Preface

Google queries for *systems biology* and *pathway analysis* fetch over 9 million and 14 million entries, respectively. These numbers speak volumes about the utility and popularity of systems data analysis in modern bioscience. These days, any gene expression or SNP-analyzing manuscript would feature a chapter on pathways, ontology enrichment, and/or biological networks. The application of systems biology approaches now spreads widely from basic research and preclinical drug discovery to translational research and personalized healthcare.

Systems biology "focuses on the systematic study of complex interactions in biological systems, thus using a new perspective (integration instead of reduction) to study them" (Wikipedia). From a practical standpoint, it translates as an integration of accumulated biological knowledge in a computer-readable format, followed by a creation of tools for the analysis of biological and chemical experimental data. Starting in the 1970s, biochemistry was the first field codified into databases such as BRENDA, EMP/MPW, and, later, KEGG. Over the years, a regulation and signaling components were added to biochemistry in the form of protein interaction databases such as HPRD and BIND. On top of that, comprehensive ontologies of cellular processes and protein functions were developed and integrated, the best known of which is Gene Ontology (GO).

Functional analysis is inseparable from high-throughput, or "omics"-driven experimental biology, which has been rapidly evolving since the late 1990s. At that time, the "genome-wide," noisy assays with thousands of data points were nearly illegible for a majority of wet lab researchers, in part, due to the "diaper stage" of development for the statistical tools which only helped to reduce data complexity, but largely failed to aid in understanding of the underlying biology. Gradually, bioinformaticians and wet lab biologists found efficient ways of communicating. As a result, wet lab biologists acquired the skill of using existing databases of pathways and processes for mapping and prioritization of experimental data (enrichment analysis). Later, biological networks were added to analysis toolboxes, borrowing from years of research in graph theory and physics.

Recent technological advances and scalability in next-generation sequencing (NGS) and other genomics technologies enable production of biological "big data" at unprecedented tera- and petabyte scales. Efficient mining of these vast and complex datasets for the needs of biomedical research critically depends on an integration of the clinical and omics information, sophisticated analytical tools, and taking into account prior knowledge about genotype-phenotype relationships and protein functionality. Experimental "omics" data has been accumulated in publicly available and private databases for over 20 years.

Analytical tools are described in hundreds of computational biology and bioinformatics publications and scattered across code repositories and commercial bioinformatics suites. Information about protein functionality is structured and accumulated in computer-readable format in several curated databases on protein-protein interactions, pathways, and network modules. Such curated content is then used for analysis of "omics" datasets, by means of ontology enrichment, interactome density analysis, pathways activation analysis, network modeling, and other approaches. In this book, we collected cutting-edge material on the latest methods and studies on "data-driven" and "knowledge-based" analysis from the internationally recognized leaders in this field.

This book represents a compilation of methods of functional analysis and their applications, written by experts from academy, governmental research organizations, pharmaceutical industry, and bioinformatics laboratories. It begins with the modeling of protein-protein interactions (PPIs) and protein-nucleic acids interactions as these are the building blocks of protein functionality and the most essential tools for functional analysis of large experimental datasets. There are several ways to extract protein interactions. Information on many of them is scattered in hundreds of thousands of experimental articles and can be extracted in both human- and machine-readable form. We have two chapters devoted to extracting PPIs from literature and an experimental one, using a modified yeast two-hybrid assays. In the former approach, a team from GeneGo (now acquired by Thomson Reuters) developed a sophisticated approach of structured manual annotations to assemble a comprehensive database of over 1 million experimentally proven interactions of different types.

In the second chapter, the authors present high-throughput, quantitative, yeast two-hybrid screening approach coupled with the NGS approach. This strategy allows identification of interacting proteins that are preferentially associated with a bait of interest and helps eliminate nonspecific interacting proteins.

As an example of the large-scaledata-driven network approach, we included a chapter on co-expression modules in cancer datasets. The analysis of differentially expressed gene sets (in a form of functionally related genes or pathways) in a form of either RNA-Seq or microarray experiments has been a method of choice for extracting the strongest signals from "omics" data. The authors combined an experimental approach of extracting co-expression modules from cancer expression datasets via meta-analysis with calculation of promoter motifs. Analysis of gene co-expression networks is a powerful "data-driven" tool, invaluable for understanding cancer biology and mechanisms of tumor development.

The most common and intuitive approach to functional analysis of "omics" datasets is ontology enrichment. Essentially, it consists of labeling each gene, protein, and RNA species on the experimental list with a certain functional category (cellular process, pathway, network module etc.), followed by grouping them according to the "collective" labels. The motivation behind using gene sets instead of individual genes is twofold. First, this approach incorporates pre-existing biological knowledge into the analysis and facilitates the interpretation of experimental results. Second, it employs a statistical hypotheses testing framework.

In this book, we include a comprehensive review of the Gene Set Analysis (GSA) approaches for testing differential expression of gene sets and several GSA approaches for testing statistical hypotheses beyond differential expression that allow to extract additional biological information from the data. Gene sets frequently can be analyzed as pathways. A novel algorithm OncoFinder evaluates the activation of molecular pathways on the basis of gene/protein expression data in the objects of interest. OncoFinder enables performing both quantitative and qualitative analysis of the intracellular molecular pathways. Another approach enables causal analysis of multidimensional "omics" dataset using an "upstream analysis" strategy which combines TRANSFAC database with analysis of the upstream signal transduction pathways that control the activity of these TFs. This analysis highlighted a substantial heterogeneity of specific TF-DNA binding sites in terms of their observed relative binding avidity and correlations between avidity for specific TF-DNA binding sites with the levels of mRNA transcription at the proximal gene target. Combined gene expression/promoter sequence analysis has been applied to extract novel insight from cancer biology.

