Chapter 1

Introduction

OVERVIEW

Today, 50 years after the elucidation of the structure of DNA, high-performance computing, data management, and the Internet are making possible the large-scale industrialization of many aspects of biomedical research. The basic steps of identifying, purifying, and cloning a gene, followed by purification and characterization of the proteins coded for by that gene, have been automated and streamlined to a degree that no one could have predicted just ten years ago. Superimposed on this trend has been a rapid evolution in the design of large-scale computing infrastructure for bioinformatics. Coupled with explosive growth in the quantity of available bioinformatic data, these improvements are driving a migration from *in vivo* (observations of real life) to *in vitro* (test tube experimentation) to *in silico* (experimentation by computer simulation).

Central to this theme is an emerging understanding of the complex relationship between genotype and phenotype. A genotype is inherited as a code; a phenotype is the physical manifestation of that code. Between the two lie an enormous number of probabilistic events. In principle, a complete understanding of the code and the rules for its expression, coupled with information about the environment, should be enough to allow the prediction of phenotype. However, the tremendous number of variables involved makes such predictions nearly impossible given today's technology.

This book discusses the relationship between genotype and phenotype with a focused view of the flow and management of biological information. Its organization parallels

that flow—gene to transcript to protein to metabolic system. It is the emerging understanding of this information flow and the migration to *in silico* research that promises the greatest advance of our time: the launch of molecular-based medicine and the first true understanding of the molecular basis of health and disease.

The theme of information flow from gene to complete organism rests on a foundation composed of a very small number of key concepts. These concepts are referred to throughout the book and form the core of many important discussions. They are illustrated in the sections that follow.

COMPUTATIONALLY INTENSE PROBLEMS: A CENTRAL THEME IN MODERN BIOLOGY

The ability to solve computationally intense problems has become a core competency and a driver of many industries and academic pursuits. Although the trend is not new, the complexity and size of the information infrastructure required to solve such problems has increased dramatically over the past few years. The result has been a noticeable increase in the rate of development of new algorithms and technologies that address technical computing problems in new and ingenious ways. For instance, today's aerospace industry is built on a computer infrastructure that facilitates the complete *in silico* design of jet aircraft, the automotive industry lives on a diet of algorithms for computational fluid mechanics and thermodynamics, and the petroleum industry relies on computer-assisted geophysical exploration to discover oil.

Each of these industries has evolved to take advantage of the rapid advances in information technology that continue to shift the balance from physical to *in silico* experimentation.

An important computational milestone was reached in 1995, when the new Boeing 777 aircraft was designed entirely *in silico* by 238 cross-functional engineering teams collaborating across 2,200 workstations using the computer-aided three-dimensional interactive application (CATIA) system. The system worked so well that the first assembled flight vehicle was only 0.03mm out of alignment and perfectly safe for carrying passengers [1].

During the past few years, biology and medicine have taken their rightful place on the list of industries that depend on high-performance technical computing. As a result, two new fields have emerged—computational biology and its derivative, information-based medicine. Computational biology is a superset of traditional bioinformatics because it includes new technical initiatives such as *in silico* molecular modeling, protein structure prediction, and biological systems modeling. Information-based medicine is a system of medical care that supplements traditional evidence-based diagnoses with new insights gleaned through computerized data acquisition, management, and analysis. Throughout this book, it will become clear that information-based medicine depends heavily on computational biology. This new and emerging era of medicine depends strongly on a broad array of new technologies such as high-throughput DNA sequencing, gene-expression profiling, detection and quantification of trace amounts of specific proteins, and new algorithms for pattern matching in large heterogeneous datasets. The successful deployment of most of these technologies depends on the availability of very high-performance computing infrastructure. Most recently the focus has been clusters of commodity-priced machines.

Computational biology is built on a new class of technical problem and associated algorithms that are still evolving. Over the past few years, it has become clear that most biological problems that can be described mathematically can also be divided into a

large number of small self-contained computations. This characteristic of biological problems makes them amenable to solution in a cluster of computers rather than on a single large machine. The high-performance computing community typically refers to such problems as being "embarrassingly parallel," and designers of bioinformatic algorithms have been quick to take advantage of these attributes by building parallel infrastructure (most often Linux-based clusters composed of large numbers of relatively small machines). Such clusters have now become a dominant force in bioinformatics and, despite its recent emergence, computational biology has become a principal driver of one of the most important trends in information technology: the migration from traditional large multiprocessing servers to clusters of commodity-priced machines.

