PATHWAY ANALYSIS FOR DRUG DISCOVERY

Computational Infrastructure and Applications

Edited by

ANTON YURYEV
Ariadne Genomics, Inc.
Rockville, Maryland

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

PATHWAY ANALYSIS FOR DRUG DISCOVERY

Wiley Series on Technologies for the Pharmaceutical Industry

Sean Ekins, Series Editor

Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals
Edited by Sean Ekins

Pharmaceutical Applications of Raman Spectroscopy
Edited by Slobodan Šašić

Pathway Analysis for Drug Discovery: Computational Infrastructure and Applications
Edited by Anton Yuryev

PATHWAY ANALYSIS FOR DRUG DISCOVERY

Computational Infrastructure and Applications

Edited by

ANTON YURYEV
Ariadne Genomics, Inc.
Rockville, Maryland

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 527-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:
Pathway analysis for drug discovery : computational infrastructure and applications / [edited by] Anton Yuryev.
p. ; cm. — (Wiley series on technologies for the pharmaceutical industry)
Includes bibliographical references and index.
ISBN 978-0-470-10705-8 (cloth)
1. Drug development–Data processing. 2. DNA microarrays—Data processing.
3. Computational biology. I. Yuryev, Anton. II. Series.
[DNLM: 1. Drug Design. 2. Computational Biology. 3. Microarray Analysis—methods. QV 744 P297 2008]
RM301.25.P374 2008
615′.190285—dc22

2008015107

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

This book is a compilation of articles by pioneers of pathway analysis. While providing some solutions and the state-of-the-art overview, the book formulates many questions that yet to be addressed by the scientific community. The phrase "drug discovery" in the title was intended to emphasize the pragmatic approach for the book. Pathway analysis attempts on the enormous task of formalizing the molecular biological knowledge to make it suitable for predictive computation. Pathway analysis is currently in its infancy and requires the framework for thinking and development of useful applications. This framework can only be based on practical solutions that have a direct impact on human life and well-being of society. Improving human health and optimizing biological organisms for human needs are two main practical applications of molecular biology that rapidly move it away from being an academic discipline toward the application science in commercial industry.

Drug discovery industry is likely to benefit most from pathway analysis. There are two major computational challenges in the drug development. First is calculating the structure of drug molecules that have specific and predictable protein targets and therefore predictable biological effects. Second is calculating the biological effects themselves. Both tasks require extensive computational resources but use very different fundamental principals. The structural drug design is based on physics of molecular structure and interaction while calculating biological effects is based on the analysis of information flow or pathways. Even though the information flows inside the living cell through the physical interaction network, the physics of the molecular interactions has limited effect on biological pathways. Instead, the combinatorial effect from players in the protein community network determines the biological outcome of the drug treatment, disease progression, and healthy signaling throughout the human body. The computational complexity in structural drug design is due to the large number of atoms participating in the molecular interaction, while the complexity of pathway analysis is due to the large number of processes that occur in a single human cell, multiplied by the large number of tissues in human organism.

Currently used approach of "computing" the drug effects using live organisms such as animal models and patients in clinical trials is expensive, error prone, and can be viewed as unethical. Therefore, everyone appears to believe that *in silico* predictions should be the safest and most economical way of improving the drug discovery pipeline. Because the excuse of having insufficient computational resources rapidly vanishes into the history, the book attempts to summarize critical elements that are necessary for successful *in silico* pathway analysis for drug development.

Anton Yuryev

CONTRIBUTORS

Gordana Apic, Cambridge Cell Networks Ltd., St John's Innovation Centre, Cambridge CB4 0WS, United Kingdom; gordana.apic@camcellnet.com

Michael J. Birrer, MD, PhD, National Cancer Institute, Center for Cancer Research, 37 Convent Drive, Bldg 37, Room 1130, Bethesda, MD 20892; birrerm@bprb.nci.nih.gov

Fedor Bokov, Ariadne Genomics Inc, 9430 Key West avenue, Suite 113, Rockville, MD 20850; masf@ariadnegenomics.com

Andrej Bugrim, GeneGo, Inc., 500 Renaissance Drive, #106, St Joseph, MI 49085; andrej@genego.com

Bojana Cosovic, Cambridge Cell Networks Ltd, St John's Innovation Centre, Cowley Road, CB4 0WS Cambridge, United Kingdom; Bojana.Cosovic@camcellnet.com

Nikolai Daraselia, Ariadne Genomics Inc, 9430 Key West avenue, Suite 113, Rockville, MD 20850; nikolai@ariadnegenomics.com

Sreenivas Devidas, PhD, Vice President Business Development, GVK BioSciences, 5457 Twin Knolls Road, Suite 101, Columbia, MD 21045; sreeni.devidas@gvkbio.com

Zoltan Dezso, GeneGo, Inc., 500 Renaissance Drive, #106, St Joseph, MI 49085; zoltan@genego.com

Sean Ekins, PhD, DSc, Collaborations In Chemistry, 601 Runnymede Ave, Jenkintown, PA 19046, USA; ekinssean@yahoo.com

John Farley, MD, Associate Professor, Department of Obstetrics and Gynecology, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814; jfarley@usuhs.mil

Craig N. Giroux, Institute of Environmental Health Sciences, Wayne State University, 2727 Second Avenue, Detroit, MI 48201; cgiroux@genetics.wayne.edu

Iaroslav Ispolatov, PhD, Departamento de Fisica, Universidad de Santiago de Chile, Casilla 302, Correo 2, Santiago, Chile. Ariadne Genomics Inc., 9430 Key West Avenue, Suite 113, Rockville, MD 20850; slava@ariadnegenomics.com

Andrey Kalinin, Ariadne Genomics Inc, 9430 Key West avenue, Suite 113, Rockville, MD 20850; kalinin@ariadnegenomics.com

Mark P. Kühnel, EMBL-Heidelberg, Cell Biology and Biocomputing Meyerhofstrasse 1, 69117 Heidelberg, Germany; mark.kuehnel@camcellnet.com

Sergei Maslov, Department of Physics, Brookhaven National Laboratory, Upton, NY 11973; maslov@bnl.gov

Goran Medic, Cambridge Cell Networks Ltd, St John's Innovation Centre, Cowley Road, CB4 0WS Cambridge, United Kingdom; Goran.Medic@camcellnet.com

