Deep computing for the life sciences

by W. C. Swope

Information technology, from software applications to special high-speed computers, networks, and storage, has transformed the life sciences. This paper is a brief overview of some of the biological processes and concepts that are central to the life sciences. It also relates how deep computing plays a key role for the life sciences. The papers in this issue of the* IBM Systems Journal*, and a companion issue of the* IBM Journal of Research and Development*, provide an outline of the methods and technologies by which an understanding of the basic molecules of life can lead to the treatment of disease and other benefits to humankind.

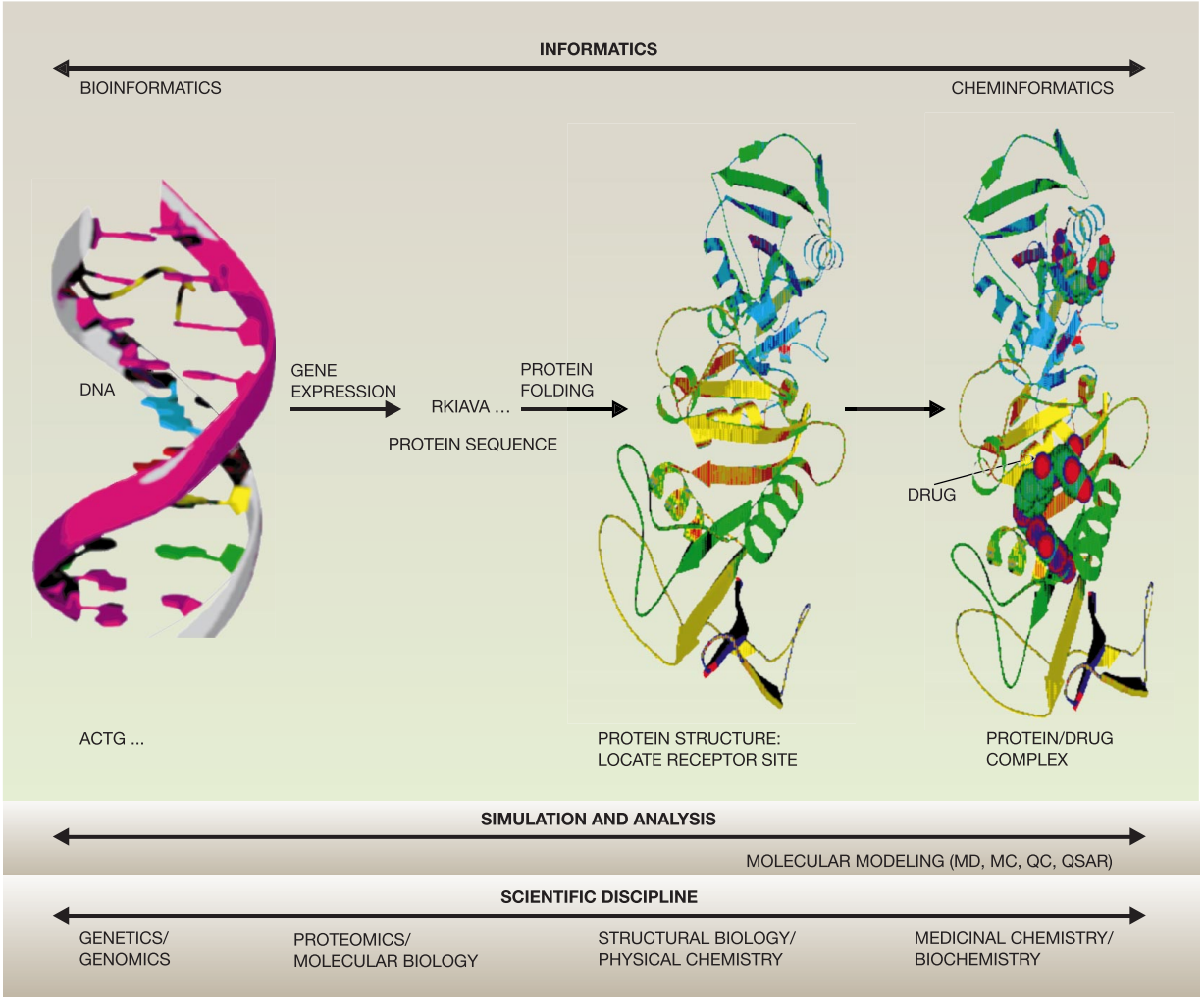
What makes us human? The human "genome" is the genetic material that characterizes every human being.[1] The genome consists of 23 pairs of chromosomes, each of which is composed of molecules of deoxyribonucleic acid (DNA). DNA molecules are shaped like a double-stranded helix that looks like a twisted ladder. Sugar and phosphate molecular components form the sides of the ladder, whereas pairs of nucleotide bases form the rungs. The nucleotide bases come in four main types: guanine, adenine, cytosine, and thymine, commonly abbreviated as G, A, C, and T. The base pairs that make up the rungs are either G–C, C–G, T–A, or A–T pairs. The human genome consists of approximately three billion base pairs, or rungs on the ladder. In the human body we have identical DNA in nearly every cell. Encoded in the DNA of the human genome is all the information that makes us human. It provides the information that shapes our bodies as embryos and controls biological function throughout our lives.

In late February of 2001, the journals *Science*[2] and *Nature*[3] both published cover articles about the human genome, to describe the significant human achievement of completing a first draft of the complete human DNA sequence (the Human Genome Project). Much work remains, of course, to analyze, interpret, and use this information. And the technology that is developed for these tasks will be applied to the genomes of other important species of plants and animals.

This paper provides first a brief overview of some of the biological processes and concepts that are central to understanding the life sciences. Then, within this context, the papers in this issue of the *IBM Systems Journal*, and a companion issue of the *IBM Journal of Research and Development*,[4] are described. The overview of life science processes and concepts is loosely organized around the chemistry and biology of the cell, starting with DNA and moving to proteins and drugs. (See Figure 1.[5]) The special role of agricultural life sciences is briefly mentioned, followed by discussion of the interplay between progress in life sciences and the issues of software applications, tools, infrastructure, middleware, data storage, data distribution, and data standards.

