Protein Structures: From Famine to Feast

Thousands of protein structures are known and accessible. Structural genomics is building on new technology to fill in the missing pieces.

Helen M. Berman, David S. Goodsell and Philip E. Bourne

The history of biology has been a journey into ever-smaller realms. Biology began with natural history—the study and classification of living organisms. Darwin’s theory of natural selection placed this in a coherent framework by showing that the diverse organisms that fill the world arose through a process of evolution. Leeuwenhoek’s microscopes looked more closely, uncovering the world of cells and leading to the understanding that all life is organized at the microscopic into cellular units. Today, we have turned our eyes to the nanoscale, and we are striving to integrate our knowledge of life at this level through molecular biology.

Modern molecular biology is an information-laden science and requires the integration of data from many sources. In the past few decades, two types of information—genomic sequences and protein structures—have become abundant and have transformed the study of molecular biology. Genome sequences contain instructions for building molecules, cells and, ultimately, entire organisms. The genome is the blueprint for all of the proteins of an organism—this is essential information because proteins control nearly all of the processes of living cells.

Protein structure determinations perfectly complement genomic information. The proteins revealed in these structures are marvelous molecular machines. Each consists of a chain of amino acids (or several chains) whose sequence is encoded in the genome. Once constructed, that chain folds into a specific three-dimensional structure to form a molecular motor, a sturdy filament, a chemical factory or any of thousands of molecule-sized gizmos. In the past few decades, structural biologists have determined the structures of thousands of proteins, revealing the details of their function at the atomic level.

Early Structural Biology

The pioneers of structural biology toiled for years to determine the first protein structures. In 1957, after 22 years of work, John Kendrew announced the determination of the three-dimensional structure of myoglobin. For the first time, the atomic structure of a protein was revealed, providing a close look at how this protein selectively stores oxygen. As with all of science, this landmark result built on earlier discoveries. In 1913, Max von Laue and the father-son team of William Henry Bragg and William Lawrence Bragg discovered that crystals diffract x rays and that the orderly pattern of this diffraction could be used to deduce the location of every atom in a crystal. Early experiments in x-ray diffraction explored the structure of inorganic salts and small organic molecules. As the size of the molecules whose structures were deciphered grew, it became apparent that even the large molecules in cells could be studied. Pioneers such as Dorothy Hodgkin, John D. Bernal, Max Perutz and John Kendrew set to work jumping many experimental hurdles. Their work opened the atomic world of large molecules to study, and Kendrew, the Braggs, von Laue, Hodgkin and Perutz were all recognized for their contributions with Nobel Prizes.

In the decade after 1957, a dozen or so structures were determined. Each new structure provided a wealth of information about a previously invisible world, and an enthusiastic scientific community greeted each publication with excitement. In 1967, when ribonuclease, the first structure solved by an American group, was announced at a packed meeting of the American Crystallographic Association, crystallographers skipped their dinner rather than miss the groundbreaking presentations by Gopinath Kartha and Hal Wyckoff. These first structures revealed many of the basic principles of protein structure and function. Protein architecture was shown to involve a complex combination of highly ordered local structures—alpha helices and beta sheets glued together by hydrogen bonds, with a geometry perfect enough to satisfy any engineer—bent and folded into an intricate globular structure. The basic mechanisms of enzyme catalysis were revealed, showing the perfect placement of key chemical groups and the use of forcible distortion to stimulate reactions.

Helen M. Berman is a Board of Governors Professor of Chemistry and Chemical Biology at Rutgers, The State University of New Jersey. She is the director of the Protein Data Bank and the Nucleic Acid Database, and a past president of the American Crystallographic Association. Her research interests are in structural biology and structural bioinformatics, with special emphasis on protein-nucleic acid interactions. David S. Goodsell is a researcher in computational biology and computer-aided molecular design. He is author of two books on biomolecular structure and function, and currently presents the “Molecule of the Month” at the Protein Data Bank. Philip Bourne is a professor in the Department of Pharmacology at the University of California, San Diego, Director of Integrated Biosciences at the San Diego Supercomputer Center and currently president of the International Society for Computational Biology. He has a background in small molecule and protein crystallography. His current interests center around structural bioinformatics and the application of information technology in the quest to understand more about living systems. Address for Goodsell: Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037. Internet: goodsell@scripps.edu
Protein crystallography became an established part of research in molecular biology in the summer of 1971 at a Cold Spring Harbor symposium entitled “Structure and Function of Proteins at the Three-Dimensional Level.” The topics covered diverse aspects of biology, including structure-function studies of proteases, glycolytic enzymes, dehydrogenases, muscle proteins, hemoglobins, immunoglobulins and even viruses. The discussions within the meeting room, on the lawn and on the beach were exciting, intense and forwardlooking. There was a sense that a new era in biology had arrived; Sir David Phillips, one of the pioneers in protein crystallography, described structural biology as “coming of age.”

