Acknowledgements

We thank M. Nelson for preparing Fig. 1 and L. Barcellos, M. Grote, J. Hollenbach, L. Louie, K. Mather, S. McWeeney, D. Meyer, E. Mignot, M. Nelson, H. Payami, D.C. Rao, J-X. She and A.M. Valdes for helpful input on the manuscript. The work was supported by NIH grant GM56688.

References

1 Goodman, L. (1998) The human genome project aims for 2003. *Genome Res.* 8, 997–999

2 Collins, F.S. *et al.* (1997) Variations on a theme: cataloguing the human DNA sequence variation. *Science* 278, 1580–1581

3 Gelehrter, T.D. *et al.* (1998) *Principles of Medical Genetics*, Williams and Wilkins

4 Thomson, G. (1997) Strategies involved in mapping diabetes genes: an overview. *Diabetes Rev.* 5, 106–115

5 Collins, F.S. *et al.* (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* 8, 1229–1231

6 Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72, 971–983

7 Lander, E.S. and Schork, N.J. (1994) Genetic dissection of complex traits. *Science* 265, 2037–2048

8 Thomson, G. (1995) HLA disease associations: models for the study of complex human genetic disorders. *Crit. Rev. Clin. Lab. Sci.* 32, 183–219

9 Thomson, G. (1995) Analysis of complex human genetic traits: an ordered notation method and new tests for mode of inheritance. *Am. J. Hum. Genet.* 57, 474–486

10 Elston, R.C. (1998) Linkage and association. *Genet. Epidemiol.* 15, 565–576

11 Levy-Lahard, E. *et al.* (1998) Recent advances in the genetics of Alzheimer's disease. *J. Geriatr. Neur.* 11, 42–54

12 Thomson, G. (1995) Mapping disease genes: family based association studies. *Am. J. Hum. Genet.* 57, 487–498

13 Payami, H. *et al.* (1998) The affected sib method. IV. Sib trios. *Ann. Hum. Genet.* 49, 303–314

14 Martin, E.R. *et al.* (1997) Tests for linkage and association in nuclear families. *Am. J. Hum. Genet.* 61, 439–448

15 Wynshaw-Boris, A. (1996) Model mice and human disease. *Nat. Genet.* 13, 259–260

16 Pugliese, A. (1999) Unraveling the genetics of insulin-dependent type 1A diabetes: the search must go on. *Diabetes Rev.* 7, 39–54

17 Mein, C.A. *et al.* (1998) A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nat. Genet.* 19, 297–300

18 Lin, L. *et al.* (1999) The sleep disorder canine narcolepsy is caused by a mutation in the *hypocretin* (*orexin*) *receptor 2* gene. *Cell* 98, 365–376

19 Thorsby, E. (1997) Invited anniversary review: HLA associated diseases. *Hum. Immunol.* 53, 1–11

20 Concannon, P. *et al.* (1998) A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat. Genet.* 19, 292–296

21 Carrington, M. *et al.* (1999) HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 283, 1748–1752

22 Lernmark, A. and Ott, J. (1998) Sometimes it's hot, sometimes it's not. *Nat. Genet.* 19, 213–214

23 Suarez, B.K. *et al.* (1994) in *Genetic Approaches to Mental Disorders* (Gershon, E.S. and Cloninger, C.R., eds), pp. 23–46, American Psychiatric Press

24 Brown, P.O. and Hartwell, L. (1998) Genomics and human disease – variations on variation. *Nat. Genet.* 18, 91–93

25 Barcellos, L.F. *et al.* (1997) Chromosome 19 single-locus and multilocus haplotype associations with multiple sclerosis: evidence of a new susceptibility locus in Caucasian and Chinese patients. *J. Am. Med. Assoc.* 278, 1256–1261

26 Gu, C. *et al.* (1998) Meta-analysis methodology for combining non-parametric sibpair linkage results: genetic homogeneity and identical markers. *Genet. Epidemiol.* 15, 609–626

27 Wise, L.H. *et al.* (1999) Meta-analysis of genome searches. *Ann. Hum. Genet.* 63, 263–272

28 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517

29 Morton, N.E. and Collins, A. (1998) Tests and estimates of allelic association in complex inheritance. *Proc. Natl. Acad. Sci. U. S. A.* 95, 11389–11393

30 Nickerson, D. *et al.* (1998) DNA sequence diversity in a 9.7 kb region of the human lipoprotein lipase gene. *Nat. Genet.* 19, 233–239