0018-8670/01/$5.00 © 2001 IBM

Figure 1 This figure illustrates gene expression, protein folding, and drug docking as biochemical processes. Also listed are some of the terms mentioned in the paper that relate to the disciplines engaged in these studies and the types of analysis software employed. Reading from left to right, we see that DNA holds the information used by the cell to manufacture proteins. The DNA codes for a sequence of amino acids that make up a protein sequence. This sequence is manufactured and folded into a unique three-dimensional shape that is intimately related to its function. Drugs interact with proteins and influence their functions, often by binding to a protein site in a way that inhibits that protein's ability to perform its normal task. The deep computing software techniques used to study these molecular systems include molecular dynamics (MD), Monte Carlo (MC), quantum chemistry (QC), and quantitative structure-activity relationships (QSAR).[5]



Genes and proteins

Every member of a biological species has almost identical DNA. But of the three billion base pairs that make up the human genome, only about 1 percent codes for proteins. The variation among different members of a species is caused by very minor differences in their DNA. For humans, that difference constitutes about one base pair per 1300, or about 0.1 percent of the 1 percent that codes for proteins.[2,3]

How much disk storage would it take to hold a representation of the genetic information of an individual? Since each rung of the DNA ladder has four pos-

sibilities, its state can be stored in two bits. Therefore, the three billion base pairs could be stored in approximately 750 megabytes (MB). If only the protein-encoding regions were stored, the resulting information could be stored in 7.5 MB. If only the information that characterizes deviation from some standard reference genome were stored, the result could be encoded in about 7.5 kilobytes (KB). Finally, and very importantly, we each actually have *two* copies of the human genome in our cells: one from our father and one from our mother, organized as pairs of chromosomes. This brings the storage required up to about 15 KB. Of course, the situation is more complex than this. For example, there are many diseases that are caused by variations in parts of the genome that do not code for proteins. But for personalized medicine, medical diagnostics, and personal health risk analyses, it might be possible to hold a fairly complete genetic description of an individual on a couple of floppy disks. (A floppy disk holds about a megabyte.)

The DNA in an organism encodes the information that the machinery of the cell uses to make *proteins*. In contrast to the DNA, which provides information content, proteins are the working molecules of an organism, carrying out most of the daily operations of a cell. They are involved in such things as motion, metabolism, cellular repair, cell-to-cell signaling, and cellular reproduction. Some even control which parts of the genome need to be activated at any particular time to make yet more and different proteins. Proteins are made from long chains of different combinations of 20 chemical building blocks called amino acids. The order and composition of the set of amino acids ultimately determines the three-dimensional shape and biological function of the protein.

Although the DNA in every cell of an individual's body is roughly identical, there are obvious differences between, say, nerve cells and liver cells. How is this possible if each cell has the same DNA? The answer lies in the fact that these cell types have different combinations of *proteins* in them. These proteins were built by *expressing* different *genes*—different parts of the same genome—in each of the cell types. Gene expression is the process of using the information encoded in the DNA to manufacture a protein from its amino acid components. The set of genes expressed not only varies from cell type to cell type, but also varies depending on the age of the cell and any stress or disease state it might be in.

Gene expression is a complex, multistep process involving an intermediate representation of the genetic information in the form of molecules called messenger RNA (ribonucleic acid) or mRNA. The process of copying information in the DNA into mRNA is called *transcription*. Experiments to capture and analyze the mRNA in a cell—and thus the genes being expressed—have become standardized using products called "gene expression arrays" from companies such as Affymetrix, Hyseq, and Incyte. The use of gene expression arrays is becoming common in many areas of biological research. In the future this technology will probably become a key component of medical diagnostics and even individualized medical treatments, which tune a medical treatment or drug protocol to the genetic makeup of an individual and his or her medical condition.

When genetic information is expressed, the mRNA produced is transferred to a cellular structure called a ribosome. The ribosome is a cellular manufacturing site where, in a process called *translation*, amino acid raw materials are assembled into proteins. The order of the amino acids in each protein is specified by the mRNA, through what is known as the "genetic code." As described above, the nucleotide base sequence along either side of the DNA ladder consists of bases denoted G, A, C, and T. These can be thought of as letters in a four-letter alphabet, which can form words, or *codons*, of three letters. There are 64 possible three-letter words from a four-letter alphabet. Each of these words corresponds, through the genetic code, to one of 20 amino acids. (Some of the amino acids relate to more than one three-letter word and some words act as punctuation.)

In the simplest view, we are told that each gene in the genome contains the information for a specific protein. However, it is much more complex than that. Most scientists now believe that many genes actually encode more than one protein. The choice of which proteins are made from a given gene is controlled by yet other proteins through processes that include post-translational modification and, to a lesser extent, post-transcriptional modification. Although commonly accepted figures for the number of genes in the human genome range from 25 000 to 40 000, many scientists now believe the number of proteins could be more than a million.

Add to this complexity these facts: proteins rely on other proteins to get themselves expressed; some proteins exist to aid in the folding of yet others; many proteins are actually dimers (protein pairs from the

same or different genes) or even multimers; and proteins often are parts of complex signaling pathways that work through very specific protein-to-protein interactions. As difficult as the genomic landscape is to describe, we see that the protein landscape is even more complex. The study of specific proteins, the chemical pathways in which they are involved, and their activities has always been a key area in the disciplines of biochemistry and molecular biology. But the study of the entire collection of proteins in the cell and the complexity of their interactions is a dynamic and evolving field now known as proteomics.

Drugs work by interacting with proteins that are, themselves, involved in some aspect of a disease or illness.

Proteins owe their function in large part to the special three-dimensional shape or *fold* they adopt in the cell. These shapes allow proteins to fit favorably together, for example, during protein-to-protein interactions, or to enable a molecule to fit into a protein's *active site*. In order for us to understand protein function, therefore, we would like to be able to obtain complete information about the relative locations of each of the atoms in a protein molecule. The two most common experimental techniques for obtaining atomic resolution information about protein structure are X-ray spectroscopy and nuclear magnetic resonance (NMR) spectroscopy. Currently, atomic-level structures are known for only about 15 000 proteins.[6] These structures include those of proteins from many organisms, not just humans. X-ray spectroscopy is very accurate, highly reproducible, and the method preferred by many scientists to generate detailed three-dimensional protein structural information. But the production of atomic coordinates by X-ray techniques can be very time consuming, sometimes taking as long as years or even decades per protein. X-ray experiments require high-quality crystals of the purified protein. Most of the time and effort is spent purifying and crystallizing the protein. Some proteins cannot be crystallized at all, and some only crystallize under very specific experimental conditions. Because they are difficult to crystallize, there are entire classes of very important proteins, such as membrane and fibrous proteins, that remain largely inaccessible to study by X-ray methods. The art of X-ray crystallography is often in finding the specific conditions of crystallization method, crystallizing agent, pH (a measure of acidity and alkalinity), temperature, and the concentrations of several solvents used in the process.

NMR spectroscopy provides an alternative way to determine detailed three-dimensional atomic structures of proteins. Liquid state NMR is not dependent on purified crystals of protein. However, the experiments can be difficult to interpret in unambiguous ways. Many proteins aggregate and precipitate at concentrations appropriate for this technique. Furthermore, even if the protein is sufficiently soluble, the NMR spectra of large proteins are often difficult or impossible to interpret.