Alexander Nikitin, Ariadne Genomics Inc., 9430 Key West Avenue, Suite 113, Rockville, MD 20850; shura@ariadnegenomics.com

Tatiana Nikolskaya, GeneGo, Inc., 500 Renaissance Drive, #106, St Joseph, MI 49085; tatiana@genego.com

Yuri Nikolsky, GeneGo, Inc., 500 Renaissance Drive, #106, St Joseph, MI 49085; yuri@genego.com

Laurent L. Ozbun, PhD, Center for Cancer Research, National Cancer Institute, 37 Convent Drive, Bldg 37, Room 1130, Bethesda, MD 20892; ozbunl@mail.nih.gov

Robert B. Russell, EMBL-Heidelberg Biocomputing Meyerhofstrasse 169117 Heidelberg, Germany. Email: russell@embl.de

Sergey Simakov, Moscow Institute of Physics and Technology, Department of Applied Mathematics, Instituskii Lane, 9, Dolgoprudny, Russia; simakovss@ya.ru

Andrey Y. Sivachenko, PhD, Senior Staff Scientist, Ariadne Genomics, Inc. 9430 Key West Ave. #113, Rockville, MD 20850; andrey.sivachenko@gmail.com

Todor Vujasinovic, Helios Biosciences, 8, Avenue du Général Sarrail, 94010 Créteil, France; todor.vujasinovic@heliosbiosciences.com

Anton Yuryev, PhD, Ariadne Genomics Inc., 9430 Key West Avenue, Suite 113, Rockville, MD 20850; ayuryev@ariadnegenomics.com

André Siniša Žampera, Helios Biosciences, 8, Avenue du Général Sarrail, 94010 Créteil, France; sinisa@heliosbiosciences.com

1

INTRODUCTION TO PATHWAY ANALYSIS

ANTON YURYEV

Table of Contents

1.1 INTRODUCTION

Pathway analysis is a rapidly developing discipline that combines software tools, database models, and computational algorithms—all of which help molecular biologists to convert molecular interaction data into a set of computational models. The models are developed for better prediction of cell behavior in response to a drug, nutrients, or other outside stimuli. The

Pathway Analysis for Drug Discovery, Edited by Anton Yuryev

development of pathway analysis was triggered by the expansion of high-throughput methods and the completion of human genome sequencing project. Because of these technological advances, the emphasis of molecular biology has shifted from reductionism to system integration. Suddenly, nearly all of the components of a living cell became known and the new goal of "putting them all together" into a working computational model of the living cell is awaiting the scientific community. This model must be built by a consensus effort of all molecular biologists and will be constantly refined for a significant period of time. The first and most important goal of pathway analysis is to provide tools and infrastructure that facilitate building a consensus cell model by the collective effort of the scientific community. These tools must enable adequate data exchange, automatic data integration, communication with central public depositories of pathways, and molecular interaction information supporting consensus knowledge base building. In this review, I discuss current approaches for constructing a molecular interaction database, explain the available pathway analysis methods from the drug discovery point of view, and place pathway analysis into the historical perspective of advances in molecular biology.

1.2 METHODS TO CONSTRUCT THE PATHWAY ANALYSIS KNOWLEDGE BASE

Several layers of consensus information are necessary for pathway analysis: (1) a generally agreed-upon list of molecules; (2) a consensus global molecular interaction network; and (3) a collection of consensus pathways for known biological processes. Even though significant portion of the work to create the molecular "inventory" of the human organism has been accomplished by sequencing the human genome, it is still far from being completed. The alternative splicing and protein modification isoforms are yet to be fully cataloged. This next major challenge for this knowledge level is being addressed now by development of exon and phosphorylation microarray technologies. The consensus interaction network is being created by a combination of efforts in high-throughput experiments, prediction of interactions, and classical molecular biological and genetic techniques aimed at elucidating the function of individual proteins. Because the results of numerous small-scale experiments usually are available only in the form of scientific publications and because no central depository for molecular interactions exists in the scientific community, special text-mining techniques have been developed to extract this information into machine-readable format [1].

Every available technique to record interactions for a global network database has some degree of a false-positive rate. High-throughput methods for detection of protein–protein interactions, such as a two-hybrid screen or an identification of protein complexes by co-immunoprecipitation followed by mass spectrometry, are currently being reassessed in a panic due to an apparently worrisome 50–70% false-positive rate [2]. Even though the sources of errors are well understood for these methods, the only way to reduce them at present is to scale down these experiments, effectively reducing them back to one of the laboratory techniques and preventing their high-throughput application. Consequently, attempts to improve the reliability of these methods continue amidst criticism [3]. As a reconciliation note for all high-throughput methods, I emphasize that proteins were designed by nature to interact with each other so as to provide the structural backbone for a living organism. They literally stick to each other to be alive. Therefore, it is not surprising that every protein can interact with many different partners, and many interactions that appear as false positives are in fact true physical interactions. Yet, many of these interactions are not biologically meaningful. Some of them never occur inside the living cell because two proteins never meet each other in space or time. Even if an interaction does occur *in vivo*, it simply may not perform a biological function—it is not followed by the cascade of molecular events that is called a "cell process." In my opinion, high-throughput methods probably produce mostly correct interaction data, but additional evaluation is necessary to sort out biologically functional and meaningful interactions. The identification of biologically meaningful interactions is a separate task from measuring them. If this is the case, our frustrated energy should be diverted away from these methods and refocused on remedying our general inability to understand what each interaction means for cell physiology. In Chapter 3, about automatic pathway inference, I show that measuring and recording regulatory interactions is one way to calculate biological meaning for physical interactions.

The prediction of physical interactions is typically accomplished using sequence homology [4,5], and regulatory interactions can be calculated from time-series gene expression data using Bayesian inference [6,7]. By holding the powerful promise to "reverse engineer" biological objects, Bayesian inference has been even proclaimed as the principal method of systems biology. However, it currently suffers from the noisy and dispersed experimental data available for analysis, a lack of understanding how to construct good training sets [8], and the emerging realization of the plasticity of biological regulatory networks [9].

Peer-reviewed scientific literature is still the most important source of reliable molecular interaction data. The sheer number of scientific publications and the fact that they are written in machine non-readable format necessitated the development of methods to extract this information into machine-readable format that can be used by computational algorithms. These attempts focused on the manual recording of interactions into a database by an army of curators and, at the same time, on the development of natural language processing algorithms to read scientific papers automatically. Manual curation turned out to be a slow and expensive process that is not error-free: humans do make mistakes when both writing and reading papers. The rate of manual recording appears unable to keep up with the rate of new articles being published.