31 Escamilla, M.A. *et al.* (1999) Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: an initial screen for bipolar disorder loci on chromosome 18. *Am. J. Hum. Genet.* 64, 1670–1678

32 Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139–144

33 Barcellos, L.F. *et al.* (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.* 61, 734–747

34 Cargill, M. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238

35 Halushka, M.K. *et al.* (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22, 239–247

36 Schork, N.A. *et al.* (1998) The future of genetic epidemiology. *Trends Genet.* 14, 266–272

37 Todorov, A.A. and Rao, D.C. (1997) Trade-off between false positives and false negatives in the linkage analysis of complex traits. *Genet. Epidemiol.* 14, 453–464

38 Paterson, A.D. and Petronis, A. (1999) Sex of affected sibpairs and genetic linkage to type 1 diabetes. *Am. J. Med. Genet.* 84, 15–19

39 Huttley, G.A. *et al.* (1999) A scan for linkage disequilibrium across the human genome. *Genetics* 152, 1711–1722

Challenges at the frontiers of structural biology

Andrej Šali and John Kuriyan

Andrej Šali
sali@rockefeller.edu

John Kuriyan*
kuriyan@rockvax.rockefeller.edu

Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology,
*Howard Hughes Medical Institute,
The Rockefeller University,
1230 York Ave,
New York,
NY 10021, USA.

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 to the determination of the structures of macromolecular assemblies of ever-increasing size and complexity. On the other hand, large consortia of structural biologists, inspired by the pace of genome sequencing, are developing strategies to determine new protein structures rapidly, so that it will soon be 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.

To understand fully the workings of the cell, we face the challenge of describing the three-dimensional structures of all the cellular components at an atomic level of detail, and relating these structures to molecular mechanisms. As work towards this aim progresses, the frontiers of structural biology will expand – but in two, almost orthogonal, directions. The newer theme, referred to as 'structural genomics', is motivated by the growing impact of genome-sequencing efforts and is aimed at accelerating the rate at which protein structures containing new folds are solved. The other thrust of research builds on

TCB, Vol. 9, N° 12 (0962-8924) PII: S0962-8924(99)01685-2
TIBS, Vol. 24, N° 12 (0968-0004) PII: S0968-0004(99)01494-2
TIG, Vol. 15, N° 12 (0168-9525) PII: S0168-9525(99)01908-3

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 not[1–3]. Large proteins comprise almost invariably a number of relatively small domains, which are usually 100–250 residues long[4–6]. 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[1–3]. This has led to international efforts to develop systematic structural genomics projects so that the Protein DataBank contains at least one example of every kind of domain fold[7–11].

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 proteins 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 translating 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 analysis[12–18].

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[19] 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[20]. 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-tein in significantly less than an hour – instead of the weeks of experimental time required for a conventional crystallographic-structure determination[21].

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 masterplan for the project[7,11]. 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 bacteria[7,9], eukaryotic organisms such as *Saccharomyces cerevisiae*[10] 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 cancer[10].