Bioinformatic problems generally fall into one of two broad technical categories: floating point or integer. Algorithms that ultimately count something (e.g., number of correct alignments in a sequence comparison) are considered to be integer in nature. Conversely, complex calculations that involve statistical analysis or operations of higher mathematics are considered to be floating point in nature. Such problems tend to execute faster on systems with more powerful floating-point processors, whereas their integer-intensive counterparts tend to execute fastest on machines with the highest possible clock speed. Floating-point bioinformatic problems typically involve complex algorithms borrowed from physical chemistry or quantum mechanics. In recent years, molecular modeling and metabolic systems simulation have become central to the drug discovery process. Such applications tend to be floating-point-intensive. Many of the newest applications that form the basis of information-based medicine, such as image processing, visualization of 3D graphics, and natural language processing (NLP), are also floating-point-intensive. Conversely, integer-based bioinformatic problems typically depend on algorithms that compare characters in sequences or search databases for matching phrases and terms. Much of contemporary molecular biology, including genome sequencing, is built on the solutions to such problems. The algorithms used to align gene sequences and search for patterns are important examples of integer-intensive bioinformatic problems. One of the most significant applications in this class is the assembly algorithm that was used to construct the human genome from millions of sequence fragments. The new technique, known as whole genome shotgun sequencing, is enormously compute-intensive. Not surprisingly, the complete assembly of the three-billion-base human genome represents one of the most complex logic problems ever solved.

Bioinformatic problems, both floating point and integer, are often well suited to solution in parallel computing environments because the operations they depend on are atomic in nature. Protein folding is an excellent example; the problem may be broken into thousands of individual calculations, each representing the attractive force between two atoms in the protein. These pairwise force calculations are each assigned to a different node in the cluster. A typical folding calculation consists of an enormous number of time steps—typically on the order of 10^{15}. During each time step, the forces acting between each pair of atoms in the molecule are calculated and a new transitional structure is generated. The process is iteratively repeated on every node for every time step, and the atomic nature of the calculations ensures that the problem will scale linearly as compute nodes are added to the cluster. This property of linear scalability distinguishes clustered computing environments from large symmetrical multiprocessing machines. Unfortunately, even today's largest clusters are too small to solve thousands of pairwise force calculations for thousands of trillions of time steps. Additionally, communication latencies can prevent such problems from scaling linearly across very large numbers of nodes. The solution has been to restrict the problem to only the most relevant portion of the molecule.

Sequence-homology and pattern-discovery problems also lend themselves to solution in clustered computing environments. In many cases, a large number of sequences need to be matched against a single genome or the contents of a sequence database. There are two distinct approaches for dividing the problem among a large number of machines. The first involves using each node to compare a single test sequence against the entire database; the second involves dividing the target sequence (sometimes an entire genome) across the cluster, using each node to compare all test sequences with a small fragment of the sequence, and managing overlap at the boundary of each node. Although fundamentally more complex, the second solution is well suited to situations where a small number of search sequences are compared to a large target.

BUILDING THE PUBLIC INFRASTRUCTURE

This book refers to data gathered during the past 25 years by the collective community of molecular biologists. This information is warehoused in a public infrastructure maintained by the scientific community and available to anyone with an Internet connection. Virtually all work performed in the life sciences community depends in some way on the availability of this ever-growing body of knowledge. Because of its importance, we have chosen to spend some time describing the infrastructure and have included sample records from some of the databases. Not surprisingly, the size of these databases has grown exponentially over the past two decades. The history of its evolution is significant because today's medical community is embarking on a similar program designed to build a public infrastructure to support medical informatics.

Early bioinformatic efforts, which focused on the construction of databases containing protein and DNA sequences, depended more on advances in the chemistry of protein and nucleotide sequencing than on advances in computer science. Regardless of the focus, biology and information science became entangled in a symbiotic relationship that has provided tremendous benefits to both disciplines. During the early part of the 1980s, while the first protein and nucleotide sequence databases were being built, physical biochemists began writing the first computer programs for Fourier transform analysis of x-ray crystallographic data, enzyme and chemical kinetics, various types of spectroscopy, statistical analysis of protein structure, and ligand binding experiments. Molecular biology, however, was still a new science in the process of maturing from its early focus on bacterial genetics into a mainstream medical discipline based on a broad understanding of eukaryotic chromosomes—most notably those of the human genome. In the mid-1980s, it would have been impossible to predict that the one-gene-one-protein model would collapse during the human genome sequencing project and that the level of complexity would skyrocket as researchers unraveled the complex control mechanisms that form the foundations of gene expression. In those days, genes and proteins were sequenced one at a time using tedious processes that limited the amount of data generated and microarrays capable of simultaneously measuring expression levels for hundreds of thousands of genes were 15 years in the future. A small group of scientists began to envision the industrialization of molecular biology and, as DNA sequencing techniques improved, leaders in the field began to articulate a plan for sequencing the entire human genome. These thoughts spawned the public infrastructure that now contains dozens of protein and DNA sequence and structure databases. The wide availabil-

ity of computer horsepower also spawned the development of algorithms for pattern discovery and sequence-homology testing. These algorithms became the foundation for today's science of bioinformatics—computational biology was becoming a recognized discipline that would soon be worthy of its own advanced degree programs.