Automatic extraction of the interaction from peer-reviewed scientific literature faces its own challenges. The main advantage of automatic text-mining algorithms is the speed that allows the processing of hundreds of thousands of articles in minutes. For example, MedScan technology from Ariadne Genomics can process the entire PubMed database that contains more than 14 million abstracts in 2 days on a regular personal computer. This speed allows comprehensive coverage of the entire body of scientific literature, as well as a fairly good assessment of interaction confidence. Interaction confidence can be estimated using the frequency with which the interaction was recovered from the literature. The high misinterpretation rate that occurs during fact extraction is the biggest problem of text-mining technologies. Natural language processing algorithms that use a full sentence parsing approach have the lowest error rate, but they also have the lowest recovery rate among all methods for automatic extraction of interactions [1]. Because errors are evenly distributed among all sentences, false-positive interactions appear as interactions with a low number of supporting references. This technique can be used as a filter to eliminate erroneous interactions at the end of an extraction. Unfortunately, it effectively increases the error rate among relations with a low reference count. Recently discovered true interactions by definition have a low number of references. Thus, automatic extraction methods make the effective confidence of novel interactions even lower due to contamination by false-positive facts. For example, some linguistic patterns used by the MedScan technology extract interactions with one supporting reference with only 70% accuracy. Hence, it is difficult to use the extracted data immediately for analyzing experimental data and building pathways. Additional efforts to curate automatically extracted data are required prior to an analysis such as this one [10,11].

Improvement of all methods for recording of molecular interactions will have to continue for some time until a clear winner can emerge. The most likely outcome, however, is that all interactions for human proteins will be found by the combined effort of all these methods before a winner is determined. After the best method is apparent, interactions for new organisms will be predicted mostly from the consensus interaction network for the human organism and other model organisms. Despite a seemingly overwhelming challenge to measure all physical interactions between proteins in the human genome, this goal will be achieved within the life span of most readers of this book. Indeed, the total number of unique interactions between N proteins is equal to $N(N - 1)/2$. There are 30,000–35,000 genes in the human genome, making the total number of all possible pairwise interactions around 500 million. This number is the upper estimate for human interactome size, which includes both true- and false-positive interactions regardless of the methods used to measure and record them to the global database. The interactions for alternatively spliced proteins can be found relatively easily by calculation using protein sequence information, known interactions of the longest isoforms, and interactions from the homologs to determine the protein interacting domains. Thus, even though alternatively splicing greatly increases the number of possible protein–protein interactions, measuring and recording the interactions between splicing isoforms should not require much time and investment. To measure all 500 million possible interactions, each of 500,000 molecular biologists will have to measure about 100 interactions to achieve this goal in 1 year. Assuming there are about 50,000 research projects worldwide actually measuring protein–protein interactions, all physical interactions will be measured in 10 years. The speed of measuring the new physical interaction should gradually increase, and the actual number of interactions is smaller than 500 million. Therefore, 10 years is a very safe upper estimate for time required for recording of all physical interactions.

1.3 ORGANIZATIONAL CHALLENGES FOR CONSTRUCTING THE KNOWLEDGE BASE FOR PATHWAY ANALYSIS

The major barrier separating humankind from measuring all physical protein–protein interactions in its own species is the lack of organization and communication among individual scientists. I have reason to assume that this book will not change this situation, so research will continue as usual by measuring the same interactions multiple times in different laboratories that are trying to prove or disprove each other's theories. Typically, the authors of these studies publish only interactions that seem to support their respective favorite scientific theory or model in hopes of securing funding in the future. High-throughput methods will also continue to contribute to a significant amount of interaction data while attempting to improve their accuracy. Taking the opportunity given to me by this book, I want to join the call to release all interaction data into the public domain [2]. The relatively small organizational challenges that accompany this call include having a central authority to maintain an interaction depository, the tools for data submission and building a consensus global network database, and a method for calculating the confidence of a specific interaction from supporting evidence submitted by multiple sources.

Several public institutions have taken an early lead in the attempt to become a central authority for pathway and molecular interaction databases. Kyoto University provides the Kyoto Encyclopedia of Genes and Genomes (KEGG) database curated by its own staff. Its main disadvantage is that a system cannot be used as a depositary by external users outside KEGG. The Signal Transduction Knowledge Environment (STKE) database is maintained by the American Association for the Advancement of Science (AAAS) and contains a collection of pathways curated by scientists considered to be the top experts in the field. This database contains a small collection of highly reliable and canonical pathways and accurately reflects the current state of the art of the pathway analysis field: very few pathways are actually known and experimentally verified at present. The slow rate of curation and the absence of any formal method to create pathways, including the absence of universal identifiers for pathway components, are the main disadvantages of the STKE database. Unfortunately, the usual leaders in storage of biological information, the National Center for Biotechnology Information (NCBI) in the United States and the European Bioinformatics Institute (EBI) in the European Union, seem to be overwhelmed by the amount of sequencing data they need to maintain. Currently, they lag behind in creating a central resource for pathway information. NCBI, for example, has limited itself by integrating protein physical interaction information from public databases: Biomolecular Interaction Network Database (BIND), Human Protein Reference Database (HPRD), BioGRID, and EcoCyc. Moreover, the constant introduction of new protein identifiers by these organizations unnecessarily complicates the issue even further. Among other public sources of pathway information, I must mention the Reactome database maintained by Cold Spring Harbor Laboratory in collaboration with EBI and Gene Ontology, and the Database of Interacting Proteins (DIP) at the University of California at Los Angeles. Currently, the largest pathway and molecular interaction databases are only available commercially from privately held companies such as Ingenuity Systems, GeneGo, and Ariadne Genomics.

