experimental data by various multivariate statistical methods. One of the widely used forms of inference is Bayesian probabilistic modeling. Larjo, Shmulevich, and Lähdesmäki (*Structure Learning for Bayesian Networks as Models of Biological Networks*) review recent progress in the development and application of these methods. Mordellet and Vert (*Supervised Inference of Gene Regulatory Networks from Positive and Unlabeled Examples*) discuss SIRENE, a machine learning method for inferring networks of transcriptional regulators and their targets from expression data and known regulatory relationships. Honkela, Rattray, and Lawrence (*Mining Regulatory Network Connections by Ranking Transcription Factor Target Genes Using Time Series Expression Data*) developed a reverse engineering approach to infer regulator target interactions and applied it to candidate targets of the p53 tumor suppressor promoter.

Historically, molecular biology has focused on proteins and nucleic acids. One of the major changes in the past decade has been a dramatic increase in understanding metabolism; this, of course, is also stimulated by the availability of whole genome sequence data. This constitutes the subject of *Protein–Chemical Substance Interactions*. Hancock, Takigawa, and Mamitsuka (*Identifying Pathways of Co-ordinated Gene Expression*) present a tutorial for the use of gene expression data to identify metabolic networks associated with a given condition.

More direct approaches to metabolism include an increased emphasis on the structure of complex carbohydrates. Aoki-Kinoshita (*Mining Frequent Subtrees in Glycan Data Using the RINGS Glycan Miner Tool*) describes an algorithmic method for finding frequently occurring tree structures with glycan databases, which are relevant to the binding of particular proteins. This can be thought of as the metabolic analogue to approaches that identify protein–protein and protein–DNA binding sites.

The chapter by Yamanishi (*Chemogenomic Approaches to Infer Drug–Target Interaction Networks*) discusses another kind of network, those formed by drug–target interactions. In this case, sequence and chemical structure databases provide the information that enable statistical classification methods to identify plausible drug–target interactions.

Functional Inference

The ability to predicatively localize proteins to one or another cellular compartment can generate important clues about their possible function. Imai, Hayat, Sakiyama, Fujita, Tomii, Elofsson, and Horton (*Localization Prediction and Structure-Based In Silico Analysis of Bacterial Proteins: With Emphasis on Outer Membrane Proteins*) evaluate localization prediction tools against a known dataset, and illustrate with an application to β -barrel outer membrane proteins in *E. coli*. For biological interpretation of large-scale datasets, visualization tools play key roles. Hu (*Analysis Strategy of Protein–Protein Interaction Networks*) explains how to use the multiple data sources and analytical tools in VisANT to identify and analyze networks of various kinds. Karp, Paley, and Altman (*Data Mining in the MetaCyc Family of Pathway Databases*) present an introduction to the contributions made by Karp and his colleagues over many years. The chapter is a rich source of tools and methods for mining this extensive, well-curated, and extremely important set of databases.

Approaches to genotype–phenotype correlations have evolved continuously over the past several decades. With the advent of whole genome sequencing, the search for correlations between genes and Mendelian traits accelerated enormously, but complex phenotypes, whether normal traits or diseases, find their genetic basis in sets of genes, and in particular combinations of alleles. Various procedures have been developed to infer such sets from variations in transcriptional variation. Hung (*Gene Set/Pathway Enrichment Analysis*) describes in detail how the so-called gene set enrichment analysis can be used to draw functional inferences from such transcriptional datasets. The method has been applied to identify processes that distinguish disease phenotypes from normal phenotypes. This leads to the final four chapters of the volume, which are all disease related.

Linghu, Franzosa, and Xia (*Construction of Functional Linkage Gene Networks by Data Integration*) discuss an approach to combining heterogeneous datasets in order to construct full genome networks in which each gene is surrounded by functionally related

neighbors, with the relationships specified by evidence-weighted links. Such functional linkage networks (FLNs) of human genes can uncover surprising genetic associations between phenotypically unrelated diseases and suggest that our current disease nosology may need to be reformulated.

The chapter by Yang, Kon, and DeLisi (*Genome-Wide Association Studies*) presents an overview of genome-wide association methods and explains how multiple data sources, including databases generated by high-throughput genotyping technologies, can be used to identify disease-associated chromosomal locations.

Kuiken, Yoon, Abfalterer, Gaschen, Lo, and Korber (*Viral Genome Analysis and Knowledge Management*) discuss three of the major infectious disease sequence-function databases—those for the human immunodeficiency, hepatitis C, and hemorrhagic fever viruses. The challenge here again is combining information from different sources, but in this case, integration and quality control are achieved by a continually upgraded community-developed infrastructure.

Kanehisa (*Molecular Network Analysis of Diseases and Drugs in KEGG*) presents another integrated approach where known disease genes and drug targets are integrated into the KEGG molecular network database and explains how to make use of this resource with the KEGG Mapper tool in large-scale data analysis.

We expect this book to be of interest to cell biologists and biotechnologists, as well as to the scientists and engineers developing the databases and mining and visualization systems that are central to the paradigm-altering discoveries being made with increasing frequency.

Uji, Kyoto, Japan
Boston, MA, USA
Uji, Kyoto, Japan

Hiroshi Mamitsuka
Charles DeLisi
Minoru Kanehisa

Data Mining for Systems Biology

Methods and Protocols

Mamitsuka, H.; DeLisi, C.; Kanehisa, M. (Eds.)

2013, XII, 279 p., Hardcover

ISBN: 978-1-62703-106-6

A product of Humana Press