Another novel method, weighted SNP correlation network analysis (WSCNA), can be used to identify SNP networks from GWAS data, create network-specific polygenic scores, examine network topology to identify hub SNPs, and gain biological insights into complex traits. An automatic annotation system (Association Rule Mining Annotator for Pathways) utilizes rule mining techniques to predict metabolic pathways across a wide range of prokaryotes. This system can be used to enhance the quality of automatically generated annotations as well as annotating proteins with unknown function.

The increasing amount and variety of data in biosciences call for innovative methods of visualization, scientific verification, and pathway analysis. sbv IMPROVER is a platform that uses crowdsourcing and verification to create biological networks with easy public access. Currently, it contains 120 networks built in Biological Expression Language to interpret data from PubMed articles with high-quality verification available for free on the CBN database. Another solution is an integrated computational platform Lynx—aweb-based database and knowledge extraction engine, which provides its users with advanced search capabilities and an access to a variety of algorithms for enrichment analysis and network-based gene prioritization. User-friendly web services and interfaces connect its users both to the Lynx integrated knowledge base (LynxKB) and integrated analytical tools.

MetaCore and Key Pathway Advisor constitute an integrated platform for functional data analysis. This platform enables analysis of sequencing data, annotation of gene variants, gene expression, proteomics, and other high-throughput (OMICs) data, which is routinely challenging because of its biological complexity and high level of technical and biological noise. We present techniques and concepts used to represent complex biomedical networks. The BioXM Knowledge Management Environment (BioMax AG, Germany) is an example of how a domain such as oncology is represented and how this representation is utilized for research. We also discuss the ArrayTrack (National Center for Toxicology Research, FDA) that is also used in the routine review of genomic data submitted to the FDA. ArrayTrack stores a full range of information related to DNA microarrays and clinical and nonclinical studies as well as the digested data derived from proteomics and metabolomics experiments.

Recent advances in genome sequencing and "omics" technologies are opening new opportunities for improving diagnosis and treatment of human diseases. The precision medicine initiative in particular aims at developing individualized treatment options that take into account individual variability in genes and environment of each person. Systems biology approaches that group genes, transcripts, and proteins into functionally meaningful networks will play a crucial role in the future of personalized medicine. By that, systems biology enables comparisons of healthy and disease-affected tissues and organs from the same individual, as well as these between healthy and disease-afflicted individuals. However, the field faces a multitude of challenges ranging from data integration to statistical and combinatorial issues in data analyses. Here, we collected computational approaches developed to tackle challenges in network analyses. Successful application of systems biology approach to psychiatric diseases opens the application part of our book. Another chapter is using an example of Alzheimer's disease to identify and analyze the candidate gene lists, and divide them up into different tiers of evidence consistency established by enrichment analysis across sub-datasets collected within the same experiment and across different experiments and platforms. Ingenuity Pathway Assistant tool was used to expand these gene lists and interpret the outputs.

One chapter is devoted to a different kind of networks, the connectome of brain cells affected in mental diseases. It has been long recognized that schizophrenia, unlike certain other mental disorders, appears to be delocalized, i.e., difficult to attribute to a dysfunction

of a few specific brain areas, and may be better understood as a disruption of brain's emergent network properties. The authors focused on topological properties of functional brain networks obtained from fMRI data, in order to demonstrate that some of those properties can be used as discriminative features of schizophrenia in multivariate predictive setting.

We have also included a chapter on in-depth clinical analysis of a particular pathway, with pleiotropic effects on key cellular functions. Wnt (Wingless-related integration site) is one of the key signaling pathways in eukaryotes, which orchestrates self-renewal programs in normal somatic stem cells as well as in cancer stem cells. Aberrant Wnt signaling is associated with a wide variety of malignancies and diseases. Although our understanding has increased tremendously over the past decade, therapeutic targeting of the dysregulated Wnt pathway remains a challenge and the effect of Wnt-targeted compounds poorly predictable. The chapter revised recent preclinical and clinical therapeutic approaches to target the Wnt pathway.

Functional data analysis is evolving quickly as a discipline. Novel network algorithms and software tools are published almost weekly, and the scope of applications expands with every new DNA, RNA, or protein assay hitting the market. Therefore, we could not and had no intention to pack as many tools as possible into this volume. Instead, we tried to focus on the established methods and software packages we see in the marketplace every day and provide readers with a broad understanding of issues and applications of this fascinating new field.

Los Angeles, CA, USA — *Tatiana V. Tatarinova*
Solana Beach, CA, USA — *Yuri Nikolsky*



http://www.springer.com/978-1-4939-7025-4

Biological Networks and Pathway Analysis

Tatarinova, T.V.; Nikolsky, Y. (Eds.)
2017, XIV, 509 p. 110 illus., 98 illus. in color. With online files/update., Hardcover
ISBN: 978-1-4939-7025-4
A product of Humana Press