Knowledge of the three-dimensional structures of proteins is the key to unlocking the full potential of genomic information. There are two distinct directions along which cutting-edge research in structural biology is currently moving towards this goal. On the one hand, tightly focused long-term research in individual laboratories is leading towards this aim, and as work progresses, the frontiers of structural biology will expand – but in two, almost orthogonal, directions. The one hand, large consortia of structural biologists, inspired by the pace of genome sequencing, are developing methodologies for combining non-parametric linkage analysis of complex traits across the human genome. On the other hand, large, concerted efforts, inspired by the pace of genome sequencing, are developing strategies to determine new protein structures rapidly, so that it will be soon possible to predict reasonably accurate structures for most protein domains. We anticipate that a small number of complex systems, studied in depth, will provide insights across the field of biology with the aid of genome-based comparative structural analysis.

References

the past triumphs of structural biology and seeks to analyse the structures of complex molecular assemblies that are ever larger and more intricate.

Although the universe of distinct protein sequences is essentially unlimited, the number of different folding patterns for these proteins is not⁹⁻¹³. Large proteins comprise almost invariably a number of relatively small domains, which are usually 100–250 residues long.⁴⁻¹⁵ Even though proteins of enormous structural variety are generated by combining individual domains, the number of distinct domain folds seems to be limited to a few thousand.⁴⁻¹⁵ This has led to substantial efforts to develop structural genomics projects so that the Protein DataBank contains at least one example of every kind of domain fold¹⁶⁻¹⁷.

The functions of proteins cannot be understood if we consider individual protein domains separate from their molecular and cellular contexts. The precise nature of the assembly of domains into larger structural biology is crucial, as is the interplay between a particular protein with others in the cell. In the past decade, molecular machineries of increasing complexity have succumbed to structural analysis. The limit of complexity that can be studied at the atomic level appears to be considerably beyond what was imagined a decade ago, and a growing excitement arises from the anticipation of fascinating structures yet to come.

The structural analysis of supramolecular assemblies involves science that is radically different in its emphasis and style from the science that drives projects in structural genomics. The former emphasizes depth, focus and the individual investigator, whereas the latter places a high premium on breadth, speed and the formation of large consortia. These differences are sufficiently deep to cause conflicts between extreme proponents of the two approaches. Nevertheless, both approaches are essential for transcending the wealth of sequence information generated by the genome projects into coherent mechanisms of function.

Structural genomics

Structural genomics is a new effort to determine rapidly the structures of proteins expected to contain new folds. It is impractical to provide experimentally derived structures for every gene in a particular genome, even for organisms with very small genomes, because of inherent difficulties in protein expression, crystallization and solubility for many proteins, particularly those associated with membranes. Instead, the generation of a set of structures representative of most of the possible folds for individual protein domains is feasible and likely to be achieved in the near term. Such a representative set will then allow useful structural characterizations of the remaining protein sequences through computational analysis¹⁸⁻¹⁹.

Structural genomics is feasible because of developments in molecular biology that allow more rapid production of sufficient quantities of pure protein as well as because of developments in X-ray crystallography and nuclear magnetic resonance spectroscopy that allow more rapid determination of protein structures. Perhaps the most significant of these developments is the ability to determine experimental phases for X-ray diffraction data by carrying out multi-wavelength experiments on synchrotron beam lines. Using protein crystals in which methionine residues are replaced by selenomethionine, it is now possible to record all the X-ray measurements required to generate an experimental electron-density map for a small pro-protein in significantly less than an hour – instead of the weeks of experimental time required for a conventional crystallographic-structure determination²⁰.

Two key strategies distinguish structural genomics from conventional structural biology. The first is the generation of one or more lists of protein targets that serve as the master plan for the project²¹. In general, structural genomics focuses on proteins for which a connection to a known protein structure cannot be made and are therefore more likely to contain new folds. However, more specialized target lists can also be drawn up by concentrating on important or convenient organisms whose genomes have been sequenced completely, including thermophilic bacteria²², eukaryotic organisms such as Saccharomyces cerevisiae²³ or pathogenic bacteria such as Mycobacterium tuberculosis. Alternatively, lists can be obtained with practical applications foremost in mind. For example, a list might include proteins identified as targets for the design of inhibitors with potential therapeutic value or proteins that are implicated in human cancer²⁴.