These trends continued through the late 1980s with the emergence of the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), and other international centers for the management of biological data, such as the European Bioinformatics Institute (EBI). Then, in October 1992, NCBI assumed responsibility for the GenBank DNA sequence database. GenBank, the first publicly available nucleotide sequence database, remains the most comprehensive source of gene sequence information to this day. During its first year in operation, the database grew by approximately 1.7 bases per hour. Rapid advances in nucleotide sequencing technology coupled with improvements in the algorithms used to analyze alignments drove this number to more than 850,000 bases per hour by the end of the decade. During the same timeframe, similar advances in protein sequencing technology enabled the development and rapid expansion of a number of protein sequence and structure resources, including the Protein Information Resource (PIR), Swiss-PROT, Protein Research Foundation (PRF), and Protein Data Bank (PDB). Since then, the public infrastructure for aggregating, managing, and disseminating biological data has grown to include hundreds of such databases which, coupled with proprietary sources, has become the foundation for modern-day, computer-driven, rational drug design.

THE HUMAN GENOME'S SEVERAL LAYERS OF COMPLEXITY

A recurring theme visible throughout this book is the incredible density of information contained within the human genome. We have come to take for granted the shocking fact that all the information needed to code for an entire person can easily fit within the nucleus of a single cell. That thought is not lost within the pages of this book, and we will spend a considerable amount of time discussing the information-processing mechanisms that enable such density of information to be achieved.

Over the past several years, our understanding of gene expression has evolved to reveal a multistep process that embodies an order of magnitude increase in complexity at each step, starting with transcription and ending with the metabolic pathways that define a functioning organism. Without this stepwise increase in complexity, it would be impossible for the human genome, which contains approximately 25,000 genes, to code for more than a million proteins. Nature, in an incredible display of efficiency, has created a system that requires only ten times as much information to code for a person as for an *E. coli* bacterium [2].

The following mechanisms achieve tremendous diversity from the relatively small number of coding regions that make up the human genome:

- Each coding region contains six distinct reading frames—three in each direction.
- A single messenger RNA molecule, through alternative splicing, can code for dozens of different messages called splice variants. Each variant codes for a different protein.
- Individual proteins are post-translationally modified. These modifications can include the removal of amino acids as well as the addition of chemical side chains—sugars, acetyl groups, carboxyl groups, methyl groups, and many others. Protein function is highly dependent on these modifications and, in some cases, improperly processed proteins are completely dysfunctional.
- Protein function is context sensitive in the sense that identical proteins can have completely different roles in different parts of an organism.
- Protein-protein interactions are the final determinant of phenotype. Hundreds of millions of such interactions form the metabolic basis of life.

Figure 1-1 contains a diagrammatic representation of this stepwise buildup of complexity.

Without molecular-level mechanisms for adding diversity at each step in the gene-expression process, the human genome would need to be considerably larger and more complex. However, because a genome organized in this fashion would contain separate coding regions for each protein, bioinformatics would become a much less complex science. Protein sequences would be determined directly from the gene sequence; the lack