1.4 FROM MOLECULAR INTERACTION DATABASE TO PATHWAY COLLECTION

The collection of physical interactions is not, however, the pathway database. It merely provides the underlying network or pool of interactions necessary for pathway and network building. This pool is not likely to be 100% accurate because of all the reasons I previously mentioned. Software tools and methods developed for pathway building must take into account the reliability of each interaction when working with such a database. At this point, it is worth emphasizing the difference between a network and a pathway, because both can be built by the same software tools. A network represents a static image of all possible physical and/or regulatory interactions between biological entities, while a pathway represents how the information propagates through the network. Because information propagation is a directional process, a pathway must have entry nodes where the information flow starts and terminal points where the information flow ends. The pathway components represent the sequence of molecular events in space or time while the biological process is occurring. A network, in contrast to a pathway, can contain any relation or entity, including those that do not participate in the information flow. For example, a physical interaction network can include structural interactions, while regulatory networks can include indirect relations that actually are mediated by a set of physical interactions. Network analysis can provide important insights into biological functions. It can, for example, identify major regulators and targets in a biological process that appear as hubs—nodes with many connections for this process in the network. Hubs also can provide an idea about the information flows within the network, starting from major hub regulators and propagating toward major hub targets. Network analysis can also identify protein complexes involved in a process. Yet, a network cannot be used for dynamic modeling because it lacks one essential ingredient of any pathway: an initial signal or input.

Because a pathway is a way to represent specific events that take place after exposing a cell to an extracellular signal or environmental condition, the main task of pathway analysis is to establish methods for converting network information into a pathway. Any biological pathway is essentially an abstraction or an approximation describing the major channels of information flow through the physical interaction network. The goal of pathway inference is generating a diagram simple enough to be used in kinetic simulations yet adequately describing what happens inside a cell after the stimulation. That known canonical pathways represent the preferred path through the network was proposed some time ago [12]. If this premise is true, then a pathway is simply a path through hubs in the global regulatory network. The evidence taken from scientific literature certainly supports this view: essentially every component in any currently known canonical pathway is a hub in the global and regulatory network. Hubs should be also the first experimentally detectable components of a pathway because they are the best targets for pathway inhibition, which is the favored method to study pathways *in vivo*. For this reason, the currently known connectivity of a hub must be artificially elevated relative to other "non-hub" proteins in both physical and regulatory networks derived from experimental literature: historically, researchers first identified hubs as principal pathway components and then began identifying other proteins interacting with them. Thus, in reality, the relative connectivity of hubs may not be as high as it appears to be in the currently known network. Nevertheless, currently known hubs will still probably remain hubs even after the entire network is known.

Figure 1.1 represents the principal task of pathway analysis: converting a network into a pathway suitable for dynamic modeling. The input for pathway analysis is a network and a stimulus used to invoke the information flow. The output of pathway analysis is a pathway suitable for dynamic modeling with adequate predictive power. Where does the input network come from? From another class of high-throughput experiments aimed at measuring the state of an entire biological system, such as gene expression microarray experiments. In spite of being noisy like all other biological high-throughput methods, they provide a snapshot of a system upon exposing cells to a stimulus. Ideally, the state of a cell must be measured on several different levels such as cell transcriptome, proteome, and metabolome [13]. The current state of the art, however, produces only gene expression data with acceptable quantity and accuracy. Yet recent developments in proteomic methods measuring protein

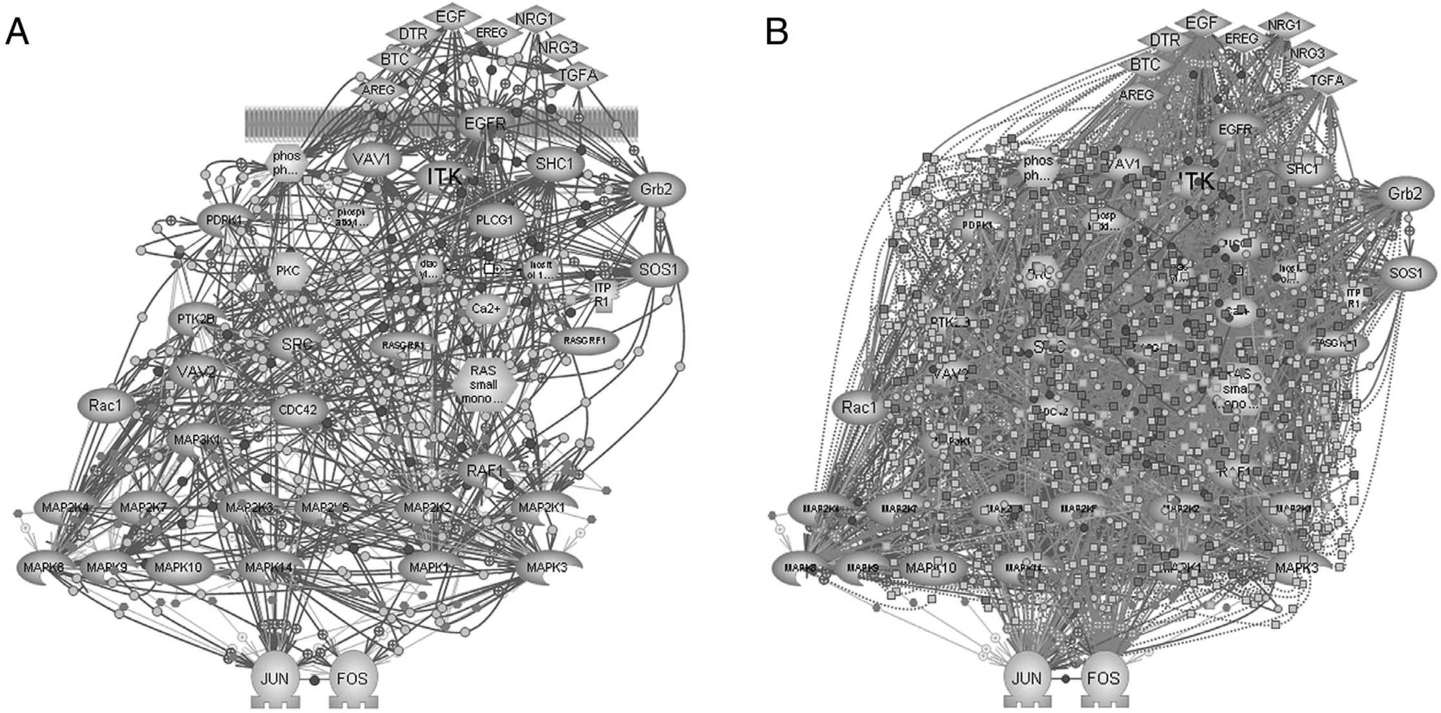
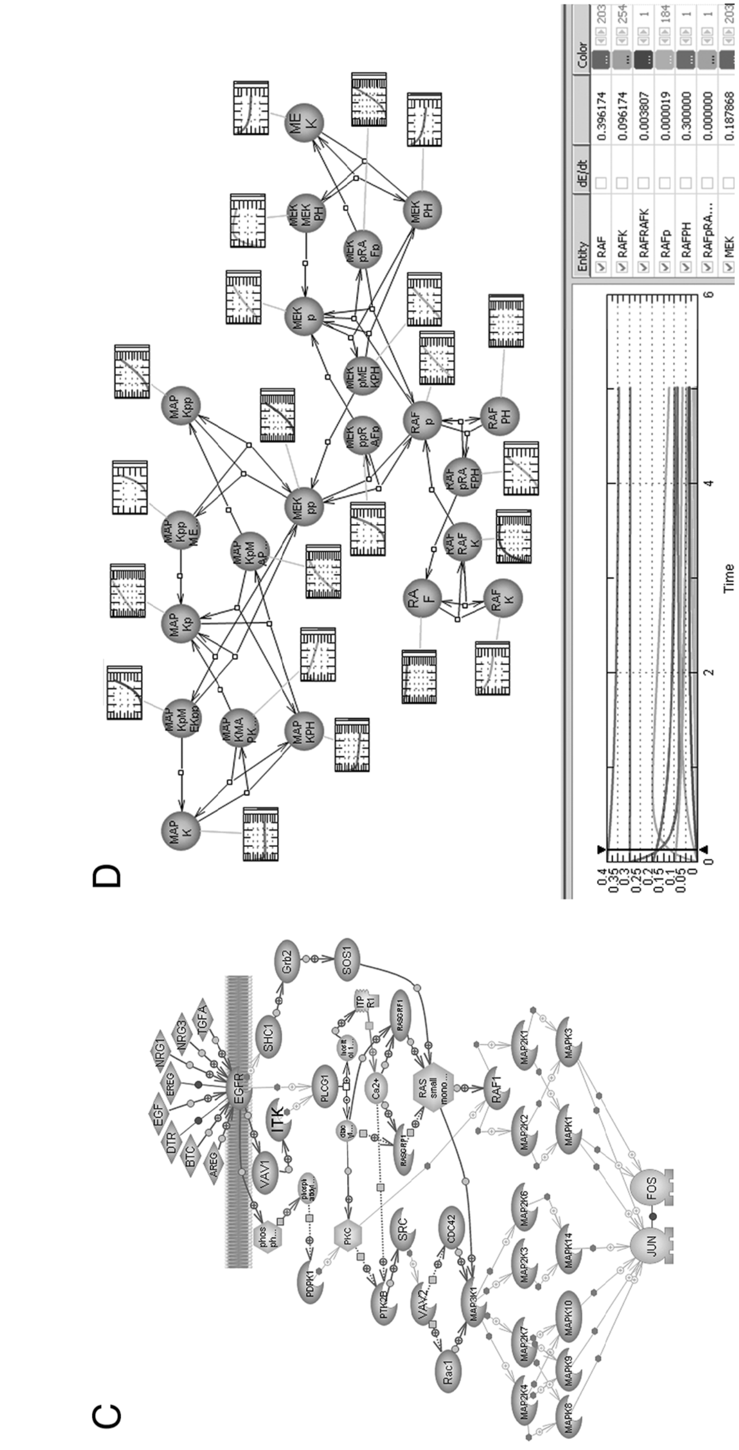