Because large numbers of amino acid sequences for proteins will be available as an important product of the various genome projects, and because the experiments to determine full protein structures are so difficult to perform, two of the most intriguing scientific issues being addressed today are (1) whether we can *predict* the three-dimensional structure of a protein (*what*) from just its amino acid sequence, and (2) what is the process (*how*) by which proteins adopt the shapes they need to perform their functions. These two issues are often called *the protein folding problem*, and they constitute a body of "grand challenge" research that has absorbed scientists for about 50 years. But interest in this area has recently been stimulated by access to higher-performing computer systems, better protein models, more experimental data, and the enormous amount of sequence data from the Human Genome Project.

Drug development

One of the benefits we hope to get from the Human Genome Project is the enhanced ability to design cures for human disease.[7] Drugs work by interacting with proteins that are, themselves, involved in some aspect of a disease or illness. For example, a drug might inhibit the effect of a protein essential for the transmission of a virus, or enhance the effect of a protein that helps the cells of the immune system to fight disease. One of the key steps in the drug development process is the identification and selection of the protein that will be the *target* for the drug. This step generally involves a very thorough molecular understanding of the *disease pathway*, i.e., the

biochemical processes and molecules involved in the causes and symptoms of the disease. From among the proteins involved in a pathway, a drug target is selected and characterized. The target must not only be on the critical part of a disease pathway, it must be able to interact with or bind to a drug molecule. It is in these steps (target identification, selection, and validation) that genomics, proteomics, gene expression analysis, and all the associated data management technologies are likely to have the largest impact.

Once one or more drug targets are selected and characterized, the next stages in the drug development process are known as *lead identification* and *lead optimization*. During these stages many molecules are *screened* to determine which ones best interact with one or more potential targets. Molecules that show a sufficient degree of interaction with the target are known as *lead compounds*, or, simply, leads. Most pharmaceutical and biotech companies have synthesized, purified, and characterized a large number—perhaps millions—of molecules over the years of the companies' existence. This body of physical material is usually referred to as a *compound library*. The process of screening often involves retrieving a small amount of material from the compound library and testing it against the targets under consideration. The test is called an *assay* and the result of the test is usually an affinity measure, related to the ability of the compound to bind to the target. Good leads should not only show high affinity for the target, they should be *selective*, interacting with only a single type of protein. Compounds that interact with multiple target proteins generally produce numerous and undesirable side effects, by interfering with multiple biochemical pathways.

Experimental technologies. During the last ten years, two important experimental technologies have transformed the lead identification and optimization process. These technologies make heavy use of computational techniques and robotics. The first is *combinatorial chemistry*, the automated synthesis of tens of thousands of prospective drug molecules using robotics. For example, it is quite possible now to synthesize, for testing, all of the molecules that can be made by chemically bonding together chains of 5 reagents selected from a set of 20 possibilities. This could represent $20^5 = 3\,200\,000$ compounds. The reagents could be the 20 amino acids if the drug company were investigating *peptide therapeutics*.

The second new technology is known as *high-throughput screening* (HTS). This also uses robotics and is often combined with combinatorial chemistry. HTS is the automated testing of each of the compounds in a library against one or more targets. Current technology can produce 10 000 to 20 000 assay results per day. Companies such as Aurora Biosciences produce equipment capable of 100 000 assay results per day, a level known as "ultra-high-throughput screening" (UHTS). Of course, these data need to be systematically stored, organized for efficient retrieval, and analyzed. At these rates, computational support for the HTS process is essential and ranges from data collection and management to sophisticated data mining and decision support.

Computational approaches. There are several important computationally oriented approaches used in the lead identification and optimization stages to design drugs today, including rational or structure-based drug design, virtual high-throughput screening, and combinatorial library design.

A traditional use of computational chemistry in drug design is in structure-based drug design, also known as *rational drug design*. In these approaches, usually one starts with a characterized protein target, perhaps even a molecule bound to its active site. This characterization typically includes the three-dimensional structure of the protein, determined perhaps by X-ray, NMR, or computational methods. Through the use of numerically intensive methods that range from purely classical (such as molecular dynamics and Monte Carlo) to fully quantum mechanical, chemists attempt to design a molecule that binds strongly to the active site and is selective against that drug target.

Another application of computational chemistry is virtual high-throughput screening (VHTS), in which libraries of compounds are characterized using computer models that predict, in some way, a measure related to binding affinity to a particular target. These techniques make heavy use of numerically intensive computation and can also be database-intensive.

One of the important new areas of computational chemistry is in the development of techniques to streamline the search for lead compounds through the design of efficient screening libraries. The number of molecules that can be made by combinatorial chemistry is enormous, and only a very few are useful as drug leads. It is, therefore, important for eco-

nomic reasons to quickly focus in on the compounds that are potential leads. There have recently been developed methods for measuring the *chemical diversity* of a compound library. Libraries can be designed that are either diverse or focused in terms of some set of chemical attributes. Libraries can be designed that balance chemical diversity against such characteristics as library size and cost of synthesis. Usually, an iterative process is followed, where one first searches for the best leads using HTS with a *diverse* compound library. Then, more *focused* libraries are designed of molecules with properties that are similar to the best ones found during the previous screening.

Agricultural life sciences

Recently, with the successes of the Human Genome Project in such high profile, it has been easy to overlook the fact that the impact of the life sciences extends deeply into agriculture as well. The life sciences tools and techniques, from DNA sequencing and gene expression to combinatorial chemistry and high-throughput screening, are playing important roles in agriculture as well. Although there is some political controversy today, we should eventually expect to see safe new varieties of plants that are more nutritious, better tasting, and resistant to spoilage in the field and in the store. Transgenic plants will be developed that cheaply produce pharmaceuticals and plastics. Plants that produce chemicals that make them immune to insect pests will decrease our reliance on dangerous insecticides and decrease harm to the environment. Using the same technology that has been developed for drug design, we are capable of developing more selective, more effective, and safer herbicides and insecticides.

Sequencing of important plant genomes is well underway. Considerable progress has been made with rice, corn, and soybeans. An especially important plant genome effort is that of the *Arabidopsis thaliana*. This is an important model organism to plant biologists, much like the mouse is to human biologists, because of its rapid life cycle. The Arabidopsis Genome Initiative (AGI) is an international collaboration to sequence and annotate the Arabidopsis genome. Although when it began, in 1996, the goal was to complete the sequencing of the DNA by 2004, the project was completed at the end of 2000.[8]

Application software, data storage, middleware, and infrastructure

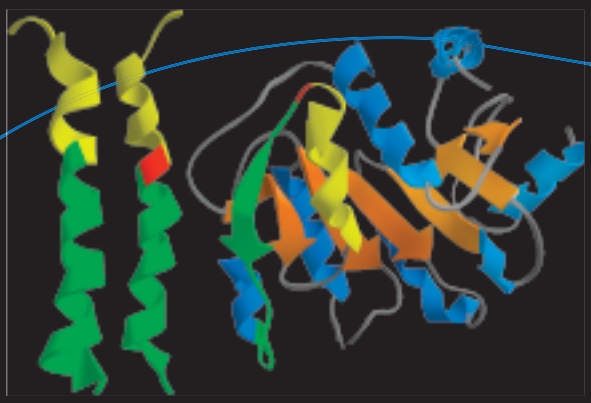
A critical factor in the rapid emergence of life sciences is not only the availability of high-performance computational capability, but the presence of the Internet for nearly immediate dissemination of high volumes of textual and numerical information. On-line journals and on-line data sources are now common. For example, the January 1, 2001, issue of the on-line journal *Nucleic Acids Research*[9] has about one hundred articles, each describing a different *database*, almost all of which are available through the Internet.