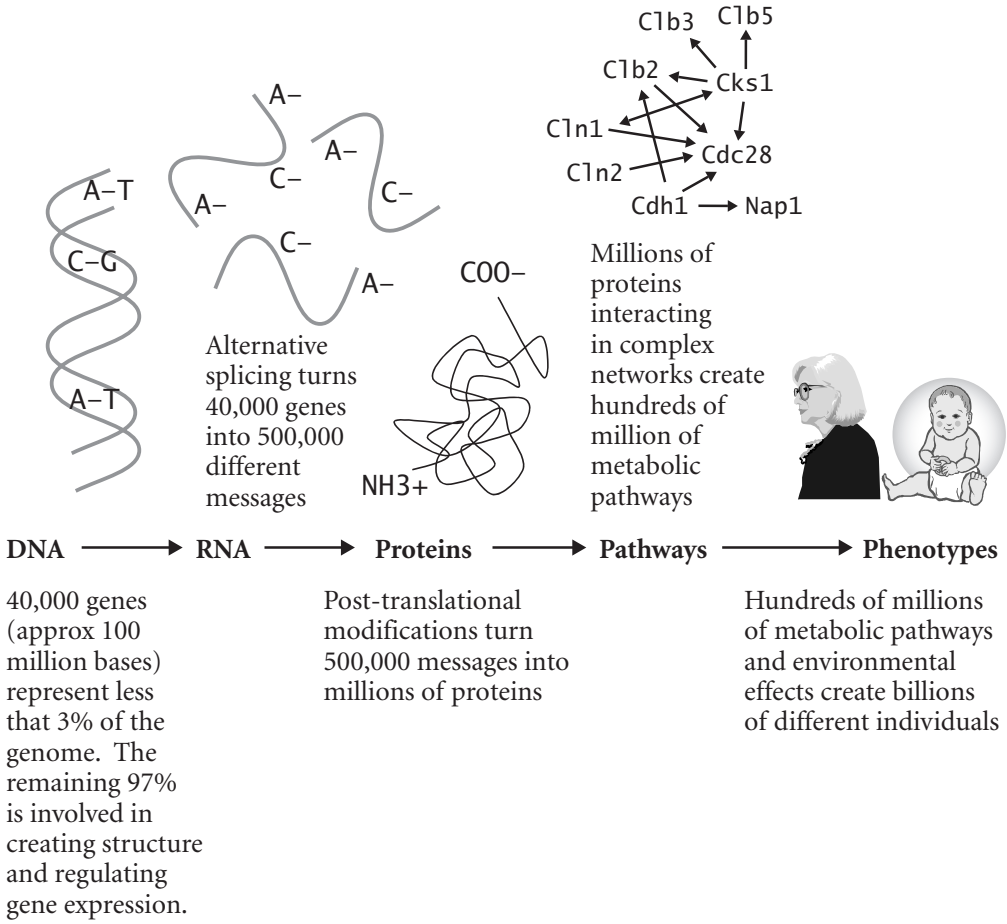


Figure 1-1 Human diversity is achieved through a stepwise progression that adds complexity at each step in the gene-expression process.

of splice sites within coding regions would ensure that start and stop signals for a given protein would always appear in the same reading frame; only one DNA strand would contain code, the complementary strand would again be referred to as the nonsense strand as it was 20 years ago; lack of splicing at the mRNA level would ensure a direct correlation between gene-expression studies and the base genome sequence; and a point

mutation in the base DNA sequence could affect only a single protein. Unfortunately, the human genome is not constructed in this way; fortunately, bioinformatics has evolved to include complex statistical analysis techniques and pattern-matching algorithms for solving these problems.

TOWARD PERSONALIZED MEDICINE

Our previous discussions have made many references to the flow of biological information from genome to complete organism. In its complete form, that information and associated computational tools form the basis of a discussion that can be meaningful only at the level of an entire organism. For our purposes, that organism is a human being, and our discussion will be framed in the context of personalized medicine.

We have chosen to place this section near the end of the book because it builds on all the concepts presented in earlier sections; personalized medicine is a sophisticated application of computational biology and basic bioinformatics. The modern drug discovery process is undeniably central to this discussion. The process, often visualized as a discovery "pipeline," begins with basic research and ends with disease-specific pharmaceuticals. Like biology and medicine, drug discovery has also become an information science. The impact of this trend has been tremendous. For example, using traditional drug development techniques it took nearly 40 years to capitalize on a basic understanding of the cholesterol biosynthesis pathway to develop statin drugs—those that inhibit the enzyme HMG-CoA Reductase, the rate-limiting step in cholesterol biosynthesis [3, 4]. Conversely, a molecular-level understanding of the role of the HER-2 receptor in breast cancer led to the development of the chemotherapeutic agent Herceptin within only three years [5]. The developers of Herceptin enjoyed the advantages of *in silico* molecular modeling, high-throughput screening, and access to databases containing genomic and proteomic information. Biochemistry and pharmacology have advanced considerably since the launch of statin drugs, and today's computational biologists enjoy the advantages of a new generation of applications for studying molecular dynamics, predicting protein tertiary structure, and identifying genes that are coregulated in various disease states and individuals. As these tools mature, larger portions of the drug discovery process will make their way from the lab bench to the computer. We

are already witnessing the beginning of this trend as evidenced by the emphasis being placed by large pharmaceutical companies on the power of their information technology platforms. These platforms have become key differentiators and true sources of value creation in the drug discovery business.

ILLNESSES ARE POLYGENIC

Systems that are capable of supporting personalized medicine must be built on an infrastructure consisting of many different clinical and research databases. The aggregation of large quantities of such content has fostered a variety of important statistical studies. One important realization that has emerged from these studies is that virtually all diseases are complex in the sense that they are polygenic.