Figure 1.1 Canonical EGFR pathway shown in three ways: (A) physical interaction network; (B) physical interaction network combined with regulatory network; (C) as a pathway ready for kinetic modeling; and (D) kinetic model of MAP kinase cascade, which is a portion of EGFR pathway. One way to visualize the goal of pathway analysis is to imagine that it changes the contrast of the network image to make the main information flow more visible by hiding relations that are nonessential for depicting the principal information flow. See color insert.



Entity	dE/dt		Color
✓ RAF	□	0.396174	... 203
✓ RAFK	□	0.096174	... 254
✓ RAFRAFK	□	0.003807	... 1
✓ RAFp	□	0.000019	... 184
✓ RAFPH	□	0.300000	... 1
✓ RAFpRA...	□	0.000000	... 1
✓ MEK	□	0.187868	... 203

Figure 1.1 (*Continued*)

concentration and modification already necessitate integration of information from different experimental types into software for pathway analysis. The important workflow described above must be supported by all pathway analysis software: it must provide access to the molecular interaction database, permit analysis of high-throughput data to identify molecular networks appearing in response to an experiment, and subsequently allow calculation of pathways suitable for molecular modeling.

An additional approach for building pathways for the human organism is by using orthologous pathway information from model organisms and paralogous information about known pathways in human tissues. One of the most important achievements of network biology in the last decade is providing further support to the duplication-divergence theory of molecular evolution (see reference [14] and references therein). The best way to evolve is to duplicate an existing mechanism and then modify one or both copies to develop new functions while keeping older functions in one of the copies, if necessary. Evolution constantly duplicates individual genes and occasionally makes a copy of entire genomes in order to mutate genes later and to develop new interactions and functions [15]. As further evidence in support of the duplication-divergence model of evolution, current efforts in studying model organisms provide crucial insights into general rules for the modular and pathway organization of a cell. They have revealed and will reveal more of the conserved, “must have” mechanisms in molecular signaling and cell physiology [16]. The following list enumerates currently known conserved principles of pathway organization:

- Pathway sub-compartmentalization using clustering in physical interaction networks [17] and scaffolding [18,19]
- Fast decay of crosstalk mediated by binding interactions [20]
- Feedback loops providing positive self-activation of a pathway [18,21]
- Feed-forward loops providing noise tolerance for a pathway [22,23]
- Cross-pathway inhibition [18,22,24]

In addition, model organisms reveal the conserved molecular interaction and regulatory blocks necessary for biologically meaningful propagation of information [25], such as the MAPK-kinase cascade, for example [18].