Another factor important to the life sciences is the availability of inexpensive disk storage and efficient database software that is critical for the storage of the massive volumes of data being produced. These include not only textual data, but also raw and processed data in the form of images, DNA and protein sequences, gene expression data, chemical structure and properties data, and HTS data.

Finally, for tying all this data and analysis capability together we are seeing the emergence of standards for data representation and interchange. Examples of this include mmCIF (macromolecular Crystallographic Information File), XML (Extensible Markup Language), and the Object Management Group's (OMG) Life Sciences Research Domain Task Force to establish CORBA** (Common Object Request Broker Architecture**) as the standard for interoperable software components. We are also seeing the emergence from a number of software vendors of application frameworks, software environments, and middleware that enable the rapid development of new analytical protocols driven, in turn, by the need to perform novel analyses on data integrated from a variety of sources, both local and remote.

Modern computational techniques
are critical in our approach to understanding and treating specific human diseases.

See pages 254 and 255.



Cystic fibrosis

The story of the genetic disease cystic fibrosis (CF) is related to many of the technologies described in this issue of the *IBM Systems Journal*.[1-5] CF is one of the most common fatal genetic diseases in developed countries, affecting about 30 000 people in the U.S. and 25 000 in Europe. Affected individuals have abnormally thick lung secretions and frequently succumb to respiratory infections.

This disease is caused by a defect in a gene on chromosome 7 that codes for a protein known as CFTR, which consists of 1480 amino acids. The CFTR gene was identified in 1989 by a large group of collaborators. The CFTR protein plays a critical role in epithelial cells, which often function in the secretion of mucous and other fluids. CFTR is a *transmembrane* protein, embedded in the cell membrane, where it mediates the transport of chloride ions and water molecules. The most common defect to the CFTR gene is the omission of a single phenylalanine amino acid at position 508 in the CFTR protein. It is currently thought that the missing amino acid results in a *misfolding* of the CFTR protein, at which point it is degraded by normal cellular housekeeping. As a result, no CFTR protein finds its way to the cell membrane, disrupting normal cell function.

A cure for CF rests partially on the full characterization of the structure, folding mechanism, and function of the CFTR protein. However, transmembrane proteins such as CFTR are nearly impossible to crystallize, preventing X-ray crystallography experiments that would reveal its 3-D structure. Structure-prediction methods based on similarity of a protein's sequence to known structures are only partially revealing. A structure for a 214-amino-acid subregion of CFTR, known as NBD1 (nucleotide binding domain one), based on purely computational approaches, was reported in 1996. This domain includes amino acid 508, which is missing in most CF patients. The rightmost ribbon diagram on this page is a skeletal depiction of the subregion's structure.

Nuclear magnetic resonance (NMR) techniques can be used to determine detailed structural information for short peptide sequences. Examples of NMR-determined CFTR peptide structures are also shown; they are two short peptides of 25 and 26 amino acids. The larger one includes, and the smaller one omits, the phenylalanine that is missing from position 508 in CF patients. The red regions in the peptide and in the theoretical NBD1 structure represent this amino acid. Corresponding regions before and after position 508 are shown in yellow (498 to 507) and green (509 to 523). These NMR structures are somewhat at odds with the theoretical structure shown for NBD1. However, the surrounding material in the larger domain may serve to stabilize different structures than are exhibited by short peptides in solution. Closing the gap between theoretical and actual structure will be the first step to understanding and conquering CF.

In February 2001, the Cystic Fibrosis Foundation announced a collaboration with Structural GenomiX involving $11 million to fund a five-year project to produce a full 3-D structure for the CFTR protein. It is hoped that this structural information will be useful to developing an understanding of how the protein works and in the design of new treatments.

AIDS

The story of the infectious disease AIDS is also related to many of the technologies described in this issue.[6-9] AIDS (acquired immunodeficiency syndrome) is one of the most devastating diseases in modern human existence. It is caused by the human immunodeficiency virus (HIV). HIV weakens the immune system, making the body unable to resist ordinarily benign infections. HIV, like most viruses, takes advantage of the cellular machinery of the cells it infects to make the proteins needed for its replication. These proteins are coded in the genetic material of the invading virus. Several HIV proteins are synthesized in one continuous chain, known as a polyprotein. The polyproteins are then cleaved into smaller chains that are reassembled into new viruses. The cleavage of the polyprotein is performed by a viral enzyme known as a protease.

In 1988, the amino acid sequence of HIV-1 protease and its role in the replication of the virus was established. The X-ray crystallographic structure of HIV-1 protease was determined in 1989, yielding its 3-D shape. The HIV-1 protease structure is shown on this page. HIV-1 protease is itself a symmetric protein, a dimer composed of two identical strands.

The HIV-1 protease works by binding to the HIV polyprotein and cleaving it at specific places. Using the tools of rational drug design, a number of drugs have been designed that bind strongly to the active site in HIV-1 protease, thereby interfering with its ability to cleave the polyprotein. Such drugs are known as HIV-1 protease inhibitors. Protease inhibitors have proved to be one of the most effective classes of drugs for the treatment of HIV infections. A second highly effective class of drugs is known collectively as reverse transcriptase inhibitors (AZT is an important example). Drugs from each of the two classes are often delivered together in the form of combination, or cocktail, therapy.

In rational drug design, chemists make extensive use of computers along with graphical representations of the 3-D molecular structure of the protease, as determined by X-ray crystallographic experiments, to design molecules that bind more strongly to the active site than the polyprotein. Not only is the shape of the protease used in this process, but also computational predictions of properties such as solubility and toxicity of the candidate molecules. Dozens of HIV-1 protease inhibitors have been designed using these techniques, including saquinavir (Invirase, from Hoffman-LaRoche, shown here bound to HIV-1 protease), indinavir (Crixivan, from Merck), nelfinavir (Viracept, from Agouron), and ritonavir (Norvir, from Abbott). With these kinds of experimental and computational advances, we may some day soon have HIV under control as a treatable disease.

William C. Swope

Cited references and notes

1. For information about the Protein Data Bank, from which the molecular diagrams shown on the cover and on these two pages were developed, see: H. M. Berman et al., "The Protein Data Bank," *Nucleic Acids Research* **28**, 235–242 (2000).