Building on this thought, it is important to note that the past several years have witnessed the collapse of two overly simplistic views of biology: the one-gene-one-protein model of molecular genetics and the one-protein-one-disease model of medicine. For most of the twentieth century, biochemical research was focused on the delineation of complex metabolic pathways. Much of this early work was concerned with identifying individual proteins, their functions, and the roles they play in metabolism. A natural consequence of such work was to uncover the relationships between metabolic defects—missing or aberrant proteins—and disease. These discoveries were the genesis of a monogenic view of many diseases. For example, diabetes was traditionally viewed as an insulin deficiency, and cardiovascular disease was thought of as being caused by improper regulation of cholesterol biosynthesis. Although these assertions are true at a basic level, most diseases are, in reality, complex polygenic disorders that can be fully understood only at a systems level. Diabetes is now thought of as a very complex disorder with a variety of environmental and genetic components. The standard nomenclature that divides the disease into two major classes—type I and type II—is now known to mask a more complex variety of genetically distinct subtypes. Not surprisingly, changes in the expression levels of many of these genes have far reaching impacts that are difficult to predict. For example, a well-documented but subtle relationship exists between type II diabetes and certain mood disorders; depression and many of its associated symptoms constitute a major risk factor for the development of the disease [6].

Likewise, as molecular-level data became available, cardiovascular disease evolved from a simple problem of arterial obstruction to a metabolic disease and finally to part of an inflammatory process.

Variations in base gene sequence represent only a small part of the story. Within the context of molecular medicine, four basic sets of parameters make each of us what we are

- Basic genetic sequence at the DNA level
- Environmental effects (including exposure to radiation and mutagens) on gene expression
- Stochastic and nonstochastic probabilistic functions that affect gene expression
- Viral infections that alter the genomes of individual cells

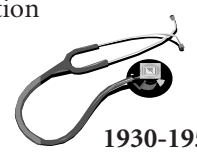
Each of these parameters plays an important role in gene expression. An individual's genome sequence contains important information about polymorphisms that are the root cause of many diseases as well as basic information that can be used to predict many physical characteristics. Over time, exposure to various elements in the environment has profound effects on gene expression. The expression levels of most genes are controlled by molecular-level feedback mechanisms, and these are often affected by the environment. Stochastic processes are sequences of probabilistic events that cause the expression levels of specific genes to vary between predictable values. Finally, viral infections can alter the genomes of individual cells by inserting new sequences into the chromosomal material. Such changes accumulate throughout an individual's life, and the effects associated with them are only vaguely understood.

In the past, lack of detailed information about each of these parameters has caused the medical community to rely exclusively on phenotypic descriptions of illnesses. The process of defining diseases by their phenotypes is giving way to a more precise set of molecular-level definitions. For example, psychiatric disorders such as schizophrenia and depression are no longer thought of as simple diseases but as broad phenotypes displayed by patients with many different gene-expression profiles, genome sequences, and medical histories. An accurate view requires an understanding of large numbers of proteins, protein interactions, and many metabolic pathways [7]. Responding to this complexity, researchers have turned to mRNA expression profiling, where the goal is to

correlate the complex patterns of gene expression and medical history with treatment outcomes. The technique involves profiling the up and down regulation of specific genes using microarray technology and analyzing the resulting data to help identify potential protein targets for drug therapy. Information technology is a crucial component at many stages of the process, beginning with target identification where a single microarray can produce hundreds of thousands of individual spots, each containing information about the expression of a single nucleotide sequence. Software products are often used to reorganize the information according to certain criteria such as cell function—protein synthesis, carbohydrate metabolism, energy production, etc. When clusters of coregulated genes are identified, heterogeneous database access tools are often used to search the public database infrastructure for the most current and accurate relevant information. Automating this process and linking the databases is a daunting challenge. After the data are retrieved, they are often compared with internal proprietary sources inside the corporate firewall. Additionally, large-scale database searches utilizing pattern-matching algorithms are used to find expression profile matches in databases containing millions of such patterns. The task is similar to that of scanning a database containing millions of fingerprint records for patterns similar to a specific reference fingerprint. In the vast majority of situations, where thousands of genes are involved, it is necessary, for a given set of expressed genes, to correlate specific up and down regulation patterns with treatment outcomes and phenotypic changes. Establishing the relevant correlations requires analyzing complex datasets using knowledge-management tools and searching data sources with different schemas and data structures.

The challenge of properly classifying illnesses is further complicated by the fact that virtually all diseases affect only a particular subpopulation of cells. To address this problem, medical research databases must be populated with cell-specific gene-expression data for specific diseases. Additionally, cell specificity is a complex affair because it is not always possible to predict all the different populations of cells that might be affected by a specific illness, and changes in gene-expression patterns across different populations of cells are likely to be relevant in almost any disease state. When this sort of exhaustive data is available for specific diseases across a large population, it will likely be possible to improve patient stratification. One possible result will be the rejuvenation of previously abandoned lead compounds that failed clinical trial or appeared to lack specificity for the target patient population. Large-scale microarray data analysis has also revealed new relationships between disease categories that were previously considered to be unrelated.

