Part 1 Introduction

1
Bioinformatics – From Genomes to Therapies

Thomas Lengauer

1 Introduction

In order to set the stage for this book, this chapter provides an introduction to the molecular basis of disease. We then continue to discuss modern biological techniques with which we have recently been empowered to screen for molecular drugs targets as well as for the drugs themselves. The chapter finishes with an overview of the organization of the book.

2 The Molecular Basis of Disease

Diagnosing and curing diseases has always been and will continue to be an art. The reason is that man is a complex being with numerous facets, many of which we do not and probably will never understand. Diagnosing and curing diseases has many aspects, include biochemical, physiological, psychological, sociological and spiritual aspects.

Molecular medicine reduces this variety to the molecular aspect. Living organisms, in general, and humans, in particular, are regarded as complex networks of molecular interactions that fuel the processes of life. This "molecular circuitry" has intended modes of operation that correspond to healthy states of the organism and aberrant modes of operation that correspond to diseased states. The main goal of molecular medicine is the identification of the molecular basis of a disease, i.e. to answer the question: "What goes wrong in the molecular circuitry?". The goal of therapy is to guide the biochemical circuitry back to a healthy state. The molecular approach has already proven

Bioinformatics - From Genomes to Therapies Vol. 1. Edited by Thomas Lengauer

ISBN: 978-3-527-31278-8

its effectiveness for understanding diseases, and is dramatically enhanced by genomics and proteomics technology [5]. It is the prime purpose of this book to explore the contributions that this technology, particularly its computational aspect, can have to advancing molecular medicine.

As already noted, the molecular basis of life is composed of complex biochemical processes that constantly produce and recycle molecules, and do so in a highly coordinated and balanced fashion. The underlying basic principles are quite alike throughout all kingdoms of life, even though the processes are much more complex in highly developed animals and the human than in bacteria, for instance. Figure 1 gives an abstract view of such an underlying biochemical network, the *metabolic network* of a bacterial cell (the intestinal bacterium *Escherichia coli*) – it affords an incomplete and highly simplified account of the cell's metabolism, but it nicely visualizes the view of a living cell as a biochemical circuit. The figure has the mathematical structure of a *graph*. Each dot (*node*) stands of a small organic molecule that is metabolized within the cell. Alcohol, glucose and ATP are examples for such molecules. Each line (*edge*) indicates a chemical reaction. The two nodes connected by the edge represent the substrate and the product of the reaction. The colors represent the role that the respective reaction plays in metabolism. These roles include the construction of molecular components that are essential for life – nucleotides (red), amino acids (orange), carbohydrates (blue), lipids (light blue), etc. – or the breakdown of molecules that are not helpful or even harmful to the cell. Other tasks of chemical reactions in a metabolic network pertain to the storage and conversion of energy. (The blue cycle in the center of Figure 1 is the citric acid cycle.) A third class of reactions facilitates the exchange of information in the cell or between cells. This includes the control of when and in what way genes are expressed (*gene regulation*), as well as such tasks as the opening and closing of molecular channels on the cell surface, and the activation or deactivation of cell processes such as replication or apoptosis (programmed cell death). The reactions that regulate cellular processes are often collectively called the *regulatory network*. Recently, molecular networks that facilitate the propagation of signals within the cell are being selectively called *signal transduction networks*. Figure 1 only includes metabolic reactions, without any regulatory reactions or signal transduction cascades. Of course, all molecular networks of a cell are closely intertwined and many reactions can have metabolic as well as regulatory aspects. In general, much more is known about metabolic than regulatory networks, even though many relevant diseases involve regulatory rather than metabolic dysfunction.

The metabolic and regulatory networks can be considered as composed of partial networks that we call *pathways*. Pathways can fold in on themselves, in which case we call them *cycles*. A metabolic pathway is a group of reactions that turns a substrate into a product over several steps (pathway) or recycles