The Protein Data Bank
At the Cold Spring Harbor symposium, a group of young crystallographers looked to the future of protein crystallography and realized that this important data must be made available to the wider research community in a public data bank of protein structures. This idea had been discussed among members of the American Crystallographic Association, but a serendipitous combination of events made it a reality at the symposium.

At an ad hoc meeting of protein crystallographers, Walter Hamilton, a chemist at Brookhaven National Laboratory, became interested in the project. Hamilton was a powerful figure in the international community of crystallographers and at the age of 40 had already served on numerous influential committees, most notably acting as president of the American Crystallographic Association and as an editor of Acta Crystallographica. At that point in his ca-
reer, he had begun to focus on two new science and technology projects: the determination of the structures of all the amino acids using neutron diffraction and the development of new computer technologies for graphics and remote computing. This was a perfect background for overseeing the development of a public database. At that meeting, a depository was proposed with identical files in the United Kingdom and in the United States. Hamilton volunteered to set up the American data bank at Brookhaven, and the Protein Data Bank (PDB) was born.

When Max Perutz returned to England, he discussed this proposal with Olga Kennard, who had founded the Cambridge Crystallographic Data Centre and had wide experience in assembling and archiving crystallographic data for small organic structures. Hamilton wrote to her with an offer of collaboration and a proposal to meet and discuss some of the details of coordinating the activities. He visited England soon after the Cold Spring Harbor symposium and by October 1971, the journal *Nature New Biology* announced the establishment of the PDB, jointly operated by the Cambridge Crystallographic Data Centre and Brookhaven National Laboratory. At first, the PDB contained fewer than a dozen structures. These early structures were relatively small, so a very simple file format was sufficient to present the necessary information about each molecule.

Crystal-structure methodology evolved rapidly in the 1980s. A short two decades later, all steps in the determination of protein crystal structures have improved vastly. Highly automated procedures allow biochemists around the world to produce 3,000 new structures per year. In the past, proteins were laboriously isolated from natural sources, but molecular-biology techniques, such as producing proteins in engineered bacteria, now allow investigators to obtain large quantities of purified proteins. New methods have helped overcome the vagaries of growing crystals, allowing investigators to find precise crystal-growing conditions using exceedingly small amounts of material. Data-collection methods have improved at all levels. Crystals, which may be damaged by x rays, are routinely made more durable by flash freezing with liquid nitrogen. Sources of x rays have become much more intense, especially with the emergence of powerful synchrotron beam lines, allowing the use of ever smaller crystals. Detectors are much more sensitive and allow very rapid collection of x-ray diffraction patterns. And methods for phase determination and refinement have improved, allowing biologists many new options for determining a three-dimensional structure of proteins based on crystal diffraction patterns. Crystallography is now a powerful part of the armament of techniques employed by biology.

Although a powerful method, x-ray crystallography is not the only experimental route to an atomic structure. Other methods complement x-ray analysis and can reveal the structures of proteins that are not amenable to x-ray crystallography. Neutron crystallography reveals the positions of hydrogen atoms, which are virtually invisible in x-ray experiments. Nuclear magnetic resonance (NMR) spectroscopy allows the determination of structures in solution. This powerful technique is providing a new view of protein mobility and currently accounts for about 15 percent of the structures released in the PDB. Even electron-microscopy resolution is extending to the atomic level and becoming increasingly capable of providing detailed information on proteins and protein complexes.

These enhancements in technology have also made it possible to determine the structures of ever-larger protein complexes. The first protein structure contained just over 1,000 atomic positions, but the elucidation of struc-

---

**Figure 2.** In 1957, after 22 years of work, John Kendrew announced that he had determined the atomic structure of myoglobin, the first protein thus revealed. He is shown here in 1959 with his model of myoglobin. (Photograph courtesy of the MRC Laboratory of Molecular Biology) The drawing shows a more conventional rendering of the myoglobin structure, from Kendrew’s 1962 Nobel lecture. (Drawing courtesy of The Nobel Foundation.)
tures 50 times larger is now commonplace. The PDB includes structures of entire viruses, enzymes with dozens of subunits, and huge complexes of DNA and protein. Researchers will study the atomic details of these complexes for many years, wrestling with the many clues about their functions, and through the PDB students and educators will explore these structures and gain an appreciation for the complexity that underlies life.