Pre-1930: History and physical examination



1930-1950: Simple diagnostic tools

Limited biochemical insight
No bio or medical informatics

1950-2000: Advanced diagnostics including sophisticated chemical tests.

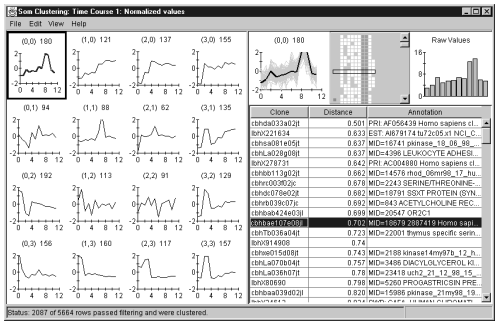
Beginning of medical informatics

Rapid advances in biochemical technology

Elucidation of basic metabolic pathways

Compute power doubles every 18 months

Launch of bioinformatics



2000-: Era of computational biology and information-based medicine. Molecular-level understanding of disease supported by advanced computing infrastructure and new tools for data analysis.

Figure 1-2 The history of medicine from simple diagnostic tools to advanced computing infrastructure and computational biology.

For example, at the genetic level some types of chronic inflammation share gene-expression profiles with certain malignancies. This discovery has led to the testing of cyclooxygenase (cox-2) inhibitors as possible treatments for lung and colon cancer. (Cox-2

inhibitors can interfere with angiogenesis and, therefore, block the vascularization process for many tumors.) The transformation of medicine from an empirical science, limited by the knowledge and personal experience of individual physicians, to a precise discipline built on databases and algorithms is depicted in Figure 1-2.

Recent gene-expression experiments have revealed that many proteins involved in specific disease states are expressed as "minor messages," mRNA transcripts that appear only in vanishingly small quantities within the cell. This discovery has added an additional level of complexity because such messages exercise their control at the single-copy-count level where accurate measurement is extremely difficult. It is becoming clear that single-digit changes in the number of such molecules can signal the onset of disease. The presence of large numbers of low-copy-count messages dictates that measurement ranges be accurately extended far below traditional levels. Table 1-1 presents a relative count of the number of distinct mRNA species in each copy-count category for a typical human monocyte. The data reveal that the vast majority of transcripts, more than 75%, are present in fewer than 100 copies, and that half are present in fewer than 10 copies. Because a typical human cell contains approximately one million mRNA transcripts at any point in time, these small copy-count species can be thought of as existing in the part-per-million range.

Table 1-1 Total Number of mRNA Species Grouped into Size Classes

Total Number of mRNA Transcripts in Each Class	Number of Different mRNA Species in Each Class
> 10,000	6
1,000–9,999	133
100–999	1,418
10–99	10,642
5–9	24,890

Source: Lynx Therapeutics, Hayward California

NEW SCIENCE, NEW INFRASTRUCTURE

Most of the discussion to this point has been concerned with computationally intense data-analysis techniques and their application to new areas—most notably information-based medicine. These new advances are strongly dependent on the power available in contemporary computing platforms. Fortunately, the platforms are rapidly evolving to meet this challenge. The trend is remarkable because it is unusual for advances in one science to become the engine of growth for another. However, that is exactly what is happening in the worlds of medicine and drug discovery, where improvements in information technology are becoming both a growth engine and a core value.

Several forces and technical trends have become evident: Expression array technology is becoming a commodity; high-throughput gene sequencing technologies are becoming available at reasonable cost; diagnostic imaging (CAT, MRI, PET, x-ray, and various ultrasound techniques) are generating large volumes of digital data; new algorithms for pattern discovery and large-scale data mining are facilitating statistical analysis across large patient populations; clusters of commodity-priced computers are replacing supercomputers at a fraction of the cost; and low-latency network connectivity has fostered the launch of a new information architecture known as a "computer grid."

Computer grids allow geographically dispersed systems to be linked into a single entity that appears to users as one system. Grids are often described as falling into one of two different categories: compute grids and data grids. Compute grids allow the parallelization of processor-intensive applications and provide a level of compute performance that cannot be achieved on individual systems. Compute grids are already impacting such diverse fields as protein folding and image processing.

Data grids provide similar functionality by linking geographically dispersed databases into a single view that can be queried from a single system. Among the most complex of these challenges is the linking of geographically dispersed heterogeneous systems with dissimilar datasets. These datasets often have different structures and internal references creating "schema" mismatches, an area of intense research within the IT community.