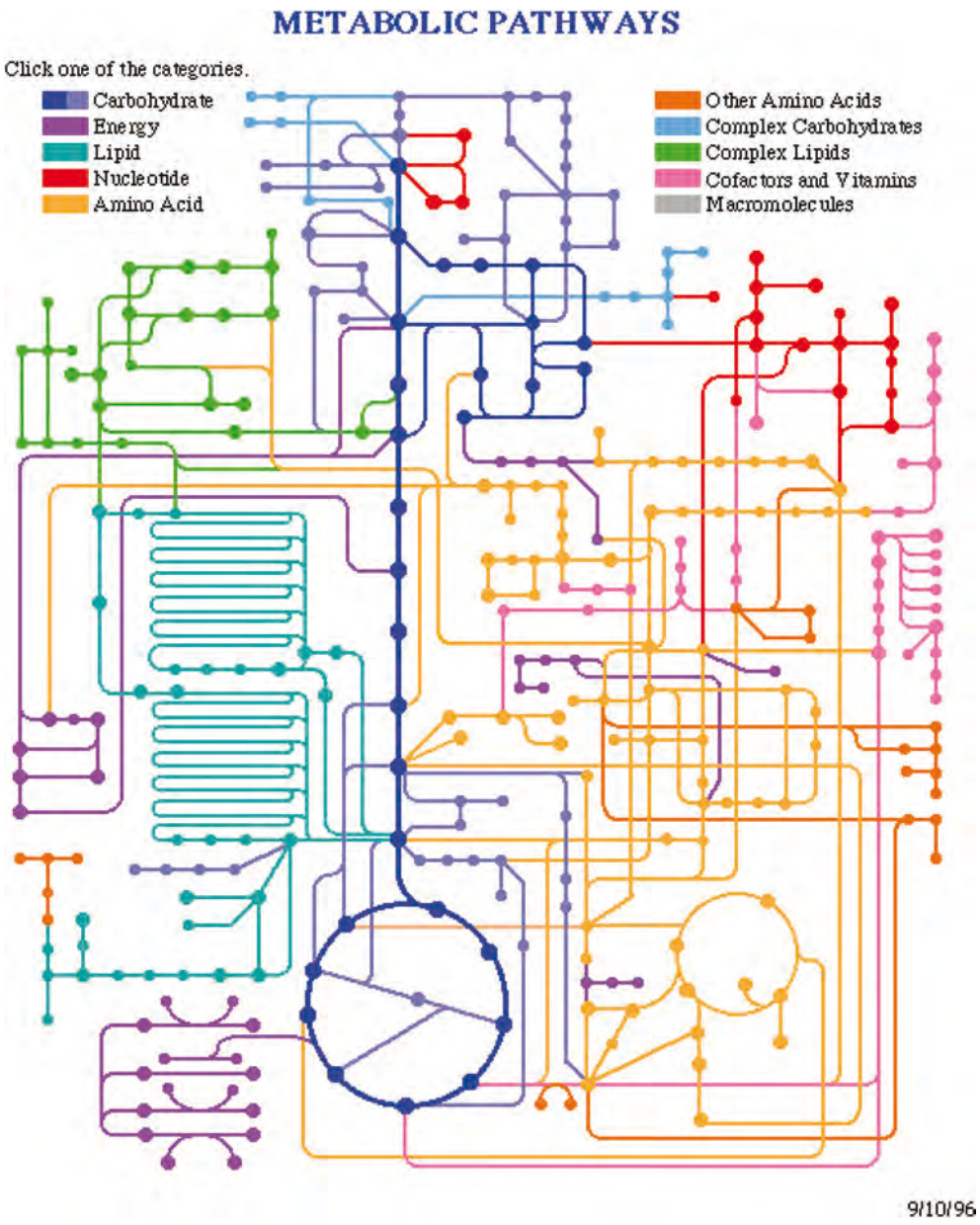


Figure 1 Abstract view of part of the metabolic network of the bacterium *E. coli* (from http://www.genome.ad.jp/kegg/kegg.html).

a molecule by reproducing it in several steps (cycle). *The glycolysis pathway* (the sequence of blue vertical lines in the center of Figure 1) is an example of a pathway that decomposes glucose into pyruvate. The *citric acid cycle* (the

blue cycle directly below the glycolysis pathway in Figure 1) is an example of a cycle that produces ATP – the universal molecule for energy transport. Metabolic cycles are essential in order that the processes of life do not accumulate waste or deplete resources. (Nature is much better at recycling than man.)

There are several ways in which Figure 1 hides important details of the actual metabolic pathway. In order to discuss this issue, we have extracted a metabolic cycle from Figure 1 (see Figure 2). This cycle contributes to cell replication; more precisely, ;t is one of the motors that drive the synthesis of thymine – a molecular component of DNA. In Figure 2, the nodes of the metabolic cycle are labeled with the respective organic molecules and the edges point in the direction from the substrate of the reaction to the product. Metabolic reactions can take place spontaneously under physiological conditions (in aqueous solution, under room temperature and neutral pH). However, nature has equipped each reaction (each line in Figure 1) with a specific molecule that catalyzes that reaction. This molecule is called an *enzyme* and, most often, it is a protein. An enzyme has a tailor-made binding site for the transition state of the catalyzed chemical reaction. Thus, the enzyme speeds up the rate of that reaction tremendously, by rates of as much as 10^7. Furthermore, the rate of a reaction that is catalyzed by an enzyme can be regulated by controlling the effectiveness of the enzyme or the number of enzyme molecules that are available.

How does the enzyme do its formidable task? As an example, consider the reaction in Figure 2 that turns dihydrofolate (DHF) into tetrahydrofolate (THF). This reaction is catalyzed by an enzyme called *dihydrofolate reductase (DHFR).* The surface of this protein is depicted in Figure 3. One immediately recognizes a large and deep pocket that is colored blue (representing its negative charge). This pocket is a *binding pocket* or *binding site* of the enzyme, and it is ideally adapted in terms of geometry and chemistry so as to bind to the substrate molecule DHF and present it in a conformation that is conducive for the desired chemical reaction to take place. In this case, this pocket is also

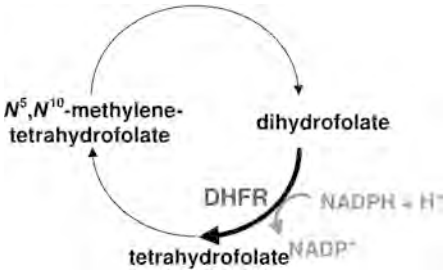


Figure 2 A specific metabolic cycle.

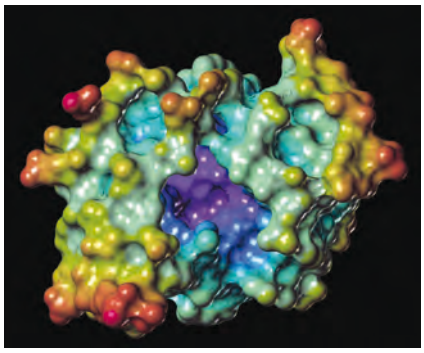


Figure 3 The 3-D structure of DHFR colored by its surface potential. Positive values are depicted in red, negative values in blue.

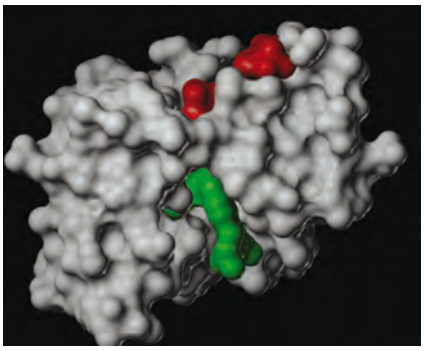


Figure 4 DHFR (gray) complexed with DHF (green) and NADPH (red).

where the reaction is catalyzed. We call this place the *active site*. (There can be other binding pockets in a protein that are far removed from the active site.)

There is another aspect of metabolic reactions that is not depicted in Figure 1 – many reactions involve *cofactors*. A cofactor is an organic molecule or a metal ion that has to be present in order for the reaction to take place. If the cofactor is itself modified during the reaction, we call it a *cosubstrate.* In the case of our example reaction, we need the cosubstrate NADPH for the reaction to happen. The reaction modifies DHF to THF and NADPH to $NADP^+$. Figure 4 shows the molecular complex of DHFR, DHF and NADPH before the reaction happens. After the reaction has been completed, both organic molecules dissociate from DHFR and the original state of the enzyme is recovered.

Now that we have discussed some of the details of metabolic reactions, let us move back to the global view of Figure 1. We have seen that each of the edges in Figure 1 represents a reaction that is catalyzed by a specific protein. (However, the same protein can catalyze several reactions.) In *E. coli* there are an estimated 1500 enzymes [6]; in human there are thought to be about least twice as many. The molecular basis of a disease lies in modifications of the action of these biochemical pathways. Some reactions do not happen at their intended rate (e.g. in gout), resources that are needed are not present in sufficient amounts (vitamin deficiencies) or waste products accumulate in the body (Alzheimer's disease). In general, imbalances induced in one part of the network spread to other parts. The aim of therapy is to replace the aberrant processes with those that restore a healthy state. The most desirable fashion in which this could be done would be to control the effectiveness of a whole set of enzymes in order to regain the metabolic balance. This set probably involves many, many proteins, as we can expect many proteins to

be involved in manifesting the disease. Also, each of these proteins would have to be regulated in quite a specific manner. The effectiveness of some proteins would possibly have to be increased dramatically, whereas other proteins would have to be blocked entirely, etc. It is obvious that this kind of therapy involves a kind of global knowledge of the workings of the cell and a refined pharmaceutical technology that is far beyond what man can do today and for some time to come.