1.5 PATHWAY ANALYSIS SOFTWARE AND THE SCIENTIFIC COMMUNITY

While providing the tools and methods for pathway building and data analysis, pathway analysis software provides additional important functions for scientific enterprise: enabling fast communication, data exchange, and education among members of the scientific community. It has been recognized that graphical and other visual information is more effective than text for learning molecular biological concepts [26]. For example, the image of the double-helix DNA structure is the most common form used by people to learn, think, and teach about DNA. This image migrates from one textbook to another. Any text describing the double-helix is merely a caption for the image. Similarly, diagram visualization in pathway analysis software allows scientists to exchange information about biological networks and work with them more efficiently. I want to demonstrate how important visualization is for pathway analysis by suggesting the following virtual experiment. First, take any pathway diagram that you find in this book and describe it in writing. Be warned that this exercise may be very boring. Second, find two of your colleagues who have never seen this pathway before. Show the diagram to the first colleague and show the text to the second one. Give both of them the same amount of time to inspect and memorize the pathway information. Then, ask them both to reproduce the pathway. You will discover that the colleague who saw the pathway will reproduce it more accurately than the person who read about the same pathway. Now, try to draw the same pathway yourself and you will realize that it is much faster to describe a pathway as text than to make a good drawing of it. Describing a pathway is easier because you most likely have a text processor program on your computer but do not have an application for pathway drawing. Imagine, however, that this pathway-drawing program exists and also enables you to send a pathway diagram to your colleagues so that they can reproduce an exact copy of your pathway, add their information to it, or compare it to their own experimental results or to other pathways. It should now be readily apparent that a pathway analysis tool can increase your ability to communicate with the scientific community and speed up your collaboration projects many times over.

Biologists usually visualize three major classes of processes as diagrams: biochemical pathways, molecular signaling cascades, and various cellular mechanisms such as a cell cycle and apoptosis. As all scientific papers are written these days using computers, pathway diagrams are drawn with computer programs as well. The degree of sophistication of these pathway-drawing programs ranges from simple vector graphics drawing tools like Microsoft PowerPoint to the database programs like Pathway Studio from Ariadne Genomics that link every pathway to the underlying global molecular interaction network and to the functional annotation of biological molecules. At the same time, these programs allow the comparison of thousands of data points from high-throughput experiments with the pathway collection in the database. The increased sophistication of pathway drawing tools has put the term “pathway analysis” on the same level with other scientific methods and disciplines that study the propagation of information inside the living cell, such as: systems biology, molecular network analysis, and dynamic modeling or kinetic simulation. In the remaining part of this introduction, I will position pathway analysis relative to these approaches in an attempt to show how pathway analysis both differs from and complements them.

1.6 PATHWAY ANALYSIS AND SYSTEMS BIOLOGY

In short, systems biology is a discipline and pathway analysis is one of its methods. Historically, the term “systems biology” was used as an umbrella to describe various attempts to understand and to model the behavior of an entire cell or organism. Since our current biological knowledge is still incomplete, systems biology focuses on the development of computational methods for analysis of high-throughput data and on designing databases and data models to store and refine the information necessary for achieving the ultimate goal of modeling cell physiology. Pathway analysis is not different in this respect from any other methods of systems biology. It allows compiling, maintaining, classification, and utilization of pathway information. The reason why pathway analysis must be isolated from other methods is evident from the following estimates. There are more than 520 signaling ligands in the human genome and 232 tissues in the human organism. Even though many tissues do not have receptors for every ligand, one hormone often can bind different types of receptors and sometimes activate different pathways [27]. Therefore, we can estimate about 100,000 signaling diagrams that are necessary in order to have a comprehensive collection of signaling pathways for the human organism. Variations of about 50 canonical biochemical pathways in 232 human tissues add another 10,000 diagrams. The number of various intracellular and physiological intercellular processes can be estimated to be about 1,000. This estimate increases the number of necessary diagrams to roughly about 200,000. Finally, there are about 2,500 complex diseases that are usually depicted as diagrams of defective pathways and cellular processes. We also must remember that pharmaceutical research typically uses animal models that require a separate pathway collection for each model organism: mouse, rat, dog, etc. All of the previous numbers estimate a daunting collection of nearly 500,000 pathways needed to build a comprehensive database for drug discovery. To create and maintain this vast pathway collection, we need a rather sophisticated software infrastructure. This software must provide interactive access to pathway diagrams, enable fast and seamless data exchange between users and databases, and allow the comparison of pathway collection with high-throughput and other experimental data. One of the goals for pathway analysis is to devise strategies and algorithms to compose such a collection and to develop an appropriate software infrastructure for its maintenance, for its utilization in analysis, and for drug discovery. The effort to create this pathway collection for drug discovery is comparable in scale to the Human Genome Project (HGP). Continuing this analogy, I would say that the current technological level of pathway analysis is about the same as the level of sequence analysis soon after Frederick Sanger proposed his method for DNA sequencing.

1.7 PATHWAY ANALYSIS AND NETWORK ANALYSIS

Briefly, network analysis studies the global properties of biological networks, while pathway analysis studies the propagation of information through a network. I have already described the pathway diagram as an intermediate step between isolating the network modules and building the dynamic model of a biological process, as shown in Figure 1.1. The method of network analysis gained momentum in 1999 after a publication by Hartwell et al. [28] that presented an intuitively clear paradigm about the modular organization inside a living cell. Since then, biological networks have undergone intense investigation. The recent efforts to analyze biological networks have yielded several important findings relevant to pathway analysis. I will list them here, since this book does not cover network analysis in much detail. First, it has been demonstrated that both physical and regulatory biological networks indeed have a modular structure that correlates well with known functional modules, as defined by humans in protein annotation [17]. Second, it was found that biological protein networks tend to have power law degree distribution, meaning that they have hubs—highly connected proteins with many interaction partners. This discovery appears to be true for both physical and regulatory networks (Figure 1.2). The main reason for having a hub or scale-free network topology is to provide robustness to the network so it does not break into independent components when a link or node is occasionally removed [29]. A node or link removal can occur because of the following: genetic mutations in evolution; somatic mutations occurring in a disease and throughout the life of the organism; and environmental conditions such as diet, trauma, and stress. The scale-free topology of a network allows biological systems to randomly try new ways of evolutionary adaptation without the danger of disintegration and to survive rather significant damage during a disease.