One of the focus areas for personalized medicine involves the analysis of imaging data from various sources mentioned earlier—CAT, x-ray, PET, MRI, and diagnostic ultrasound. Someday, in the not-too-distant future, doctors will compare medical images of their patients to millions of other images, and fast pattern-matching algo-

rithms capable of spotting similar images will be used in conjunction with other clinical and genomic information to identify close matches. Information contained in the medical records of these matching patients will be used to select the best treatment strategies. The process will also involve comparing base DNA sequence information and ongoing gene-expression profiles for large numbers of patients—both healthy and ill. The collection of gene-expression profiles will occupy an enormous amount of storage space and, because data will be collected across many time points, algorithms for comparing mRNA expression profiles in a time-dependent fashion will form a core component of the clinical tool set. During the next few years, high-throughput sequencing techniques, supported by a new generation of fragment-assembly algorithms, are likely to make the collection of complete genome sequences a standard component of routine medical care. As previously mentioned, the computational challenges associated with sequencing are significant. For instance, fragment assembly for the human genome project required constant operation of two maximally configured supercomputers for more than one year.

The computer infrastructure required to support widely available clinical genome sequencing will be significantly larger than today's hospital information systems, and each will be local to an individual medical institution. Results from sequencing operations are likely to be shared in a data grid along with the images and gene-expression profiles mentioned earlier. After a patient is classified as belonging to a particular group composed of metabolically similar individuals, an ongoing record of the treatment results, expression profiles, and other relevant data will be added to the ever-growing pool of information.

Several technical challenges must be overcome to make this vision a reality. Huge amounts of data must be collected and shared across a large dispersed infrastructure—most likely a computer grid, tools for analyzing and comparing the data must be made available to researchers and clinicians, new database schemas must be designed and deployed, and the medical community must come to agreement on a new set of standards for the representation of clinical medical records. These technical challenges form the core of an infrastructure development effort that is proceeding at major medical centers and government installations all over the world today.

These enhancements to the clinical infrastructure are beginning to drive the development of a new generation of applications that improve the human/computer interface.

Included are systems for natural language processing, automated workflow management, and the generation of ontology-based semantic queries against disparate data sources. The last example is particularly significant for this discussion because biology and medicine are technical disciplines described using a variety of objects and concepts. An ontology is an explicit specification of such a conceptualization. As discussed in Chapter 2, precise ontologies are rapidly becoming the basis for database query systems across all life science disciplines.

Perhaps the most urgent item on this list is the need to generate queries across heterogeneous databases. In response to this need, researchers have begun to create a new generation of tools capable of translating the standard SQL queries generated by most desktop tools into the language of each underlying data source and joining the results into a consolidated response. Various vendors have created tools to facilitate this process—most are based on "wrapping" individual data sources with code that provides a standard interface to a single query tool. Systems that use such virtualizations of data are routinely referred to as being "federated." Alternative approaches involve the construction of large "data warehouses" containing restructured and consolidated information from the various heterogeneous sources. Each approach has strengths and weaknesses, but together they can form the basis of a complete information-management platform for bioinformatics.

THE PROACTIVE FUTURE OF INFORMATION-BASED MEDICINE

Predictive power is the ultimate test of any science. With this view in mind, we have decided to close our discussion of personalized medicine on the subject of predictability. The area is characterized by two major dimensions. The first involves predicting clinical outcomes for illnesses that have already been diagnosed. The second, sometimes referred to as presymptomatic testing, is the focus of the present discussion. As one might expect, presymptomatic testing involves proactively predicting the onset of disease before any physical symptoms become apparent. The databases and computational tools that comprise the infrastructure for information-based medicine coupled with a new generation of ultrasensitive chemical techniques for identifying trace amounts of

circulating proteins and minor messages within cells are all part of the arsenal of tools required for this new and emerging discipline. The presymptomatic testing process depends on the development of disease-specific databases populated with cell-specific metabolic and gene-expression data, clinical and demographic information, and extensive statistical information regarding the onset of disease in treated and untreated individuals. Figure 1-3 depicts the logistics of presymptomatic testing.

The logistics of presymptomatic testing for disease include complex queries against disease-specific databases; new algorithms for pattern recognition and data mining; and algorithms for joining clinical, demographic, and gene-expression data in a single query. Solutions will ultimately be deployed at thousands of sites.

Presymptomatic testing for disease is destined to become a core component of the healthcare system because it has the potential to reduce downstream medical costs.