A second strategy that underlies structural genomics is the emphasis given to working rapidly through a list. Difficult proteins might be skipped altogether, and knowledge of these folds might be obtained from other proteins that are predicted to be structurally similar. This is a crucial distinction from conventional structural biology, which focuses on a particular target and grapples with it until it succumbs to structural analysis. To be effective, the targets for structural genomics have to be chosen to allow calculation of useful models for most protein domains, while minimizing the total experimental effort. We first propose a useful level of accuracy for the models based on the experimental structures and then estimate how many structures need to be determined experimentally to achieve the required level of accuracy.
The SrC kinases are regulated by the coordinated action of two peptide-binding modules, known as the SrC-homology domains SH2 and SH3. Shown here is the crystal structure of the SrC kinase Hck, determined in complex with an Inhibitor molecule-bound at the ATP-binding site. The catalytic domain of the kinase is shown in blue, whereas the SH2 and SH3 domains are coloured yellow and green, respectively. The SrC kinase contains two sites of tyrosine phosphorylation, one at the active site of the enzyme (Tyr416) and one at the C-terminal tail (Tyr527). The inactive form of the protein, shown here, contains phosphorylated Tyr527 but has no phosphate group on Tyr416. The phosphorylation on the C-terminal tail engages the SH2 domain of the protein, which then sets up a polyproline type II helix (PP-II) to which the SH3 domain binds. The catalytic activity of the kinase is turned off because the activation segment (red) blocks the substrate-binding site and because catalytic residues not shown are displaced from the active site. The challenge in determining the structures of signalling molecules such as Hck lies in defining the states of the system that are appropriate for structural analysis. When the SrC kinase becomes activated, Tyr527 is displaced from the SH2 domain, which, along with the SH3 domain, moves away from the protein and binds to external targets. In this state, the intact SrC protein cannot be crystallized easily because it is very flexible. The structure of the active form of the catalytic domain has been determined by crystalizing that domain separately. We anticipate that the structures of many more-complicated signalling assemblies will be determined once they are fully characterized by detailed biochemical analysis.

Using comparative or homology modelling a three-dimensional model of a protein sequence is constructed, based on known structures of related proteins. The accuracy of a model tends to increase with the sequence similarity between the modelled sequence and the related known structures. To obtain a reasonable level of accuracy, the models must be based on alignments with few errors. This is usually possible when the sequence identity between the modelled sequence and at least one known structure is higher than 30%. Thus, structural genomics should determine protein structures such that all known structures that need to be produced by structural genomics. To estimate this number, we have to consider how protein sequences and structures cluster with each other. The major evolutionary mechanism for generating complexity involves gene duplication followed by sequence divergence, rather than an unlimited increase in the number of distinct folds. Thus, the number of distinct protein families per genome does not increase in proportion to the number of proteins, even though the number of proteins per genome does increase with the complexity of the organism. For example, most of the 479 proteins in the very simple genome of Mycoplasma genitalium are expected to have unique folds, whereas more than 60% of the 20 000 proteins of Caenorhabditis elegans share a domain with another protein in the same genome. Protein domains that have similar folds, but not necessarily detectably similar sequences, are grouped into fold families. A reasonable guess as to the number of fold families covering almost all protein domains is a few thousand, of which ~1 000 are already known. Within each fold family, there are sequence families (smaller groupings of domains that are related in terms of their sequence). When a 30% sequence-identity cut-off is imposed on the sequence family, it is estimated that there are five times as many sequence families as there are fold families, of which ~2000 have already been structurally defined. It is likely therefore that structural genomics will have to produce structures for at least 10 000 protein domains. If successful, experimental structure determination of 10 000 properly chosen proteins should result in useful three-dimensional models for domains in hundreds of thousands of other proteins.

Determining the structures of molecular assemblies

Our ability to predict how macromolecules interact with each other is negligibly inadequate. Indeed, the problem of predicting how a protein domain engages another protein domain or a segment of DNA or RNA is perhaps even more difficult than the protein-folding problem. For example, the structure of the tyrosine kinases of the SrC family is modular, and each of these signalling proteins consists of three major components: two peptide-recognition modules, known as Src-homology 2 (SH2) and SH3 domains, and a catalytic tyrosine-kinase unit. The structures of all three domains have been determined independently, but an understanding of how the SH2 and SH3 domains cooperate to turn off catalytic activity required the determination of the fully assembled, inactive form of the protein (Fig. 2).