1.8 PATHWAY ANALYSIS OF DISEASE

Network biology logic suggests that, in order to become stable and robust, a disease network must acquire a scale-free topology with hubs. A disease is developed if its sub-network or module that carries out the malignant function becomes robust and perpetual. These perpetual networks may take years to develop in an organism. Genetic predisposition, diet, lifestyle, infection, accidental trauma, and stress all may contribute to the development of a disease network. Because of these multiple contributions, the networks causing the same disease most likely differ among individual patients, thus necessitating personalized drug intervention. To cause the same disease in different

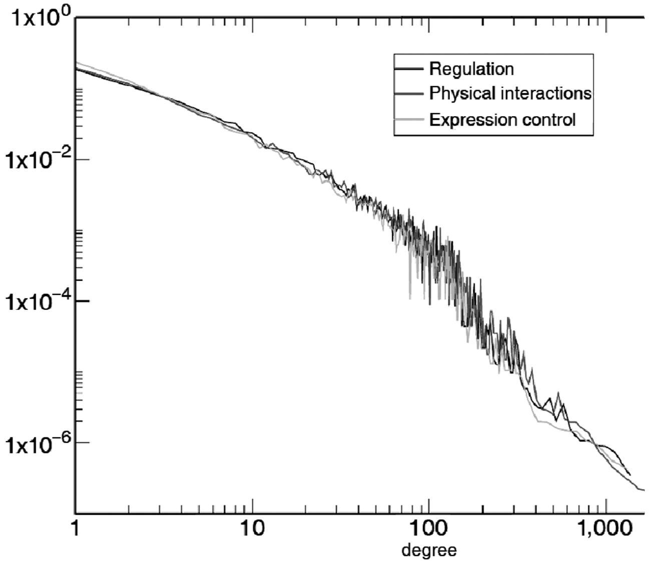


Figure 1.2 Degree distribution among known physical and regulatory interactions in ResNet database from Ariadne Genomics. The networks were built by automatically extracting relations from scientific literature using MedScan, a natural language processing technology [16].

individuals, individualized disease networks must, however, have in common the misdirected information flow that should occur through the sharing of the pathway components. Since hubs are the best targets to disrupt a scale-free network, it is yet to be seen whether individualized disease networks have similar hub composition or they overlap only in the "peripheral" nodes that are responsible for malignancy. Once formed, the disease network should be resilient to drugs precisely because of the robustness that has enabled its existence in the first place [30]. Therefore, from the network biology perspective, a drug design strategy must be a strategy to disrupt the robustness of a disease network. The network disruption itself cannot be adequate and must be supplemented by other changes that will make the disruption irreversible. Irreversibility can be achieved by either using other drugs or changes in lifestyle and diet. It is highly unlikely, however, that drugs will restore the original "normal" *network*, due to the general complexity and evolutionary nature of the development of the malignant network. Nonetheless, drugs must restore the "healthy" information flow, which is a desirable clinical outcome. Therefore, from the pathway analysis perspective, drugs must redirect a malignant information flow to restore the normal, healthy *pathway*. In many cases, this restoration can mean returning to the original pathway disrupted by a disease. In some cases, for example, in diabetes or obesity, it may mean "improving" the original pathway and making it even "healthier."

I must mention that no disease network has been completely established, and the previous paragraph is only a speculation. This book will offer examples of the current efforts toward building such networks. Here I want to mention the networks recently identified for inherited ataxia [31] and the angiogenic switch in human pancreatic cancer [32]. Most complex diseases, such as cancer or diabetes, are multistep processes of malignant transformation. Each step is probably characterized by a different robust malignant network. Therefore, it is not entirely correct to speak about a cancer disease network, for example. It is quite appropriate to speak about and study a cancer pathway, however. This pathway represents the path of cancer development that has a start, a direction, and a final stage. Every step on this pathway represents a biological process whose defect causes the progression of the disease (Figure 1.3). Each step has its own malignant network. This is another example of the difference between a pathway diagram and a network diagram: the depiction of disease history as a chain of events in time.

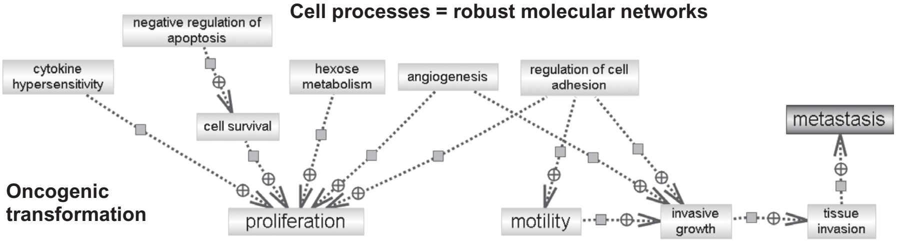


Figure 1.3 Typical cancer development pathway. This diagram depicts the sequence of cell transformation events occurring from the onset of the disease until the death of the patient. Each transformation is triggered by defects in one or several cellular processes shown at the top level of the diagram. Depending of the type of cancer, some steps may be irrelevant or skipped. For example, angiogenesis is not necessary for oncogenic transformation of blood cells. The exact order of events contributing to malignant cell proliferation is not known. In principle, this sequence can vary in different patients or cancer types, but tumor-induced angiogenesis occurs at the later stages of cancer after micro-tumors have reached a certain size. Angiogenesis in the context of cancer development means the induction of blood vessels growth by tumors. Therefore, the angiogenesis network in cancer cells mediates the production of cytokines responsible for angiogenesis in endothelial cells and the angiogenic switch network in endothelial cells mediates proliferation and differentiation of endothelial cells.

A defect in every process contributing to cancer development occurs through the establishment of the robust molecular network performing the malignant function in the cell [73]. These networks are beginning to be identified [31] for each step of cancer progression. In principle, these networks should be different in every human tissue, and each tissue may have several networks for each cancer step depending on individual genetic and environmental factors that lead to the development of cancer in the first place.

1.9 PATHWAY ANALYSIS AND DYNAMIC MODELING IN DRUG DEVELOPMENT

In essence, dynamic modeling or pathway kinetic simulation requires building a pathway diagram prior to developing a kinetic model. The mathematical model developed from experimental data is considered to be the final triumph of the effort to understand biological processes. The main criterion for successful mathematical modeling is the correct simulation of experimentally observed behavior. For this reason, good models can be built only when a system or a process is studied well enough to have known reproducible behavior in response to a stimulus. Very few examples of such processes in molecular biology exist, imposing a major limitation on the development of kinetic models. Experimentally observed cycling, oscillations, and threshold behavior appear as the most attractive targets for kinetic modeling. Hence, the first models developed in molecular biology were for the cell cycle [33], the circadian cycle [34], and the oscillation in NF-kappaB pathway [35]. Recently, the models simulating development of apoptosis [36–41], signaling in the EGFR receptor [42], and the JAK-STAT pathway [43] were developed. The criteria for a successful apoptosis model were selected based on known drug effects [44,45], threshold behavior or bistability between cell apoptotic responses and survival responses to cytotoxic stress [39], and tissue-specific differences in bistable (irreversible) and monostable (reversible) apoptotic responses [40]. For the EGFR signaling model, the criterion for success was the known activation profile of EGFR downstream targets. For the JAK-STAT pathway, the criterion for success was the cycling of STAT protein between cytoplasm and nuclei.