Molecular Machines and Life
Protein structures provide us with a unique understanding of our own bodies and a sobering appreciation of the similarity of our bodies with those of all other organisms on Earth. Protein structures have revealed atomic details about molecules throughout the body, shedding light on everything from the mechanisms of immunity to the powering of muscle contractions, from the control of blood sugar level all the way to the molecular basis of thought and emotion. The proteins are complex and intricate, often with wildly unimaginable shapes. They fit together like lock and key; they spin like wheels; they build, destroy, repair and edit; they stretch like rubber; they are sturdy as concrete. And now, thousands of these structures are available in the PDB for use in research, education or simply for personal inspiration.

One of the most exciting results in the study of structural biology has been the elucidation of the structure of the ribosome. The ribosome is a nanoscale assembler—a machine that builds proteins atom by atom using the information in the genome. To perform this complex task, ribosomes must coordinate the action of many players: They must bring together the genomic information, the building materials and the energy to create new proteins.

The structure of the ribosome has been a subject of intense study for many years. Initially, the electron microscope was used to examine thousands of individual ribosomes, revealing their basic shape and geometry. Crystallographic structures of many pieces of the ribosome followed, slowly building up a more detailed picture. Most recently, crystallographic structures of the entire ribosome have been released. The two ribosomal subunits are bewilderingly large: The large subunit alone, a complex of protein and RNA, consists of several thousand nucleotides and 28 protein chains. Ribosomes interact with transfer RNA, messenger RNA and many initiation, elongation and termination factors. With the structures of these factors also becoming revealed, a coherent picture of cellular protein synthesis is emerging. These complex machines are busy at work at this very moment throughout your body.

Protein Structure and Human Health
Protein structure plays a central role in our current understanding of the processes of life and provides concrete information that allows this understanding to be used to improve our quality of life. Perhaps the most powerful application of protein structure is in medicine. By understanding the molecular mechanisms of biological processes, we gain the ability to modify them through pharmaceutical intervention.

In the past, new therapeutic drugs were obtained through large, random searches of natural products or synthetic compounds. Today, it is possible to design a drug to target a given molecule specifically, hitting the problem exactly at its source and minimizing the deleterious side effects on other molecules. In most cases, the targets are enzymes, proteins that catalyze particular reactions. Drugs seek out the active site machinery of enzymes, blocking their action. For instance, aspirin binds deep in the pocket of the enzyme cycloox-
genase, blocking the normal formation of pain-signaling molecules.

The goal of rational drug design is to create drugs that block such pockets perfectly and specifically. The first step of this process is to obtain a protein structure. Often the active site is apparent, a deep crease in the structure or a pocket lined with reactive amino acids. Computer simulations can show how an enzyme binds to the molecules that it normally acts on, and other simulations can help design new drug molecules that block this action. The work is most often done in iterative cycles with a research team that includes biologists, chemists and structural biologists. Computer simulation is used to design a few candidate drugs, which are then synthesized by chemists and tested by biologists. The best are then subjected to another round of improvement, and so on until the best drug is found.

One of the greatest successes of this approach is HAART (highly active anti-retroviral therapy), the major medical weapon against HIV infection. Structures have been solved for nearly every protein in HIV, the human immunodeficiency virus that causes AIDS, so we have a detailed picture of the virus and how it attacks and evades the immune system. These structures have been instrumental in the search for new drugs to fight HIV infection and AIDS. Successful drugs have been designed against two parts of the virus, reverse transcriptase and protease, and research is actively under way to find compounds to attack the other components. In a decade, these drugs have changed AIDS from a uniformly fatal disease into a manageable disease for many infected individuals. Many other human ailments are under the scrutiny of structural biology—Alzheimer’s disease, hemophilia, diabetes, malaria, cancer—and investigators are optimistic about possible treatments.

Molecular Engineering

Knowledge about protein structures provides the unprecedented opportunity to modify them and to design new structures—proteins with new and improved functions. We are becoming increasingly aware, however, of the hubris of such endeavors. Billions of years of evolution have perfected the amazing molecular machines in cells. It is remarkably difficult for us, with a mere half-century of knowledge, to change proteins in specific ways that lead to a functional product. That said, protein designers have taken some truly remarkable steps.

The best results are often obtained when we combine existing functions. Chimeric proteins may be created by engineering a linker between two known proteins. This approach is used widely to create marker proteins. It has become common to connect green fluorescent protein, a small protein from fluorescent jellyfish, to a protein under study. Researchers can add the combined protein to cells and track the normally colorless protein by following the fluorescence. Chimeric proteins have also been created to fight cancer. Cancer chemotherapy requires two functions: a targeting function to find cancer cells and a toxic function to kill them once they are found. Immunotoxins have been created by combining two proteins, each specializing in one of these functions. A specific antibody is used to target the cancer cell, and a toxin, such as the poisonous protein in castor beans, is attached to it. The resulting immunotoxin then specifically attaches to cancer cells and kills them.