2. PDB ID: 1NBD. F. J. Hoedemaeker et al., "A Model for the Nucleotide-Binding Domains of ABC Transporters Based on the Large Domain of Aspartate Aminotransferase," *Proteins: Structure, Function, and Genetics* **30**, 275–286 (1998).

3. PDB IDs: 1CKW, 1CKX, 1CKY, and 1CKZ. M. A. Massiah et al., "Cystic Fibrosis Transmembrane Conductance Regulator: Solution Structures of Peptides Based on the Phe508 Region, the Most Common Site of Disease-Causing F508 Mutation," *Biochemistry* **38**, No. 23, 7453–7461 (1999).

4. D. M. Orenstein, *Cystic Fibrosis, A Guide for Patient and Family*, Second Edition, Lippincott-Raven Publishers, Philadelphia, PA (1997).

5. M. J. Welsh and A. E. Smith, "Cystic Fibrosis," *Scientific American* **273**, 52–59 (1995).

6. PDB ID: 1HXB. A. Krohn et al., "Novel Binding Mode of Highly Potent HIV-Proteinase Inhibitors Incorporating the (R)-Hydroxyethylamine Isostere," *Journal of Medicinal Chem*istry **34**, No. 11, 3340–3342 (1991).

7. P. Y. Lam et al., "Rational Design of Potent, Bioavailable, Nonpeptide Cyclic Ureas as HIV Protease Inhibitors," *Science* **263**, 380–384 (1994).

8. The National Cancer Institute maintains an HIV protease database available over the Internet at the Web site http://www.ncifcrf.gov/CRYS/HIVdb.

9. Some of the material on HIV-1 protease was adopted from the highly informative Web site on rational drug design by D. Eric Walters of the Chicago Medical School: http://www.finchcms.edu/biochem/walters/walters_lect/walters_lect.html.

The big computational challenges

As just described, the molecular-level approach to the understanding of the life sciences presents a long and complex path from DNA to the promised social impact and benefits. This path can be followed only through very sophisticated computational techniques. In this issue of the *IBM Systems Journal*, two papers stand out as being broadly applicable and potentially significant from a scientific perspective. The first, by ■ Head-Gordon and Wooley, is "Computational Challenges in Structural and Functional Genomics." This paper outlines some of the research that still needs to be done for the world to move forward and capitalize on the incredibly valuable information being derived from the Human Genome Project and other genome projects. It ties together work in topics as diverse as functional genomics, analysis of data from gene expression arrays, protein folding, combinatorial chemistry library design, and high-throughput X-ray structure determination and drug screening. The paper gives a wonderful perspective on the interrelation of many scientific areas in life sciences and it could almost serve as a blueprint for the public funding of life science research at a national level.

The second paper that almost defies classification is not a scientific one, but a software infrastructure and tools paper, by ■ Haas et al., "DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources." DiscoveryLink is a *middleware* software product from IBM. Middleware is software that is invoked by scientific application software (the upper layer), to simplify the access to data in one or more data sources (the lower layer). The DiscoveryLink software can be used to build a "federated" database system, simplifying access to a range of local and remote data sources. It enables scientific programmers to more easily build the next generation of complex applications that will permit concurrent scientific analyses of data in many different sources. This software has already been applied in life sciences settings where data from biological and chemical experiments were combined and queried to support experiments with new approaches to drug design. The DiscoveryLink middleware could be applied to problems in almost any domain of the life sciences to facilitate the development of applications that use data mining and analysis of data spanning multiple sources.

DNA sequence analysis and classification

Given the recent completion of the *first draft* of the human genome, it is hardly surprising that there are several papers in this issue, and the companion issue of the *IBM Journal of Research and Development*, that address the fields of DNA analysis in one form or another. One of the most popular questions of late is "How many genes are in the human genome?" This has been a controversial issue and the reasons for that are very interesting. As described earlier, genes encode the information that the body uses to manufacture proteins. The problem with counting genes is finding them within the three billion base pairs that encode proteins. This is harder than it sounds, since only about 1 percent of the human genome codes for protein. Between the regions that code for protein are long stretches of DNA commonly referred to as "junk" DNA, because we do not know what most of it is used for. It is not always easy to find the protein encoding regions in the junk. Also, the encoding regions of a gene are usually not continuous. Interspersed between the protein encoding regions (exons) are regions that do not code for protein (introns).

A paper in the companion issue, by ■ Davison et al., "Brute Force Estimation of the Number of Human Genes Using EST Clustering as a Measure," describes their estimate of the number of genes. Their procedure produces a number that is higher than is generally accepted by the community, but this only serves to highlight the controversy around the issue. Their prediction places the number of genes at around 130000, whereas commonly quoted figures today are in the 25000 to 40000 range. However, it is known that many genes produce more than one type of protein through a process known as alternative splicing. The expressed sequence tags (ESTs) used in Davison's estimate reflect the alternative splicing, so the apparent discrepancy may simply imply that on average each gene produces three or four ESTs through alternative splicing.

One class of analysis procedures used for finding the coding regions within a DNA sequence uses a mathematical formalism known as hidden Markov models. A paper in the companion issue by ■ Birney, "Hidden Markov Models in Biological Sequence Analysis," provides a brief review of this approach and is quite useful for its pointers into the current literature.

A large variety of genomic data sources have emerged. Many of these grew from independent studies that were organism-based, such as for *Escherichia coli*, or disease-based, such as for cancer research. Much of this information is publicly available over the Internet. Comparison and unification of data in these multiple sources is critical for much of the sequence analysis that remains to be done. However, the fact that these data sources were developed for different purposes, by different researchers using different methods, often makes the data difficult to unify. Tools such as IBM's DiscoveryLink can help with the mechanical integration of data sources. However, there are often issues that need to be addressed that have to do with making sense of data that are described with different terminology by the different data sources. The paper by ■ Goble et al., "Transparent Access to Multiple Bioinformatics Information Sources," addresses this issue. The paper describes the TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) project, that addresses many problems with the effective use of multiple independently developed data sources, especially the attendant use of unclear or inconsistent terminology and concepts across the sources. The project has many interesting components, including an example of a biological domain-specific ontology and its consistent use through a terminology server. This ontology is used to represent a global schema against which source-independent queries can be made. These queries are then re-expressed into source-dependent ones, which can be executed by a software layer that communicates with multiple sources. It should be clear that the problem that TAMBIS addresses is not limited to bioinformatics, and that the technology could be broadly applicable.

A number of papers describe software technologies that allow one to combine data in multiple sources in order to perform analyses. One important example is described in the paper by ■ Davidson et al., "K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources." K2 is data integration technology; GUS is a warehouse approach and tools to download, clean, integrate, and annotate data. The approach described is applied to genomic analysis, but one can easily see its generalizability to other life sciences areas.