There are many examples of molecular assemblies where careful consideration of the functional states has led recently to remarkably informative structures. These include structures of the CaM protein bound to Holliday junction intermediates, the T-cell receptor bound to MHC-peptide complexes, DNA or RNA polymerases bound to template-primer complexes, various transcriptional complexes, and the K+ channel. These examples illustrate the kinds of project in which a large investment in
understanding the basic biochemistry of particular systems has resulted in a substantial payoff in terms of mechanistic insights. The discovery that the large (50S) ribosomal subunit from a thermophilic organism can be crystallized and that X-ray diffraction data to 3.0 Å can be measured from these crystals heralded a new era in structural biology35. This ribosomal subunit has a molecular mass of approximately $1.5 \times 10^6$ Daltons, and contains 3050 nucleotides of RNA and 53 different proteins. The large subunit yields high-quality crystals, despite the lack of internal symmetry. This finding—a landmark in structural biology—suggests that the atomic arrangement of molecular machines will be deciphered eventually.36

A partial list of large and fascinating molecular assemblies that have had their structures determined recently includes blue tongue virus, for which the entire genomic content has been visualized37, the integral membrane protein cytochrome c oxidase38, F$_{1}$-ATPase39, and the nucleosome core particle40. The structure of the nucleosome core particle provides an interesting illustration of how the analysis of large molecular assemblies has progressed. The nucleosome is the fundamental unit of DNA packaging in eukaryotes, and it consists of a central spiral arrangement of proteins around which the DNA double helix is coiled. The protein scaffold, known as a histone octamer, is made up of two copies each of four different kinds of histones, known as H2A, H2B, H3, and H4. These histones are very similar in their amino acid sequences and three-dimensional structures. DNA packaging involves the repeated winding of genomic DNA onto tandem arrays of nucleosome cores.

The interaction between the core and the DNA is non-specific. At first glance, this would appear to make nucleosome cores very difficult targets for crystallization. Indeed, the earliest crystals consisted of histone–DNA complexes that were purified directly from the nuclei of eukaryotic cells and diffracted X-rays only to low resolution ($\sim 7$ Å). The key to achieving high resolution was to remove heterogeneity in the complexes by using recombinant histone proteins and artificially prepared DNA samples of defined length and sequence. This resulted in high-resolution views of the human histone octamer bound to 146 base pairs of DNA (Fig. 3). The structure is both breathtaking in its beauty and deeply informative about the mechanisms of DNA packaging and in its regulation.40

Will the advent of structural genomics remove the thrill of seeing new protein structures? The combination of physics, chemistry, biology and natural history that underlies protein-structure analysis makes structural biology uniquely attractive to many of us. Although the mechanistic goal of understanding protein function in terms of physics and chemistry is ultimately of overriding concern, the mechanistic goal of understanding protein function in terms of physics and chemistry is ultimately of overriding concern, the mechanistic goal of understanding protein function in terms of physics and chemistry is ultimately of overriding concern, the mechanistic goal of understanding protein function in terms of physics and chemistry is ultimately of overriding concern, the mechanistic goal of understanding protein function in terms of physics and chemistry is ultimately of overriding concern. The key to achieving high resolution was to remove heterogeneity in the complexes by using recombinant histone proteins and artificially prepared DNA samples of defined length and sequence. This resulted in high-resolution views of the human histone octamer bound to 146 base pairs of DNA (Fig. 3). The structure is both breathtaking in its beauty and deeply informative about the mechanisms of DNA packaging and in its regulation.40
The combination of the tools and principles of chemistry, together with the tools of modern molecular biology, allow us to create complex synthetic and natural molecules, and processes with novel biological, chemical and physical properties. This article illustrates the tremendous opportunity that lies at this interface of chemistry and biology by describing a number of examples, ranging from efforts to expand the genetic code of living organisms to the use of combinatorial methods to generate biologically active synthetic molecules.