Structural biologists have also been successful at modifying existing pro-
proteins to tailor their function to new applications. For instance, most organisms make a collection of digestive enzymes that chop up proteins or carbohydrates into small, absorbable pieces. These enzymes are designed to operate under specific conditions in the stomach or in special vesicles inside cells. These enzymes are also very useful in industry for processing food, for use with detergents to remove stains and for bioremediation. For these applications, the enzymes need to be stabilized so that they can operate under harsh conditions. Protein engineers have been working on modifying proteins to strengthen their protein folds without compromising the catalytic machinery.

**Structural Genomics**

Up until now, dedicated scientists have solved protein structures one at a time. With the availability of increasingly sophisticated methodology, however, a few enterprising scientists are exploring a more comprehensive approach to solving protein structures, an approach termed structural genomics. Structural genomics builds naturally from the results of the human genome project and shares with it a grand vision for expanding human understanding. It seeks to provide three-dimensional structures of as many of the proteins encoded in a given genome as possible. The ultimate goal is the determination of the proteome, an organism’s complete complement of proteins.

At the dawn of the era of structural genomics, the PDB contains more than 17,000 protein structures of varying complexity, solved individually by investigators. What impact will this new initiative have in the next decade?

---

**Figure 5.** Currently, the ribosome holds the record for the largest non-symmetrical structure in the PDB. It consists of two large subunits, each made of RNA (orange) and a collection of proteins (blues and greens). Coordinates were taken from entries 1jfj and 1jj2 at the PDB.

**Figure 6.** HIV is arguably the best understood organism (if viruses may be called organisms) known to science. The entire HIV-1 viron is shown surrounded by blood serum (center). Nearly all the proteins that make up the virus have been studied at atomic detail, as shown in the structures at right and left. In these structures, each atom is shown as a sphere with carbon atoms in white, oxygen in red and pink, nitrogen in blue, and sulfur in yellow. (The many hydrogen atoms in these proteins are not observed.) Nucleic acids are shown in green. Many of the viral components, including reverse transcriptase and protease (see Figure 7), have been the target of rational drug design in the search for effective treatments. More recently, integrase and the coat protein gp120 have been targeted. Coordinates were taken from PDB entries 1ex4, 1hys, 7hvp, 1f6u, 1g9m, 2ezp, 1ej6, and 1hiw.
Nineteen groups worldwide are now embarking on ambitious large-scale structure-determination projects. These programs involve large teams of multidisciplinary investigators that comb the genome for good candidate genes, synthesize and purify the proteins they encode and finally solve their structures. Clearly, this high-throughput approach will generate many more structures. A National Institutes of Health initiative alone is planning to generate 10,000 novel protein structures in the next 10 years. The word “novel” is important here. The 17,000 protein structures in the PDB include many duplicate structures, closely homologous structures and structures of proteins with small variations, such as proteins with point mutations and those that are bound to inhibitors. The Protein Data Bank currently includes fewer than 1,400 structurally unique protein domains. An additional 10,000 unique structures should be sufficient to fill out the unknown regions in protein fold space.

For the most part, the reception for these structures won’t be as splashy as that greeting the structure of the ribosome. Structural genomics—where proteins picked somewhat blindly from the proteome will be subjected to structural analysis—will be an excellent method for looking at many individual proteins but cannot be used to determine the atomic details of large complexes like the ribosome. Instead, structural genomics will likely reveal the structures of a large collection of single domains, compact functional components. The domains will then be analyzed to reveal how they act collectively in performing biological functions. Because structural genomics starts directly from genome sequence, the biological function of many of these proteins will not be known in advance. The overarching goal of structural genomics is not so much to determine the structure-function relationships of a given protein, but instead to determine the entire range of possible protein structures.

You might think that it would be impossible to define the entire range of protein structures, given that protein chains are so large and so diverse. However, protein fold space is not as daunting as it might first appear. One of the early surprises for structural biologists was the similar folding pat-
terns of many proteins. Nature is amazingly efficient and reuses components in a highly redundant way. Consider that a typical protein might consist of 300 amino acids. Theoretically, this leads to $20^{300}$ possible combinations—more than all the atoms in the universe. But only a tiny fraction of these protein sequences exist in nature. The evolutionary process has weeded out all of the protein sequences that are not stable, that cannot be built and that do not fold into stable, functional structures. So we can safely ignore all but the million or so different protein sequences that are made by living organisms on Earth. But the number of protein sequences is still orders of magnitude more than the estimated number of unique protein folds, which has been estimated to total in the hundreds to thousands, depending on who is counting.