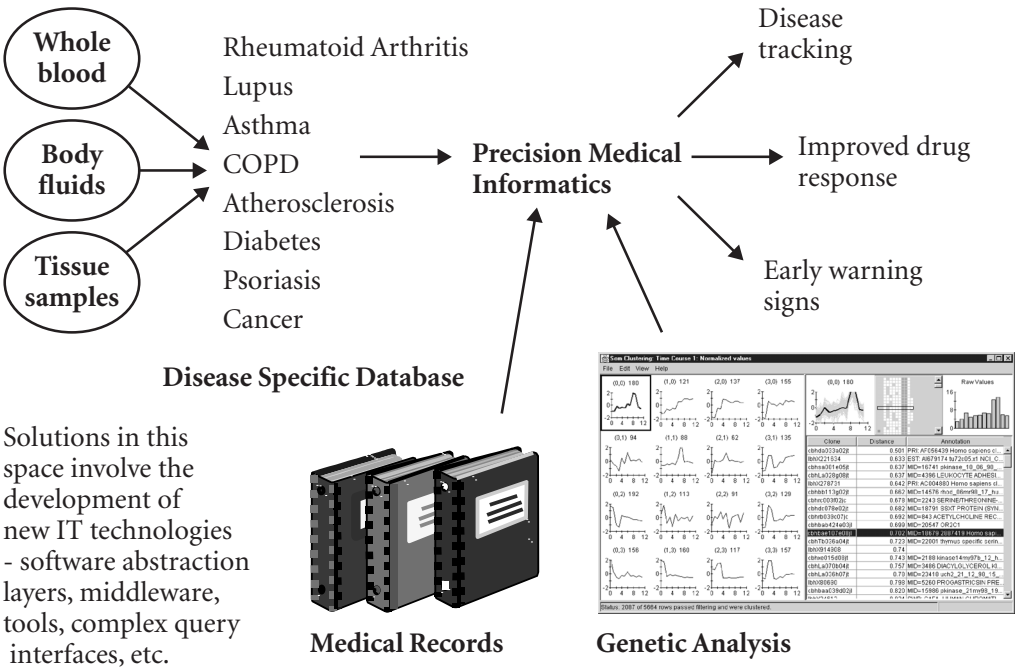


Figure 1-3 The logistics of presymptomatic testing.

However, because the testing occurs at an early stage, before the onset of symptoms, it is necessary to accurately measure very small amounts of biologically active substances—proteins and nucleotides. Such measurements require a more sophisticated set of chemical tools and techniques for mathematical analysis than those commonly used for medical diagnostics or gene-expression profiling. One of the greatest challenges is extending the range of measurement. Gene-expression studies must scale from copy counts in the single-digit range to tens of thousands, and circulating protein levels must be measurable from milligrams to nanograms. These measurement ranges have mathematical implications as well. For example, the statistical analysis that underlies gene-expression profiling depends on a variety of algorithms for clustering together coregulated messages. Most of the current crop of statistical algorithms are designed to function in the high-copy-count range where microarrays are the measurement technology of choice. Counting individual copies of expressed genes requires both new chemistry and new statistical methods. Furthermore, these experiments are time-dependent in a very sensitive way, and new mathematical methods are being developed to analyze the large number of datasets that must be generated to cover a meaningful number of time points. Presymptomatic testing is likely to become one of the most influential forces in modern computational biology. Advanced molecular diagnostics, time-dependent gene-expression studies, and new techniques for detecting trace amounts of metabolites are central to the newest biological science—systems biology. The systems approach is central to information-based medicine. Over the next few years, standardization efforts will allow the technology to become a true source of value creation for the clinical physician.

ENDNOTES

1. Norris, G., and Wagner, M. *Boeing 777: The Technological Marvel (Jetliner History)*. Motorbooks International, 2001.
2. The reference is to the number of coding regions rather than the number of total bases. The human genome contains nearly 1,000 times as many bases (3.2 billion versus 4.6 million) but fewer than 10 times as many coding regions as the *E. coli* genome. The structure of the human genome is covered in detail in a later section.
3. Corsini A., Bellosta S., Baetta R., Fumagalli R., Paoletti R., Bernini F. 1999. New insights into the pharmacodynamic and pharmacokinetic properties of statins. *Pharmacology & Therapeutics* 84: 413–428.

4. Corsini A., Maggi F. M., Catapano A. L. 1995. Pharmacology of competitive inhibitors of HMG-CoA reductase. *Pharmacology Research* 31: 9–27.

5. Slamon D. J., Godolphin W., Jones L. A., et al. 1989. Studies of the HER-2/neuproto-oncogene in human breast and ovarian cancer. *Science.* 244: 707–712.

6. Dominique L. M., Betan E., Larsen H., Phillips L. S. 2003. Relationship of depression to diabetes types 1 and 2: epidemiology, biology, and treatment. *Biological Psychiatry* 54 (3): 317–329.

7. Kitano H. 2002. Systems biology: a brief overview. *Science* 295: 1662–1664.