3 The Molecular Approach to Curing Diseases

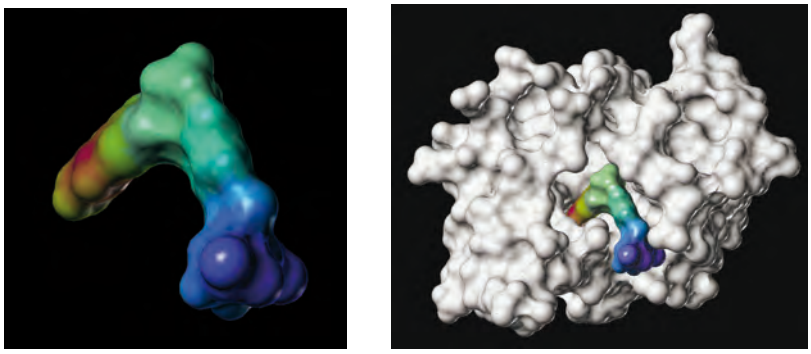
For this reason, the approach of today's pharmaceutical research is far more simplistic. The aim is to regulate a single protein. In some cases we aim at completely blocking an enzyme. To this end, we can provide a drug molecule that effectively competes with the natural substrate of the enzyme. The drug molecule, the so-called *inhibitor*, has to be made up such that it binds more strongly to the protein than the substrate. Then, the binding pockets of most enzyme molecules will contain drug molecules and cannot catalyze the desired reaction in the substrate. In some cases, the drug molecule even binds very tightly (covalently) to the enzyme (suicide inhibitor). This bond persists for the remaining lifetime of the protein molecule. Eventually, the deactivated protein molecule is broken down by the cell and a new identical enzyme molecule takes its place. Aspirin is an example of a suicide inhibitor. The effect of the drug persists until the drug molecules themselves are removed from the cell by its metabolic processes and no new drug molecules are administered to replace them. Thus, one can control the effect of the drug by the time and dose it is administered.

There are several potent inhibitors of DHFR. One of them is *methotrexate (MTX)*. Figure 5 shows MTX (color) both unbound (left) and bound (right) to DHFR (gray). MTX has been administered as an effective cytostatic cancer drug for over two decades.

There are many other ways of influencing the activity of a protein by providing a drug that binds to it. Drugs interact with all kinds of proteins:

- With receptor molecules that are located in the cell membrane and fulfill regulatory or signal transduction tasks.
- With ion channels and transporter systems (again protein residing in the cell membrane) that monitor the flux of molecules into and out of the cell.

The mode of interaction between drug and protein does not always have the effect of blocking the protein. In some cases, the drug mimics a missing small molecule that is supposed to activate a protein. We call such drugs agonists.



(a) (b)

Figure 5 MTX (colored by its surface potential, see Figure 3): (a) unbound, (b) bound to DHFR (gray).

In general, we are looking for drugs that bind tightly to their protein target (effectiveness) and to no other proteins (selectivity).

Most drugs that are on the market today modify the enzymatic or regulatory action of a protein by strongly binding to it as described above. Among these drugs are long-standing, widespread and highly popular medications, and more modern drugs against diseases such as AIDS, depression or cancer. Even the lifestyle drugs that have come into use in recent years, e.g. Viagra and Xenical, belong to the class of protein inhibitors.

In this view, the quest for a molecular therapy of a disease decomposes into three parts:

- *Question 1: Which protein should we target?* As we have seen, there are many thousands of candidate proteins in the human. We are looking for one that, by binding the drug molecule, provides the most effective remedy of the disease. This protein is called the *target protein*.

- *Question 2: Which drug molecules should be used to bind to the target protein?* Here, the molecular variety is even larger. Large pharmaceutical companies have compound archives with millions of compounds at their disposal. Every new target protein raises the question of which of all of these compounds would be the best drug candidate. Nature uses billions of molecules. With the new technology of combinatorial chemistry, where compounds can be synthesized systematically from a limited set of b uilding blocks, this number of *potential* drug candidates is also becoming accessible to the laboratory.

- *Question 3: Given a choice of different drugs to administer to a patient, in order to alleviate or cure a specific disease, what is the best selection of drugs to give to that individual patient*? Questions 1 and 2 have been posed without the specifics of an individual patient in mind. Target protein and drug were selected

for all putative patients collectively. With Question 3 we are entering the more advanced stage of *personalized medicine* – we want to understand the different ways in which different patients react to the same drug.

Question 3 has only come into the focus of research recently. The inclusion of the discussion of this question presents a major new feature of this book over its predecessor.

We will now give a short summary of the history of research on all three questions.

4 Finding Protein Targets

Let us start our discussion of the search for target proteins by continuing our molecular example of DHFR/MTX. As mentioned, DHFR catalyzes a reaction that is required for the production of thymine – a component of DNA. Thus, blocking DHFR impairs DNA synthesis and, therefore, cell division. This is the reason that MTX, an inhibitor of DHFR, is administered as a cytostatic drug against cancer. Is DHFR the "right" target protein in this context? The frank answer to this question must be "no". DHFR is active in every dividing cell, tumor cells as well as healthy cells. Therefore, MTX impairs the division of all dividing cells that it can get to. This is the cause of the serious side-effects of the drug such as loss of hair and intestinal lining. We see that in this case the limits of the therapy are mostly dictated by the choice of the wrong target protein. Why then is this protein chosen as a target? The answer to this question is also very simple: we cannot find a better one. This example shows how central the search for suitable target proteins is for developing effective drug therapies.

Target proteins could not really realistically be searched for until a few years ago. Historically, few target proteins were known at the time that the respective drug had been discovered. The reason is that new drugs were developed by modifying natural metabolites or known drugs, based on some intuitive notion of molecular similarity. Each modification was immediately tested in the laboratory either *in vitro* or *in vivo*. Thus, the effectiveness of the drug could be assessed without even considering the target protein. To this day, all drugs that are on the marketplace worldwide target an estimated set of not much more than 500 proteins [3]. Thus, the search for target proteins is definitely the dominant bottleneck of current pharmaceutical research.