With the emergence of multiple heterogeneous remote and local data sources and the need for complex analyses of them, several software environments and application frameworks have evolved to facilitate the understanding of genomic data. These generally combine graphical interfaces and complex analytical and mining tools with Web-based access to one or more remote data sources on the Internet. Three papers from the National Center for Genome Resources (NCGR) illustrate a few approaches to this kind of capability. One key point of these papers and the approaches they describe, is that the field requires *not just the integration of data* sources, but the *integration of multiple application tools* with those data sources. See the papers by ■ Inman et al., "A High-Throughput Distributed DNA Sequence Analysis and Database System," and by ■ Siepel et al., "An Integration Platform for Heterogeneous Bioinformatics Software Components."

Fortunately, nature has reused in the higher organisms what it has learned from the design of simple organisms, such as bacteria.

The third paper from NCGR, by ■ Mangalam et al., is "GeneX: An Open Source Gene Expression Database and Integrated Tool Set." This paper describes a system that helps with the storage, organized retrieval, and analysis of gene expression data. Workers in this field are often frustrated by the fact that each of the several gene expression *technologies* comes with its own unique data representation, storage, retrieval, and analysis systems. Furthermore, there are many *species-specific* public repositories for gene expression data that also have their own unique data and analysis systems. Among the notable features of the GeneX system is that it supports multiple species and gene expression technologies. It is a Web-based system that makes heavy use of emerging standards such as MAML (MicroArray Markup Language), based on XML. The data model for gene expression and associated data storage, instantiated in GeneX with a relational database management system, includes extensive support for annotation and meta-data about the conditions under which the gene expression experiments were performed. For data organization and storage, an important part of the GeneX system is the Curation Tool, which supports the use of a controlled vocabulary, essential if one wants to meaningfully combine data from different organisms and gene expression technologies. And

GeneX includes not just support for data storage and retrieval, but an interface to Perl for building it into new applications, and integrated statistical analysis and visualization capabilities.

Protein sequence analysis and classification

Among the most important software tools for understanding DNA and protein sequences are sequence similarity and alignment tools such as BLAST.[10] These tools allow one to compare an unknown sequence with a database of sequences from other organisms that are better understood. Fortunately, nature has reused in the higher organisms what it has learned from the design of simple organisms, such as bacteria. Scientists have been studying the bacteria *E. coli*, for example, for many years and have mapped out most of its biochemical pathways, the proteins involved, and the genes that were expressed to make those proteins. So, we can often deduce the function of an uncharacterized human gene or protein by searching for the most similar genes or proteins in the databases of bacterial genes. The BLAST search procedure was built to do this kind of search. A paper in the companion issue by ■ Floratos et al., "DELPHI: A Pattern-Based Method for Detecting Sequence Similarity," provides an alternative approach that might prove to be better, especially for weak but biologically important similarities. Their approach works by building a set of patterns obtained from the underlying database through a one-time computation. These patterns are subsequently searched when a query sequence is presented to the system.

Another potentially useful protein searching and comparison technique is described in the paper by ■ Robson, "Fastfinger: A Study into the Use of Compressed Residue Pair Separation Matrices for Protein Sequence Comparison." This paper describes a way to characterize a protein sequence that attempts to capture a signature of the functional and structural domains that the sequence might contain. This characterization is easily computed and is a regular data structure that might also allow support of rapid query of a database of sequences.

The paper by ■ Wang et al., "New Techniques for Extracting Features from Protein Sequences," provides a description of some techniques developed and used at the Protein Information Resource (PIR) to extract features from protein sequences and use them to help classify proteins into functional hierarchies. Their method identifies and uses protein features as inputs for Bayesian neural networks trained for protein sequence classification. The paper evaluates the performance of the approach by comparing it with other protein classifiers that are based on sequence alignment and machine learning methods.

Another paper that conveys a method for protein classification is by ■ Liu and Califano, "Functional Classification of Proteins by Pattern Discovery and Top-Down Clustering of Primary Sequences." This paper introduces an unsupervised clustering technique that determines functionally related clusters of proteins and their functional motifs by coupling a pattern discovery algorithm, a statistical framework for the analysis of discovered patterns, and a motif refinement method based on hidden Markov models.

An interesting project is described in the paper by ■ Lee and Irizarry, "The GeneMine System for Genome/Proteome Annotation and Collaborative Data Mining." This paper describes GeneMine, a Web-based system into which the user submits a gene or protein sequence. The system then submits this sequence to a number of free sequence analysis servers on the Internet and collates the results. Often, from seeing a consensus of opinions from the various servers, one can develop an annotation and a degree of confidence for the possible function of a sequence. This is a very interesting model, particularly given the incomplete reliability of some of the Web-based data sources. But what might evolve in the future when everyone has access to scientific analysis portals of this sort, that provide essentially free computing, analysis, and high-speed data source access?

A very interesting paper related to protein structure is by ■ Jurisica et al., "Intelligent Decision Support for Protein Crystal Growth." This paper describes a very powerful application of case-based reasoning tools to the scientific decision process involved with the production of protein crystals to feed high-throughput protein structure determination by X-ray methods. Case-based reasoning tools were developed to attempt to capture the thinking of experts. The software system described in the paper takes negative and positive results of protein crystallization attempts and "learns" from them to help guide and automate crystallization for high-throughput protein structure determination.

Protein simulation, folding, and structure prediction

Several papers review progress on the protein folding problem, provide descriptions of projects, or describe novel approaches. A very thorough review of the field of computational protein folding is provided by ■ Duan and Kollman, "Computational Protein Folding: From Lattice to All-Atom." These authors should be recognized for the work they reported in 1998 in what has become a very famous *Science* paper that describes a computer simulation, by molecular dynamics, of one microsecond during the folding of a small peptide.[11] The calculation was one of the most ambitious of its type ever performed, and it has served to stimulate the field; a number of research groups around the world are now trying to perform bigger, longer, and more accurate simulations, all to a great extent inspired by the one-microsecond simulation described by Duan and Kollman in 1998.

IBM Research has its own effort to study the protein folding process. This project, known as Blue Gene, was announced in December of 1999, and has as one of its most exciting aspects the construction of a petaflop (10^{15} floating-point operations per second) computer. This project, including the scientific goals, critical aspects of the application software, and the hardware effort are described in the paper by ■ Allen et al., "Blue Gene: A Vision for Protein Science Using a Petaflop Supercomputer." Nearly every aspect of the Blue Gene project is ambitious. The computing capacity of the hardware is approximately ten times the aggregate of the top 500 supercomputers in the world today. The application software that will perform molecular dynamics simulations will have to be designed very carefully to exploit this massively parallel hardware architecture effectively. But among the most important challenges in a project of this sort is to engage the scientific community in academia and in government and commercial labs to ensure that scientifically meaningful research is carried out.

One of the many challenges described in the paper about IBM's Blue Gene project is highlighted in the paper by ■ Straatsma and McCammon, "Load Balancing of Molecular Dynamics Simulation with NWChem." This paper describes some of the different ways of parallelizing molecular dynamics applications for IBM SP* parallel computers and the challenges to be faced in order to obtain efficient use of the hardware. The key issues involve not only processor speed, but also interprocessor communication speed and latency.