A mathematical theory for modeling of signaling pathways has been put forward by Heinrich et al. [46]. The theory was designed to model rate, duration, and amplitude of a signal in linear kinase-phosphatase cascades, coupled to feedback interactions and crosstalk with other signaling pathways. Undoubtedly, these modeling attempts contribute to our general understanding of biological processes dynamics. For example, they showed that phosphatases affected the rate and duration of signaling, whereas kinases controlled signal amplitude in the EGFR pathway [46]; that RAF-1 signaling was the most important regulator of EGFR phosphorylation [47]; that EGF-induced responses were stable over a wide range of ligand concentration; and that the initial velocity of receptor activation determines signaling efficiency through the EGFR pathway [42].

The behavior of a biological system, by definition, must be probabilistic in order to cope with novel environmental factors previously unmet in evolution. The only way a cell can find the optimal response to a novel stimulus is by presenting all possible responses while looking for the first optimal one that may become selected. For example, that cell transcriptional response to previously unknown challenges is fundamentally random was recently shown for yeast [9]. Drug treatment, by all means, represents the unmet challenge for a cell and therefore the cell response to a drug must be a stochastic process, which varies among the cell population of one or several human tissues. Transcriptional and epigenetic reprogramming of an entire system in response to a drug may provide strong limitations on using dynamic models in drug development. Drug-induced cell reprogramming can be caused by both the inhibition of an intended drug target and the side effects of a drug. If a drug causes global transcriptional reprogramming in a cell, its efficacy cannot be predicted by solely using the kinetic models. Instead, the new cell state first must be determined through using experimental measurements in every individual patient. The reprogrammed cell will have changed relative concentrations of proteins involved in the process that is to be modeled. Therefore, mechanistic dynamic modeling can be useful in evaluating how close the new cell state and the desired clinical outcome are by calculating the dynamics of affected processes in a new state. In the case when a drug does not cause the global reprogramming, or reprogramming is mild, it should be possible to predict the cell response by a dynamic model. Due to the variability of the initial conditions among cells in the body, the biologically meaningful outcome of any modeling should be a solution space of all possible cell responses to a drug treatment with a specific probability score assigned to every response curve in the solution space. Even though computational approaches to address this problem are being developed [48–50], they currently suffer from the general lack of knowledge about intracellular events, which is necessary for creating the model. These approaches also have knowledge gaps about cell behavior that is necessary for model validation.

A global cell model is the ultimate goal of pathway analysis. Its general complexity with millions of freely adjustable parameters probably will require more experimental constraints than modern molecular biology can ever provide. Thus, the uncertainty due to the lack of knowledge about the system must be added to the fundamental variability of the cell response, making the predictions by available models even less reliable. Nevertheless, current efforts will ultimately yield a computational model for the entire cell. The useful application of the global cell models, however, will be made in the even more distant future than the creation of complete database of molecular interactions for the human cell. Therefore, the best success stories of dynamic modeling are likely to happen beyond the life span of most readers of this book.

One goal of current dynamic modeling efforts is to isolate the minimal set of components and relations that can be used to correctly predict the behavior of a system and then to predict a system response, such as drug action, to the perturbations. While creating a model that uses pathway analysis methods, one can identify essential interactions from the pool of all cellular interactions mediating the modeled process. Once principal interactions are identified from the pool of all known interactions, it may become evident that the pathway is incomplete and actually misses some interactions. Pathway analysis can point an investigator to missing pathway components and help in the design of an appropriate experiment for identifying missing components. An assumption that the correct evaluation and prediction of drug efficacy requires the kinetic simulation of pathways containing drug targets automatically necessitates simulating hundreds of thousands of pathways for the same reasons described in previous sections. Such a large number of models will require automatic assembly using pathway analysis tools. As you will see in upcoming chapters of this book, pathway analysis may provide some alternatives to brute force kinetic modeling in evaluating drug efficacy and toxicity, as well as in selecting drug targets.

1.10 STEADY-STATE ANALYSIS OF METABOLIC NETWORKS

Flux balance analysis (FBA) [51] and extreme pathway analysis [52] are two main methods of global analysis of metabolic networks developed by Bernard Palsson and his colleagues in the last 10 years. They take advantage of a basic physical principle of mass balance with the reasonable assumption that intracellular metabolic reactions are in a steady state and therefore have constant fluxes. Certainly, metabolism changes multiple times throughout the day in the human organism [53]. These changes are rapid and the periods between changes can be viewed as a steady state. Steady-state analysis predicts a set of allowed metabolic phenotypes or phenotypic planes within the space of all possible fluxes while avoiding any kinetic modeling. The allowed flux space is calculated using standard linear algebraic methods and has the form of a convex polyhedral cone. The cone edges are called "extreme pathways" [54] and the space between them is a phenotypic plane [55]. The initial FBA solution space turned out to be very large, and the problem of calculating extreme pathways was found to be NP-hard. For genome-scale networks, this solution space has proven to be infeasible [56]. For those reasons, additional constraints on thermodynamics [57], expression regulation [58], compartmentalization [59], maximum capacity, and reaction irreversibility [60] had to be taken into account to narrow down the solution space.

In my view, metabolic reconstruction represents the most advanced method available for pathway analysis. It uses the power of the complete genome sequence, modern mathematical and computational approaches to model metabolism using basic physical principles. Yet, it possesses several limitations for drug discovery that do not allow it to become the main focus of this book. First, metabolic control in multicellular organisms is under tight hormonal control. Second, the scale of changes detectable by this method may be too detailed and irrelevant for drug discovery. It is more important to predict global metabolism homeostasis of an entire organism than the metabolism of the local cell population. The difficulties of finding a physiological interpretation of the results from extreme pathway analysis were acknowledged by Bernard Palsson himself [61].

FBA has proven to be a very good method to model the metabolic state and to predict growth conditions and metabolite yields of the microorganisms