Today, new experimental methods of molecular biology, the first versions of which were developed just a few years ago, provide us with a fundamentally new way – the first systematic way – of looking for protein targets. The basis for all of these methods was the technological progress made in the context of the quest for sequencing the human genome [1]. Based on this

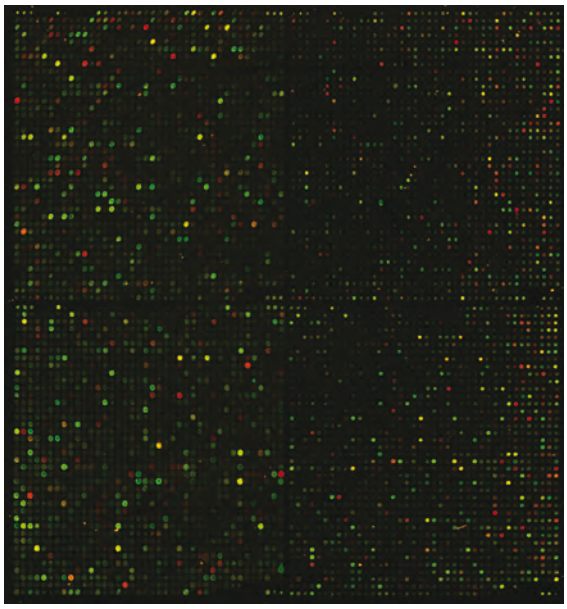


Figure 6 A DNA chip (from http://cmgm.stanford.edu/pbrown/explore/).

technology, additional developments have been undertaken to be able to measure the amount of expressed genes and proteins in cells. We exemplify this progress using a specific DNA chip technology [2]; however, the general picture extends to many other experimental methods under development.

Figure 6 shows a DNA chip that provides us with a differential census of the gene expressed by a yeast cell in two different cell states – one governed by the presence of glucose (green) and one by the absence of glucose (red). In effect the red picture is that of a starving yeast cell, whereas the green picture shows the "healthy" state. Each bright green dot indicates a protein that is manufactured (expressed) in high numbers in the "healthy" state. Each bright red dot indicates a protein that is expressed in high numbers in the starving cell. If the protein occurs frequently in both the healthy and the starving state, the corresponding dot is bright yellow (resulting from an additive mixture of the colors green and red). Dark dots indicate proteins that are not frequent, the tint of the color again signifies whether the protein occurs more often in the healthy cell (green), equally often in both cells (yellow) or more often in the diseased cell (red).

At this point, the exactly nature of the experimental procedures that generate the picture in Figure 6 is of secondary importance. What is important is how much information is attached to the colored dots in the picture. Here, we can make the following general statements.

(i) The identity of the protein is determined by the coordinates of the colored dot. We will assume, for simplicity, that dots at different locations also represent different proteins. (In reality, multiple dots that represent the same protein are introduced, on purpose, for the sake of calibration.) The exact arrangement of the dots is determined before the chip is manufactured. This involves identifying a number of proteins to be represented on the chip and laying them out on the chip surface. This layout is governed by boundary conditions and preferences of the experimental procedures, and is not important for the interpretation of the information

(ii) Only rudimentary information is attached to each dot. At best, the experiment reveals the complete sequence of the gene or protein. Sometimes, only short segments of the relevant sequence are available.

Given this general picture, the new technologies of molecular biology can be classified according to two criteria, as shown in the following two subsections.

4.1 Genomics versus Proteomics

In genomics, it is not the proteins themselves that are monitored, but rather we screen the expressed genes whose translation ultimately yields the respective proteins. In proteomics, the synthesized proteins themselves are monitored. The chip in Figure 6 is a DNA chip, i.e. it contains information on the expressed genes and, thus, only indirectly on the final protein products. The advantage of the genomics approach is that genes are more accessible experimentally and easier to handle than proteins. For this reason, genomics is ahead of proteomics, today. However, there also are disadvantages to genomics. First, the expression level of a gene need not be closely correlated with the concentration of the respective protein in the cell, although the latter figure may be more important to us if we want to elicit a causal connection between protein expression and disease processes. Even more important, proteins are modified post-translationally (i.e. after they are manufactured). These modifications involve glycosylation (attaching complex sugar molecules to the protein surface) and phosphorylation (attaching phosphates to the protein surface), for instance, and they lead to many versions of protein molecules with the same amino acid sequence. Genomics cannot monitor these modifications, which are essential for many diseases. Therefore, it can be expected that, as the experimental technology matures, proteomics will gain importance over genomics (see also Chapter 45).

4.2 Extent of Information Available on the Genes/Proteins

Technologies vary widely in this respect. The chip in Figure 6 is generated by a technology that identifies (parts of) the gene sequence. We are missing information on the structure and the function of the protein, its molecular interaction partners, and its location inside the metabolic or regulatory network of the organism. All of this information is missing for the majority of the genes on the chip.

There are many variations on the DNA chip theme. There are technologies based on so-called *expressed sequence tags (ESTs)* that tend to provide more inaccurate information on expression levels and various sorts of microarray techniques (see Chapter 24). All technologies have in common that the data they produce require careful quality control (Chapter 25). In general, it is simpler to distinguish different disease states from gene expression data (Chapter 26) than to learn about the function of the involved proteins from these data (Chapter 27). Proteomics uses different kinds of separation techniques, e.g. chromatography or electrophoresis combined with mass spectrometry, to analyze the separated molecular fractions (see Chapter 28). As is the case with genomics, proteomics technologies tend to generate information on the sequences of the involved proteins and on their molecular weight, and possibly information on post-translational modifications such as glycosylation and phosphorylation. Again, all higher-order information on protein function is missing. It is not feasible to generate this information exclusively in the wet laboratory – we need bioinformatics to make educated guesses here. Furthermore, basically all facets of bioinformatics that start with an assembled sequence can be of help. This includes the comparative analysis of genes and proteins (Chapter 37), protein structure prediction (Chapters 9–13), protein function prediction (Chapters 30–34), analysis and prediction of molecular interactions involving proteins (Chapters 16 and 17) as well as bioinformatics for analyzing metabolic and regulatory networks (Chapters 20–22). This is why all of bioinformatics is relevant for the purpose of this book.

If, with the help of bioinformatics, we can retrieve enough information on the molecular networks that are relevant for a disease, then we have a chance of composing a detailed picture of the disease process that can guide us to the identification of possible target proteins for the development of an effective drug. Note that the experimental technology described above is universally applicable. The chip in Figure 6 contains all genes of the (fully sequenced) organism *Saccharomyces cerevisiae* (yeast). The cell transition analyzed here is the diauxic shift – the change of metabolism upon removal of glucose. However, we could exchange this with almost any other cell condition of any tissue of any conceivable organism. The number of spots that can be put on a single chip goes into the hundreds of thousands. This is enough to put all of

the human genes on a single chip. Also, we do not have to restrict ourselves to disease conditions; all kinds of environmental conditions (temperature, pH, chemical stress, drug treatment, diverse stimuli, etc.) or intrinsic conditions (presence or absence of certain genes, mutations, etc.) can be the subject of study.