There are a number of papers that are, perhaps, less concerned with the *process* of protein folding, and more concerned with the *prediction of the final folded state*. The first of these, by ■ Eastwood et al., "Evaluating Protein Structure-Prediction Schemes Using Energy Landscape Theory," is in the companion issue. This paper comes from some of the top research groups working in the field of protein folding theory. It proposes that the (often simplified) energy functions that are developed to be used for predicting protein structure should be evaluated using the same properties that are used to characterize potentials in what is called "Energy Landscape Theory." This theory was developed to characterize and explain experimentally observed folding phenomena and has been very successful at providing a popular "funnel" picture of the folding process. The paper works from the premise that we think nature's energy function does not exhibit, for example, deep minima that would stabilize non-native conformations. This is because, if it did, proteins could get "trapped" and would not be able to reach their native structure. Since this does not happen, we know such minima do not occur and so hypothesized energy functions that exhibit them should be discarded as unrealistic. The authors' ideas are actually applied in the paper to determine an appropriate set of parameters for one particular choice of functional form for the energy.

Another paper, also from a very highly respected research group that has made considerable contributions to the field of protein structure prediction, by ■ Fain et al., "Determination of Optimal Chebyshev-Expanded Hydrophobic Discrimination Function for Globular Proteins," is in the companion issue. This paper describes the design of a potential, or scoring, function that can be used to evaluate predictions of three-dimensional models of proteins.

A paper in the companion issue by ■ Kumar et al., "Protein Flexibility and Electrostatic Interactions," discusses protein flexibility. Protein flexibility is often a key aspect of protein-to-protein interactions and protein-to-drug interactions. This is because the binding is often associated with protein conformational changes. This paper explores the role of protein electrostatic interactions in determining protein flexibility.

Another paper in the companion issue from the same research group is by ■ Tsai et al., "A Hierarchical, Building-Block-Based Computational Scheme for Protein Structure Prediction." The paper describes a procedure for predicting protein structure that is based on the identification of small, locally rigid and stable domains in the proteins. Once identified, by analysis of a database of experimentally determined protein structures, these domains or building blocks can then be used to predict structure as well as purported folding pathways, by assuming that the most stable building blocks form early during the folding process.

A very entertaining paper on protein structure prediction, by ■ Siew and Fischer, "Convergent Evolution of Protein Structure Prediction and Computer Chess Tournaments: CASP, Kasparov, and CAFASP," is included here. This paper describes similarities between computer-human chess championships and the competition for the critical assessment of protein structure prediction methods (CASP). We would like to use these competitions as a way to chart our advances in computer technology. However, one of the points of the paper is that we are simultaneously measuring advances in computer technology *and* our ability to exploit it through both better software and a deeper understanding of the phenomena. The authors make a case for a fully-automated version of the CASP competition, known as CAFASP, where no human intervention is employed to predict structure. This way hardware and software alone are pitted against experiment without the aid of a handful of expert overseers. This makes sense; only in a highly automated way will we be able to get structures for the millions of proteins for which sequences will eventually be available from the Human Genome Project and other genome projects.

Drug design

The most accurate way to study molecular systems such as drug molecules and drug-protein complexes would be through the use of quantum mechanical methods. However, high-quality quantum-based methods are very expensive and usually limited to molecular systems of a few dozen atoms or smaller. Out of necessity, due to the fact that most interesting biological systems have at least hundreds of atoms, most studies of biological systems use classical techniques to *approximate* the interatomic interactions. This approximation comes at a considerable cost in that it is practically impossible to study chemical reactions that occur in biological systems. Furthermore, even when reactions are not of specific interest, there are other important phenomena that cannot be addressed by classical methods, such as the electronic effects of charge transfer and polarization. Considerable effort has been spent on the development of hybrid approaches, and two papers bring us up to date on the work of two groups working in this area.

A paper in the companion issue by ■ Morokuma et al., "Model Studies of the Structures, Reactivities, and Reaction Mechanisms of Metalloenzymes," reviews a body of work in the application of quantum chemical techniques to biological problems, specifically those related to metalloenzymes (methane mono-oxygenase and ribonucleotide reductase, as examples). The paper also presents a detailed description and application of the ONIOM approach, which is a systematic way to combine in the same calculation computational approaches with different degrees of approximation. This approach can be used, for example, to combine classical molecular mechanics approaches with higher quality quantum chemical-based methods. It is likely to be one of the handful of methods exploited in the future to address complex but critical chemical issues such as enzyme- and solvent-induced chemical reactions.

A paper in the companion issue by ■ Andreoni et al., "DFT-Based Molecular Dynamics as a New Tool for Computational Biology: First Applications and Perspective," gives a brief review of some of the experiences the authors have had with the application of density functional theory (DFT) methods to problems of biological interest. This includes the use of molecular dynamics driven by DFT gradients as well as new approaches that include a partitioning of the system under study into classical and quantum mechanical domains. This kind of hybrid classical-quantum partitioning has been explored extensively with other quantum chemical methods, but the use of this kind of technique with plane-wave-based DFT methods represents an important step forward in the ability to address biological problems.

A paper in the companion issue by ■ Huang et al., "Quantum Crystallography, a Developing Area of Computational Chemistry Extending to Macromolecules," discusses the experiences the authors have had with their method, which uses data from X-ray spectroscopy to generate electron densities that also have a quantum mechanical basis. The resulting densities can be used to compute electrostatic fields or to assign partial atomic charges (to use, for exam-

ple, in classical simulations). The method also can build up approximate electron densities for large molecular systems from consideration of molecular fragments.

The paper included here by ■ Waszkowycz et al., "Large-Scale Virtual Screening for Discovering Leads in the Postgenomic Era," gives a very nice perspective on the computational methods that will be useful in the high-throughput environment of today's drug companies. It describes the development and use in an important case study of a computational method for virtual screening of compounds for drug design.

A paper in the companion issue by ■ Agrafiotis, "Multiobjective Optimization of Combinatorial Libraries," discusses how to take several factors into consideration in the design of combinatorial chemistry libraries. Combinatorial chemistry techniques have become a "workhorse" technology in modern drug design and are the key component of many high-throughput screening programs. As described earlier, the technique allows the synthesis of hundreds of thousands of molecules simultaneously by robotic techniques. Although millions of compounds can be made, library design as described in the paper addresses the important issue of *which* million should be made, out of the nearly infinite number that are possible.

Another paper in the companion issue, by ■ Platt et al., "QSAR in Grossly Underdetermined Systems: Opportunities and Issues," describes regression analysis as a commonly applied tool in the field of drug design, where it often appears as the underlying mathematical formalism in quantitative structure-activity relationships (QSAR) to relate molecular structure with biochemical activity. But these techniques can be applied much more broadly than just to drug design. Other areas of application can be as diverse as gene expression analysis and prediction of protein folding rates. This paper discusses new relationships between QSAR and principal components analysis, the application of error estimation, and the potential for other new applications of QSAR-related methods.