Just as it’s possible to tie either a piece of string or a piece of yarn into an identical overhand knot, many different protein sequences can fold into very similar protein shapes. If at least 30 percent of the amino acids in two protein chains are identical, there is a

Figure 9. Many recent estimates suggest that the number of ways that a protein chain may fold is remarkably limited. A few folds are shown here. The protein chains are shown as ribbons, and two important configurations of the chain are highlighted: alpha helices with purple corkscrew ribbons and beta strands with yellow ribbons. The classifications were taken from the CATH database (http://www.biochem.ucl.ac.uk/bsm/cath_new) and coordinates were taken from PDB entries 2ccy, 1mbn, 1lrv, 1ppr, 1ccm, 1fbr, 1vie, 1pmn, 4bcl, 1stm, 1hcd, 1jpc, 1rie, 1got, 1air, 1ndd, 1tim, 1kvd, 1fua and 2dnj.)
high likelihood that they will fold into basically the same three-dimensional shape. But often very different proteins, ones with less than 10 percent similarity in amino acid sequences, also are found to have nearly identical structures once the structure is determined experimentally. Structure analysis is essential, because we cannot currently predict that two such different proteins will fold into similar structures simply based on the genome sequence. This unpredictability is currently being battled in the computer. Computational biologists are attempting to predict the three-dimensional structures of proteins from their primary amino acid sequences. But until this protein-folding problem is solved, we must look to direct structural analysis to provide the basic information on how proteins fold.

The limited extent of protein fold space, where many protein chains are found to adopt similar folded structures, is good news for structural genomics. It reduces the problem, requiring us to solve only the structure of each fold type, not of each and every protein sequence. Once structures of all of the folds are determined, the structures of all of the other protein sequences may be inferred using comparative or homology modeling. Homology modeling builds models of all other proteins with similar amino acid sequences, for instance, all sequences with more than 30 percent similarity. Hence a goal of structural genomics is to fill out protein fold space so that suitable templates exist for all proteins. Current estimates predict that 30 to 50 percent of all possible folds are represented today in the PDB, so we are well on our way toward our goal of determining the entire proteome.

The Future

Structural genomics is quintessentially reductionistic. But like most reductionist approaches, it provides only a fragile foundation for knowledge. Just as the release of the human genome sequence has not spelled the end of research in human biology, the results of structural genomics will not render structural molecular biology obsolete. In both cases—genome analyses and proteomic structural genomics analyses—the result is a catalogue of possibilities. The challenge then becomes understanding the interactions that turn the molecular parts list into working, living organisms.
Traditional structural science, aided by the technological advances provided by structural genomics, will be essential for understanding the assembly of individual folds into functional complexes. For example, structural genomics probably would not have aided much in the understanding of ribosomal structure and function. Many large protein structures, such as flagellar motors or nuclear pores, still hold deep mysteries. Their structure is currently studied only through electron microscopy; combining microscopy with detailed knowledge about the structures of specific components may be the best approach to understanding the structures of the complexes.

Computational science will play an essential role in predicting and analyzing molecular interactions. It is impossible, technically and in terms of time and effort, to isolate and study every interaction. Computational prediction allows specific questions to be posed and answered, narrowing the range of questions to those whose answers will provide the most information. Computational science will never replace experimental science—nature is too full of surprises—but it can expand the scope of understanding.

Finally, basic research in biochemistry and cell biology is tying everything together. Many of the most important processes in life are not the consequence of a single molecule. Instead, they are the consequence of complex regulation, activation, assembly and disassembly of multiple players. Emergent properties appear, in a way that is impossible to predict from the properties of the individual parts. Two proteins may fit together precisely, resulting in a carefully regulated complex. A series of enzymes, each activating the next, can create a sensitive signal amplifier. A protein that flexes when it binds adenosine triphosphate becomes a motor that flexes muscles when placed in the proper context. It is at this level of study that we can understand how protein functions combine to become life.

We encourage you to explore with us this exciting, ever-growing world of protein structure. The Protein Data Bank, and many user-friendly tools for looking at the structures that it contains, is available on the World Wide Web at http://www.pdb.org. Many familiar favorites—hemoglobin, insulin, DNA—are available for examination. You will also find thousands of others, each a unique example of nature’s handiwork.

**Bibliography**