The paradigm of searching for target proteins in genomics data has met with intense excitement from the pharmaceutical industry, which has invested heavily in this field over recent years. However, the first experiences have been sobering. It seems that we are further away from harvesting novel target proteins from genomics and proteomics data than we initially thought. However, in principle, a suitable novel target protein can afford a completely new approach to disease therapy and a potentially highly lucrative worldwide market share. For a critical review of the target-based drug development process, see Sams-Dodd [7].

Providing adequate bioinformatics support for finding new target proteins is a formidable challenge that is the focus of much of this book. However, once we have a target protein, our job is not done.

5 Developing Drugs

If the target protein has been selected, we are looking for a molecule that binds tightly to the relevant binding site of the protein. Nature often uses macromolecules, such as proteins or peptides, to inhibit other proteins. However, proteins do not make good drugs – they are easily broken down by the digestive system, they can elicit immune reactions and they cannot be stored for a long period of time. Thus, after an initial excursion into drug design based on proteins, pharmaceutical research has basically gone back to looking for small drug molecules. Here, one idea is to use a peptide as the template for an appropriate drug (peptidomimetics).

Due to the lack of fundamental knowledge of the biological processes involved, the search for drugs was, until recently, governed by chance. However, as long as chemists have thought in terms of chemical formulae, pharmaceutical research has attempted to optimize drug molecules based on chemical intuition and on the concept of molecular similarity. The basis for this approach is the lock-and-key principle formulated by Emil Fischer [4] over 100 years ago. Figures 3 to 5 illustrate that principle: in order to bind tightly, the two binding molecules have to be complementary to each other both sterically and chemically (colors in Figures 3 and 5). The drug molecule fits into the binding pocket of the protein like a key inside a lock. The lock-and-key principle has been the dominating paradigm in drug research ever since its proposal. It has been refined to include the phenomenon of induced fit, by

which the binding pocket of the protein undergoes subtle steric changes in order to adapt to the geometry of the drug molecule.

For most of the past century the structure of protein-binding pockets has not been available to the medicinal chemist. Even to this day the structure of the target protein will not be known for many pharmaceutical projects for some time to come. For instance, many diseases involve target proteins that reside in the cell membrane and we cannot expect the three-dimensional (3-D) structure of such proteins to become known soon. If we have no information on the structure of the protein-binding site, drug design is based on the idea that molecules that are similar in terms of composition, shape and chemical features should bind to the target protein with comparable strength. The respective drug-screening procedures are based on comparing drug molecules, either intuitively or, more recently, systematically with the computer. The resulting search algorithms are very efficient and allow searching through compound libraries with millions of entries (Chapter 18).

As 3-D protein structures became available, the so-called *rational* or *structure-based* approach to drug development was invented, which exploited this information to develop effective drugs. Rational drug design is a highly interactive process with the computer originally mostly visualizing the protein structure and allowing queries on its chemical features. The medicinal chemist interactively modified drug molecules inside the binding pocket of the protein at the computer screen. As rational drug design began to involve more systematic optimization procedures interest arose in *molecular docking*, i.e. the prediction of the structure and binding affinity of the molecular complex involving a structurally resolved protein and its binding partner (Chapter 16). Synthesizing and testing a drug in the laboratory used to be comparatively expensive. Thus, it was of interest to have the computer suggest a small set of highly promising drug candidates. After an initial lead molecule has been found that binds tightly to the target protein, secondary drug properties have to be optimized that maximize the effectiveness of the drug and minimize side-effects (Chapter 19).

With the advent of *high-throughput screening* the binding affinity of as many as several hundred thousands drug candidates to the target protein can now be assayed within a day. Furthermore, *combinatorial chemistry* allows for the systematic synthesis of molecules that are composed of preselected molecular groups that are linked with preselected chemical reactions. The number of molecules that is accessible in such a combinatorial library can, in principle, exceed many billions. Thus, we need the computer to suggest promising sublibraries that promise to contain a large number of compounds that bind tightly to the protein (Chapters 16 and 18).

As in the case of target finding, the new experimental technologies in drug design require new computer methods for screening and interpreting the

voluminous data assembled by the experiment. These methods are seldom considered part of bioinformatics, since the biological object, i.e. the target protein, is not the focus of the investigation. Rather, people speak of *cheminformatics* – the computer aspect of medicinal chemistry. Whatever the name, it is our conviction that both aspects of the process that guides us from the genome to the drug have to be considered together and we will do so in this book.

6 Optimizing Therapies

How is it that different patients react differently to the same drug? Reasons for this phenomenon can be manifold. Some are easier to investigate with methods of modern biology and bioinformatics than others. Here, we distinguish between infectious diseases and other diseases.

The molecular basis of any infectious disease is the interplay of a usually large population of a pathogen with the human host. The pathogen takes advantage of the human host or, in the case of virus, even hijacks the infected cells of the patient. Chapter 23 relates a story about the interplay of a viral pathogen with the infected host cell.

With infectious diseases, the drug often targets proteins of the infecting pathogen rather than the human host. The reason is the hope that drugs for such targets harbor less serious side-effects for the patient. However, in all infectious diseases, there is a constant battle going on between the host, whose immune system tries to eradicate the pathogen, and the pathogen that tries to evade the immune system. If the disease is treated with drugs, the administered drugs impose an additional selective pressure on the pathogen. On the road to resistance the pathogen constantly changes its genome and, thus, also the shape of the target proteins for drug therapy. Changes that are beneficial for the pathogen are those that render the drugs less effective, i.e. the pathogen becomes resistant. The results of this process are widely known. With bacteria, we observe increasingly resistant strains against antibiotic therapies (Chapter 41). With viral diseases such as AIDS the drug therapy has to be adapted continually to newly developing resistant strains within the patient (Chapter 40). Therapy selection must be individualized, in both cases, at least by taking the present strain of the pathogen into account and, at best, by also considering the individual characteristics of the host. Since the pathogen is a much simpler organism than the human host, the former is significantly easier than the latter.

Although the drug acts on its intended protein target, the drug has to find its way to the site of action and, eventually, has to be metabolized or excreted again. Along that path there are multiple ways in which the drug can

interact with the patient. The resulting side-effects depend on the molecular and genetic status of the individual patient. Furthermore, the protein target often has different functions, such that its inhibition or agonistic activation can incur side-effects on molecular processes that were not intended to be changed. Again, the form and magnitude of such side-effects depends on the individual patient. This process of bringing about different reactions to drugs in different patient is much harder to analyze. The reason is that larger, often widespread, networks of interactions in the patient have to be taken in account. Analyzing them necessitates complex and accurately assembled patient histories and diverse molecular data that are seldom collected in today's clinical practice. Therefore, this approach to personalized medicine is still in its infancy (Chapter 39).