Cellular structures

The companion issue also has two papers on the study of cellular structures. These typically elude simulation techniques because they are too big to model atomistically, but too small to treat with continuum classical methods.

The paper by ■ Ayton et al., "Interfacing Molecular Dynamics with Continuum Dynamics in Computer Simulation: Toward an Application to Biological Membranes," describes a method that can be used in simulations of systems that exhibit behaviors that span very disparate time and space scales. The example application of the method is to biological membranes, which exhibit both microscopic behavior (in length and time scales related to their thickness), and macroscopic behavior (in length and time scales related to movement perpendicular to their surface). The method uses a feedback approach to couple behavior between the disparate scales. Methods like this one could be very useful in characterizing and simulating membranes and other important mechanical biological structures such as cilia and muscle motion.

The paper by ■ Baker et al., "The Adaptive Multilevel Finite Element Solution of the Poisson-Boltzmann Equation on Massively Parallel Computers," also describes a significant attempt to bridge microscopic and macroscopic simulation of structures of cellular dimensions. It describes a new algorithm and implementation for solving the Poisson-Boltzmann equations for the electric field around a charge distribution that is embedded in a continuum solvent with a distribution of counter ions. This method takes advantage of an adaptive mesh, as is commonly used in finite-element mechanical simulations, and it is particularly efficient and separable into parallel tasks for very large systems. The method is applied to the simulation of a microtubule, which is a hollow cylindrical structure in cells. Microtubles play a role in cell structure, motility, material transport, and division. These structures are highly electrically charged and the electric field around them is believed to be key to their functioning. The microtubule under consideration was 40 nanometers long and consisted of over 600000 atoms. The electric field around this structure has been computed and shows very interesting structure. It is important to note that this kind of approach has applicability outside the realm of biology; it is also useful for studying phenomena on a scale that is of interest in nanotechnology. This work was performed using a large IBM SP2 installed at the Supercomputer Center at the University of California at San Diego.

Organ simulation

The paper here by ■ Winslow et al., "Mapping, Modeling, and Visual Exploration of Structure-Function Relationships in the Heart," illustrates an *integra-*

tive model of cardiac function that requires treatment of several levels of detail in order to work. The levels of detail range from the treatment of subcellular to tissue behavior. This is necessary because the complex interactions of all the system components contribute to cardiac function. The work combines complex system modeling with medical imaging, computer graphics, and parallel processing.

Acknowledgment

The author would like to thank Jed W. Pitera for a careful reading of the manuscript and many helpful suggestions. Thanks go to Julia Rice for the original version of the figure that illustrates the relationships between DNA, proteins, drug molecules, and informatics. Thanks also go to Joan M. Zimmerman for a critical reading of the sidebar pages that describe the cover art. Finally, the author wishes to acknowledge the IBM Journals staff for many helpful suggestions and for their creativity and patience during the making of these journal issues.

Cited references and notes

1. For a wonderfully readable and informative account of the human genome, see M. Ridley, *Genome*, HarperCollins Publishers, New York (1999).
2. Special issue on the human genome, *Science* **291**, No. 5507 (February 16, 2001).
3. Special issue on the human genome, *Nature* **409**, No. 6822 (February 15, 2001).
4. *IBM Journal of Research and Development* **45**, Nos. 3 and 4 (2001), whole issue. This issue will be available after June 2001 from IBM as G322-0227 (see ordering information on the inside back cover of the *IBM Systems Journal*) and viewable on the Web at http://www.research.ibm.com/journal/.
5. On the left side of the figure is a representation of DNA, obtained from http://www.mol.biol.ethz.ch/wuthrich/software/molmol/gallery.html. The relevant figure has the following caption: "Schematic picture of the DNA complex of the Antennapedia homeodomain (image done in cooperation with M. Billeter, structure solved by Y. Q. Qian et al.)." The data were rendered using the MOLMOL and POV-Ray™ visualization software. See R. Koradi, M. Billeter, and K. Wüthrich, "MOLMOL: A Program for Display and Analysis of Macromolecular Structures," *Journal of Molecular Graphics* **14**, 51–55 (1996). At the right end of the figure is a representation of the protein dihydrofolate reductase, viewed both without and with the ligand folate. Data for these are from the Protein Data Bank (see H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research* **28**, No. 1, 235–242 [2000]). The data are from PDB file designation "1DHF." The primary citation is J. F. Davies II, T. J. Delcamp, N. J. Prendergast, V. A. Ashford, J. H. Freisheim, and J. Kraut, "Crystal Structures of Recombinant Human Dihydrofolate Reductase Complexed with Folate and 5-Deazafolate," *Biochemistry* **29,** No. 40, 9467–9479 (1990). The data were captured from a display rendered with the VRML plug-in for the Netscape browser. The plug-in is from Computer Associates; see www.cai.com. The plug-in allows for inclusion, or not, of the ligand bound to the protein.
6. Protein structures are archived for public use in an Internet-accessible database known as the Protein Data Bank. See http://www.rcsb.org/pdb/ and H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research* **28**, No. 1, 235–242 (2000).
7. For a highly entertaining, yet informative book that describes the exciting development of an important drug, see B. Werth, *The Billion-Dollar Molecule: One Company's Quest for the Perfect Drug*, Touchstone Books, Carmichael, CA (1995).
8. The Arabidopsis Genome Initiative, "Analysis of the Genome Sequence of the Flowering Plant *Arabidopsis Thaliana*," *Nature* **408**, No. 6814, 796–815 (2000).
9. See http://nar.oupjournals.org/content/vol29/issue1/#ARTICLES.
10. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology* **215**, 403–410 (1990).
11. Y. Duan and P. A. Kollman, "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution," *Science* **282**, No. 5389, 740–744 (1998).

Accepted for publication April 2, 2000.

William C. Swope *IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120 (electronic mail: swope@almaden.ibm.com).* Dr. Swope is a research staff member currently helping with the Blue Gene Protein Science project. He started his career in IBM at IBM Instruments, Inc., an IBM subsidiary that developed scientific instrumentation, where he worked in an advanced processor design group. He also worked for six years at the IBM Scientific Center in Palo Alto, California, where he helped IBM customers develop software for numerically intensive scientific applications. In 1992 Dr. Swope joined the IBM Research Division at Almaden, where he has been involved in software development for computational chemistry applications and in technical data management for petroleum and life sciences applications. He obtained his undergraduate degree in chemistry and physics from Harvard University and his Ph.D. degree in quantum chemistry from the University of California at Berkeley. He then performed postdoctoral research on the statistical mechanics of condensed phases in the chemistry department at Stanford University. He maintains a number of scientific relationships and collaborations with academic and commercial scientists involved in the life sciences and, in particular, drug development.