Another issue with diseases is the genetic predisposition of the human individual to the disease. Monogenetic diseases have been known for a long time and are relatively easy to analyze. Here, a defect in a single gene gives rise to the disease. However, these diseases are rare, in general. The major diseases like cancer, diabetes, and inflammatory and neurodegenerative diseases are based on a complex interplay between environmental and genetic factors with probably many genes involved. With data on the genomic differences between individuals just coming into being, the analysis of the genetic basis for complex diseases is embarking on a route that hopefully will lead to more effective means of prognosis, diagnosis and therapy.

7 Organization of the Book

This book is composed of three volumes. It is organized along the line from the genotype to the phenotype.

Volume 1: *The building blocks: sequences and structures*. This volume discusses the analysis of the basic building blocks of life, such as genes and proteins.

Volume 2: *Getting at the inner workings: molecular interactions*. This volume concentrates on the "switches" of the biochemical circuitry, the molecular interactions, as well as the circuits composed by these switches, the biochemical networks. In the former context, it partly also ventures into applied issues of drug design and optimization.

Volume 3: *The Holy Grail: molecular function*. This volume ties the elements provided by the first two volumes together and attempts to draw an integrated picture of molecular function – as far as we can do it today. The volume also discusses ramifications of this picture for the development and administration of drug therapies.

Each volume is subdivided in parts that are summarized below. The total book has 11 parts. *Volume 1* covers Parts 1–4.

Part 1 consists only of this chapter, and gives an introduction to the field and an overview of the book.

Part 2, consisting of Chapter 2, discusses bioinformatics support for assembling genome sequences. This is basic technology which is required to arrive at the genome sequence data that are the basis for much of what follows in the book. Major advances have been made in this area, especially during the finishing stages of completing the human genome sequence. The field has not lost its importance as we are embarking on sequencing many complex genomes, including over a dozen mammalian genomes. Furthermore, the technology is employed in projects that sequence closely related species, such as over a dozen species of *Drosophila*, in order to obtain a more effective database for functional genomics[1]. The authors of the chapter were part of the team that developed the assembler for the draft of the human genome sequence generated by Celera Genomics.

Part 3 is on molecular sequence analysis and comprises Chapters 3–8. Chapter 3 introduces the basic statistical and algorithmic technology for aligning molecular sequences. This technology forms the basis of much that is to follow. The author of the chapter has made seminal contributions to the field. Chapter 4 discusses methods for inferring ancestral histories from sequence data. This is one of the mainstays of comparative genomics. Similar to people, one can learn a lot about genes and proteins from looking at their ancestors and relatives, arguably more so with today's methods than from inspecting the gene or protein by itself. This attributes particular significance to this chapter in the context of this book. The authors of the chapter have made important contributions to the development of methods for inferring phylogenies and applied them to analyzing the evolution of *Homo sapiens*. Chapter 5 discusses the first major step from the genotype to the phenotype, i.e. the identification of protein-coding genes. The author of the chapter has developed one of the leading gene-finding programs. The ongoing debate on exactly what is the number of genes in the human chromosome years after the first draft of the human genome sequence was available shows that the issue of this chapter is still quite up-to-date. Furthermore, genes are a primary unit of linkage between the human genome and disease, as Chapter 38 discusses. Going into the gene's structure, most of the linkage with disease happens not in the coding regions of the genes that affect the structure of the coded protein. In general, proteins are far too well refined to be tampered with. Mostly, changing a protein means death to the individual

1) see http://preview.flybase.net/docs/news/announcements/drosboard/GenomesWP2003.html for the respective community white paper

and only a few severe diseases (such as sickle cell anemia, Huntington's chorea or cystic fibrosis) are linked to changes in the coding regions of genes. More subtle influences of the genotype on disease involve polymorphisms in the noncoding regulatory regions of the disease gene that do not affect the structure of the protein, but the mechanism and level of its expression. This lends special importance to Chapter 6, which addresses bioinformatics methods for analyzing these regions. The author of the chapter has led the development of a widely used set of software tools for analyzing regulatory regions in genomes. The analysis of regulatory regions ventures into the more difficult to analyze noncoding regions of genes. However, the really dark turf of the human genome is presented by the long and mysterious repetitive sections. Up to 40% of the human genome is covered with these regions whose relevance (or irrelevance?) is under hot debate, especially since some of these regions seem to harbor potential silenced retroviral genes that may become active again at some suitable or unsuitable time. The identification of these regions (although not the elucidation of their function) is discussed in Chapter 7. The authors of this chapter have made seminal contributions and provided widely used software for computational gene finding, genome alignment and repeat finding. Chapter 8, finally, discussed the algorithmic and statistical basis of analyzing major genome reorganizations that happened as the kingdoms of life evolved, and that include splitting, fusing, mixing and reshuffling at a chromosomal level. Again, we are just beginning to understand the evolutionary role of these transactions. The author of this chapter has provided important contributions to the methodical and biological side of the field, many of them together with David Sankoff and Pavel Pevzner.

Part 4 of the book is on molecular structure prediction and comprises Chapters 9–15. The part starts with a chapter on a half-way approach to protein structure prediction which only aims at identifying the regions of secondary structure of the protein (α-helices and β-strands) and related variants. The resulting information on protein structure is very important in its own right and, in addition, helps guide or verify tertiary structure prediction. The authors of the chapter have made seminal contributions to protein structure prediction starting in the early 1990s that increased the prediction accuracy significantly (from around 65 to well over 70%).

The most promising approach to identifying the fold of a protein, today, selects a template protein from a database of structurally resolved proteins and models the structure of the protein under investigation (the target protein) after that of the template protein with sequence alignment methods. If the sequence similarity between the template and the target is high enough (roughly 40% or larger), then this alignment can even serve as a scaffold for providing a full-atom model of the protein structure. The respective structure prediction method is called homology-based modeling and is described in Chapter 10.

The author of this chapter has developed one of the most advanced homology-based structure prediction tools to date. If the sequence similarity between the template and target is below 40% then generating full-atom models for the target using the template structure becomes increasingly difficult and risky. In such low-sequence-similarity ranges aligning the backbone of the target protein to that of the template protein becomes the critical issue. If this is done correctly, one obtains a 3-D model of target backbone that can serve as an aid for structural classification of the target protein. Chapter 11 describes this process. The author of Chapter 11 has codeveloped a well-performing Internet server for this structural alignment task.

Homology-based modeling can only rediscover protein structures since it models the target on the basis of a known template structure. In *de novo* structure prediction, we try to come up with the structure of the protein, even if it is novel and has never been seen before. This subject is still a major challenge for the field of computational biology, but significant advances have been made in the past 10 years by David Baker's group (University of Washington, Seattle, WA) and the author of the chapter was one of the major contributors in this context. Today, there are several projects that aim at resolving protein structures globally, e.g. over whole proteomes. The approach is a combination of experimental structure resolution of a select set of proteins that promise to crystallize easily and fold into new structures, and homology-modeling other proteins using the thus increased template set. Chapter 13 describes these structural genomics projects. One author of the chapter codirects the Protein Data Bank (the main repository for publicly available proteins structures) and the other directs a major structural genomics initiative.

The last two chapters discuss structure prediction of another important macromolecule in biology – RNA. In contrast to DNA, which basically folds into a double-helical structure, RNA is structurally diverse. There is a well-understood notion of secondary structure in RNA, i.e. the scaffold that is formed by base pairs within the same RNA chain. This algorithmically and biologically well-developed field is presented in Chapter 14. The authors of the chapter have contributed a major software package for analyzing RNA secondary structures. The last chapter in this part looks at tertiary structure prediction for RNA, a comparatively much less mature field, and its author is one of the major experts in that field, worldwide.

Volume 2 covers Parts 5–7. Based on the knowledge about molecular building blocks afforded by Volume 1, Volume 2 ventures into questions of molecular function.

Part 5 starts by considering atomic events in molecular networks, i.e. the interactions between pair of molecules. Molecular interactions are important in two ways. First, understanding which molecules bind in an organism, when and how, is fundamental for understanding of the dynamic basis of life.

Second, as we have seen in the first parts of this chapter, modifying molecular interactions in the body with drugs is the main tool for pharmaceutical therapy of diseases. Drugs bind to target proteins. Understanding the interactions between a drug and its target protein is a prerequisite for rational and effective drug therapy. Part 5 addresses both these questions. The part comprises four chapters. Chapter 16 discusses protein–ligand docking, with the implicit understanding that the ligands of interest are mostly drugs or drug candidates. The chapter discusses how to computationally dock known ligands into structurally resolved protein-binding sites and also how to computationally assemble new ligands inside the binding site of a protein. The senior author is the developer of one of the most widely used protein–ligand docking tools, worldwide. Chapter 17 discusses molecular docking if both docking partners are proteins. This problem is of lower pharmaceutical relevance, as most drugs are small molecules and not proteins, but of high medical relevance, as the basis of a disease can often be an aberration of protein–protein binding events. Furthermore, the chapter also discusses protein–DNA docking, which is at the heart of gene regulation. (Here, the protein is a transcription factor binding to its site along the regulatory region of a gene, for instance.) The authors of this chapter have developed advanced software for protein–protein docking. The last two chapters in this part discuss problems in finding drugs. As described above, the drug design process is decomposed into a first step, in which a lead structure is sought, and a second step, in which the lead is optimized with respect to secondary drug properties. If the binding site of the target protein is resolved structurally, lead finding can be done by docking (Chapter 16). Otherwise, one takes a molecule that is known to bind to the binding site of the target protein as a reference and searches for similar molecules as drug candidates. Here, the notion of similarity must be defined suitably such that similar molecules have similar characteristics in binding to the target protein. Chapter 18 discusses this type of drug screening. Finally, Chapter 19 addresses the optimization of drug leads. The authors of Chapter 19 are from the pharmaceutical industry. They are experts in applying and advancing methods for drug optimization in the pharmaceutical context.

Part 5 has advanced considerably beyond fundamental research questions and into pharmaceutical practice.

In *Part 6* we take a step back towards fundamental research. This part discusses the biochemical circuitry that is composed of the kind of molecular interactions that were the subject of Part 5. Understanding these molecular networks is clearly the hallmark of understanding life's processes, in general, and diseases and their therapies, in particular. However, the understanding of molecular networks is in its infancy, and is not advanced enough, in general, to be directly applicable to pharmaceutical and medical practice. Still, the vision is to advance along this line and the four chapters in this part present

various aspects of this process. Chapter 20 is on metabolic networks, the kind discussed in a little more detail in the beginning of this chapter. Metabolic networks are quite homogeneous with respect to the roles of the participating molecules. In general, we have a substrate that is converted to a product by a chemical reaction that is catalyzed by an enzyme, possibly with the aid of a cofactor. This homogeneity makes metabolic networks especially amenable to theoretical analysis. In addition, much is known about the topology (connection structure) of metabolic networks. However, we are still lacking much of the kinetic data needed to accurately simulate the dynamics of metabolic networks. The chapter presents methods for analyzing networks both statically and dynamically. The authors are among the main methodical contributors to the analysis of metabolic networks, worldwide. Chapter 21 analyzes gene regulation networks. These networks are more heterogeneous, since they incorporate different kinds of interactions – direct interactions, as when transcription factors bind to the regulatory regions of genes, and indirect interactions, as when transcription factors regulate the expression of genes that code for other transcription factors. Furthermore, proteins, as well as DNA and RNA, are involved in gene regulation. Inferring gene regulation networks necessitates much genomic information which is just on the verge of becoming available and, thus, the field is less mature than the area of analyzing metabolic networks. The author of Chapter 21 is one of the prime experts in the field of analyzing gene regulation networks. A very special type of molecular networks is concerned with transmission of information inside the cell. Usually, these signaling networks can be analyzed in terms of smaller modules than regulatory or metabolic networks. The special methods for analyzing these networks are presented in Chapter 22 by a group of outstanding experts in the field. Chapter 23 finally moves beyond the single cell and discusses interactions between a viral pathogen and its infected host cell – a major step from basic research to its application in a medical setting. This is a very young field and the author is one of its main proponents.

Part 7 is focused on a special types of experimental data that form the basis of much research (and debate) today – expression data. We have discussed the microarray (mRNA) expression data in the chapter above, when we addressed the quest of finding new target proteins for drug therapy. Expression data were the first chance to venture beyond the genome, which is the same in all cells of an organism, and analyze the differences between different cells, tissues and cell states. Therefore, these data have a special relevance for advancing molecular medicine and this justifies dedicating a separate part of the book with five chapters to them. Chapter 24 gives a summary of the whole field, from the experimental side of the technology of measuring mRNA expression and its implications on computational analysis methods to the bioinformatics methods themselves. Since expression data are typically

quite noisy, with many sources of variance residing both in the technology (which can be improved, in principle) and the underlying biology (which can and should not be changed), issues of quality control of the data play a prominent role in this chapter. The author is a global expert in the field of analysis of expression data. The following four chapters go into more detail on computational issues. Chapter 25 presents statistical methods for pretreating the data so as to arrive at an optimally interpretable dataset and it is written by a leading group of researchers in the area. The following two chapters discuss two fundamentally different kinds of analysis of mRNA expression data. Chapter 26 discusses methods that analyze and group different datasets (microarrays), generated under different circumstances (e.g. from different patients or from the same patient at different time points). Such methods afford the distinction of healthy from sick individuals as well as the analysis of disease type and disease progression, thus providing effective help in disease diagnosis. Chapter 27 groups data differently. Here, we are not interested in distinguishing different experiments, but in understanding the roles of (groups of) genes in, say, the progression of a disease that has been monitored with a sequence of microarray experiments. The results of the analysis are supposed to afford insight into the disease process and clues for drug therapy. This is a much harder task than just grouping microarray datasets and it has turned out that it cannot be solved, in general, just on the basis of expression data. Therefore, this chapter also prepares for later chapters that discuss the analysis of gene and protein function in a more general context (Part 8). The authors of Chapters 26 and 27 participate in a joint German national project that aims both at advancing the methods, and at applying them to biological and medical datasets. mRNA expression data (so-called transcriptomics data, because the data assess the expression level of mRNA transcripts of genes) have the advantage of being generated comparatively easily, due to the homogeneous structure of DNA (to which the mRNA is backtranscribed before measuring expression levels). However, these data correlate only weakly with the expression level of the actual functional unit, i.e. the synthesized and post-translationally modified protein. Measuring expression directly at the protein level is a more direct approach, but experimentally significantly more challenging. Therefore, the state of the field of proteomics, which analyzes protein expression directly, is behind that of transcriptomics, as far the experimental side is concerned. Nevertheless, proteomics is rapidly emerging, with several promising experimental technologies and the respective computational methods for data assembly/analysis. Chapter 28 presents the state of this field. It is written by a leading academic group engaged in software development for the field of proteomics.

Volume 3 builds on Volumes 1 and 2, and aims at embarking along an integrated picture of molecular function, and its consequences for the development and administration of drug therapies. The volume covers Parts 8–11.

Part 8 comprises eight chapters and is devoted to molecular (mostly protein) function. We have has already addressed aspects of molecular function (e.g. the chapters on molecular interactions and molecular networks, as well as the chapters on expression data), and, along the way, it has become increasingly evident that molecular function is a colorful term that has many aspects and whose elucidation relies on many different kinds of experimental data. In fact, molecular function is such an elusive notion that we dedicate a special chapter to discussing exactly this term, and the way it is and should be coded in the computer, respectively. This Chapter 29 is written by two authors that are main proponents of advancing the state of ontologies for molecular biology. Then we dedicate four chapters to inferring information on protein function from different kinds of data: sequence data (Chapter 30), protein interaction data that are based on special experimental technologies that can measure whether proteins bind to each other or not, and do so proteome-wide, in the most advanced instances (Chapter 31), genomic context data, affording an analysis based on the comparison of genomes of many species (Chapter 32), and molecular structure data (Chapter 33). Since all of these data still do not cover protein function adequately, we add another chapter that addresses methods for inferring aspects of protein function directly from free text in the scientific literature (Chapter 34). Chapter 35 presents methods for fusing all the various kinds of information gathered by the methods presented in the preceding chapters to arrive at a balanced account of the available knowledge on the function of a given protein. Finally, Chapter 36 discusses the druggability of targets, i.e. the adequacy of proteins to serve as a target for drug design. This quality encompasses properties such as a suitable shape of the binding pocket to suit typical drug molecules and a certain uniqueness of the shape of the binding pocket, such that drugs that bind to this pocket avoid binding to other proteins that are not targets for the drug. Again, all of these chapters are written by outstanding proponents of the respective fields.

With Parts 1–8 we have covered the space from the genotype (the genome sequence) to the phenotype (the molecular function). However, we can still take additional steps to making all of this knowledge work in applied medical settings. This is the topic of *Part 9*. To this end, Part 9 focuses on the analysis of relationships and differences between genomes. In the first chapter, Chapter 37, the topic is rolled up in a general fashion by asking the question: "What can we learn from analyzing the differences between genomes?". Then, we focus on the medically most relevant differences between genomes of individuals of the same species. Specifically, we are interested in the human and in pathogens infecting the human. Chapter 38 discusses what we can

learn from genetic differences between people about disease susceptibility. Chapter 39 then addresses the topic of personalized medicine: how can we learn from suitable molecular and clinical data how a patient reacts to a given drug treatment? The final two chapters address the evolution of pathogens in the human host (mostly to become resistant against the host's immune system and drug treatment). Chapter 40 discusses viral pathogens, specifically HIV, the virus that leads to AIDS. Chapter 41 covers the bacterial world. The authors of all chapters have made seminal contributions to the topic they are describing.

Part 10 is an accompanying section of the book that addresses important informatics technologies that drive the field of computational biology and bioinformatics. There are three chapters. Chapter 42 is on data handling. Chapter 43 discusses visualization of bioinformatics data; here, molecular structures are not the center of the discussion, since their visualization is in a quite mature state, but we focus on microscopic images data, molecular networks and statistical bioinformatics data. Chapter 44 focuses on acquiring the necessary computational power for performing the analysis from computer networks (intranets and the Internet). There is a special research community that provides the progress in the underlying informatics technologies and the authors of these chapters are outstanding proponents of this community.

In *Part 11*, finally, Chapter 45 addresses in a cursory manner emerging trends in the field that were too new at the time of the conceptualization of the book to receive full chapters, but have turned out to become relevant issue at the time that the book was written. Thus, this chapter gives a cautious and anecdotal look into the future of the field of bioinformatics.

The goal of this book is to provide an integrated and coherent account of the available and foreseeable computational support for the molecular analysis of diseases and their therapies. The authors that have contributed to the book represent the leading edge of research in the field. We hope that the book serves to further the understanding and application of bioinformatics methods in the fields of pharmaceutics and molecular medicine.

References

1 COLLINS, F. S., E. D. GREEN, A. E. GUTTMACHER AND M. S. GUYER. 2003. A vision for the future of genomics research. Nature **422**: 835–47.

2 DERISI, J. L., V. R. IYER AND P. O. BROWN. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278**: 680–6.

3 DREWS, J. 2000. Drug discovery: a historical perspective. Science **287**: 1960–4.

4 FISCHER, E. 1894. Einfluss der Configuration auf die Wirkung der Enzyme. Ber. Dt. Chem. Ges. **27**: 2985–93.

5 PAPAVASSILIOU, A. G. Clinical practice in the new era. A fusion of molecular biology and classical medicine is

transforming the way we look at and treat diseases. EMBO Rep. 2001. **2**: 80–2.

6 RILEY, M., T. ABE, M. B. ARNAUD, et al. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot – 2005. Nucleic Acids Res. **34**: 1–9.

7 SAMS-DODD, F. 2005. Target-based drug discovery: is something wrong? Drug Discov. Today **10**: 139–47.