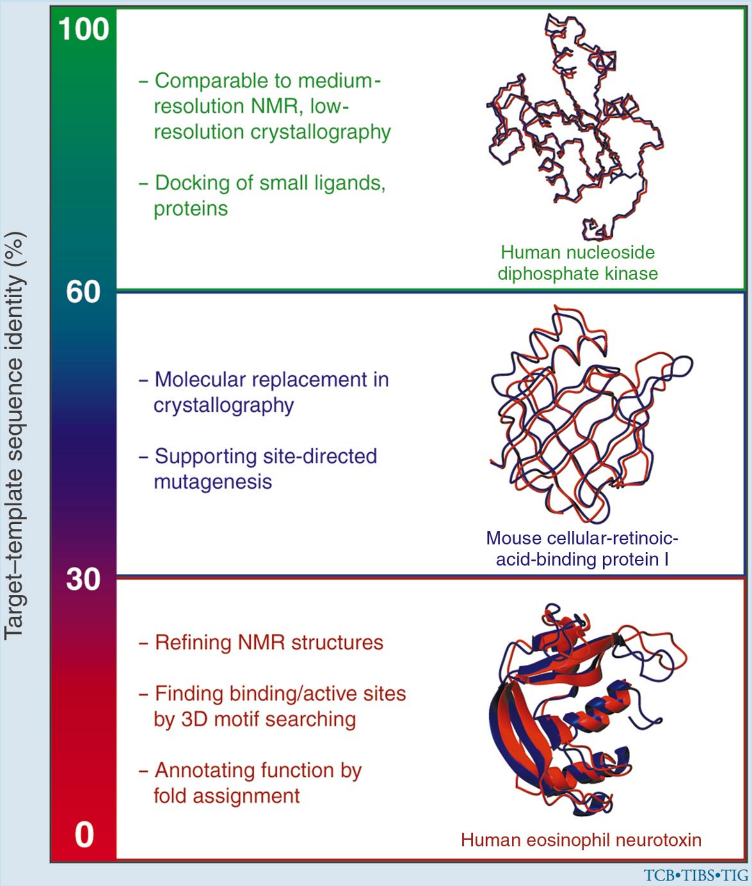

100
60
30
0
Target–template sequence identity (%)
– Comparable to medium-resolution NMR, low-resolution crystallography
– Docking of small ligands, proteins
Human nucleoside diphosphate kinase
– Molecular replacement in crystallography
– Supporting site-directed mutagenesis
Mouse cellular-retinoic-acid-binding protein I
– Refining NMR structures
– Finding binding/active sites by 3D motif searching
– Annotating function by fold assignment
Human eosinophil neurotoxin
TCB•TIBS•TIG

FIGURE 1. Schematic diagram showing the range of accuracy obtained by comparative modelling[23]. The potential uses of comparative models depend on their accuracy. This in turn depends significantly on the sequence identity between the sequence modelled and the known structure on which the model was based. Sample models (red) are compared with the actual structures (blue).

A second strategy that underlies structural genomics is the emphasis given to working rapidly through a list[7]. Difficult proteins might be skipped altogether, and knowledge of their 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.

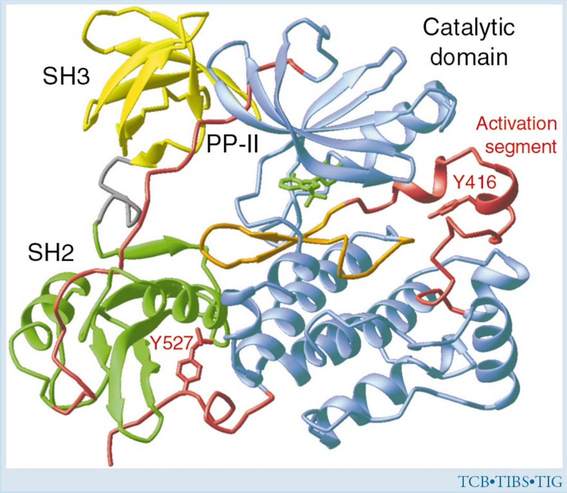


FIGURE 2. Crystal structure of a Src kinase in the inactive form[42,43]. The Src kinases are regulated by the coordinated action of two peptide-binding modules, known as the Src-homology domains SH2 and SH3. Shown here is the crystal structure of the Src kinase Hck, determined in complex with an inhibitor molecule bound at the ATP-binding site[42]. The catalytic domain of the kinase is shown in blue, whereas the SH3 and SH2 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 phosphotyrosine 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 crystallizing that domain separately[44]. 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 (Fig. 1)[22,23]. The accuracy of a model tends to increase with the sequence similarity between the modelled sequence and the related known structures[13]. 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 sequences in the genome databases match at least one structure in the structural database with an overall sequence identity of no less than 30% (Refs 8 and 23). If this degree of sampling is achieved, most of the models will be based on sequence identity in the range of 30–50%. Such models tend to have more than 85% of the Cα atoms within 3.5 Å of their correct positions[23]. For functional analysis, the accuracy of the models is frequently higher because the active-site regions generally exhibit stronger structural conservation than the rest of the protein. The models based on more than 30% sequence identity are usually suitable for a number of applications[23], including the testing of ligand-binding modes by designing site-directed mutants with altered binding capacity and computational screening of databases of small molecules for potential inhibitors or lead compounds[24]. A fraction of the models will be based on more than 50% sequence identity. The average accuracy of such models approaches that of low-resolution X-ray structures (3 Å resolution) or medium-resolution NMR structures (ten long-range restraints per residue)[23]. In addition to the applications listed above, high-quality models can be used for more-reliable calculations of ligand docking and drug design.

The requirement that each protein domain share at least 30% sequence identity with a known structure determines the number of protein 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[14], rather than an unlimited increase in the number of distinct folds. Thus, the number of distinct protein folds 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 80% 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[4–6]. A reasonable guess as to the number of fold families covering almost all protein domains is a few thousand, of which ~1000 are already known[1–3]. 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[1–3], 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[18,25].

Determining the structures of molecular assemblies

Our ability to predict how macromolecules interact with each other is woefully 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[26,27]. 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[28]. 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).

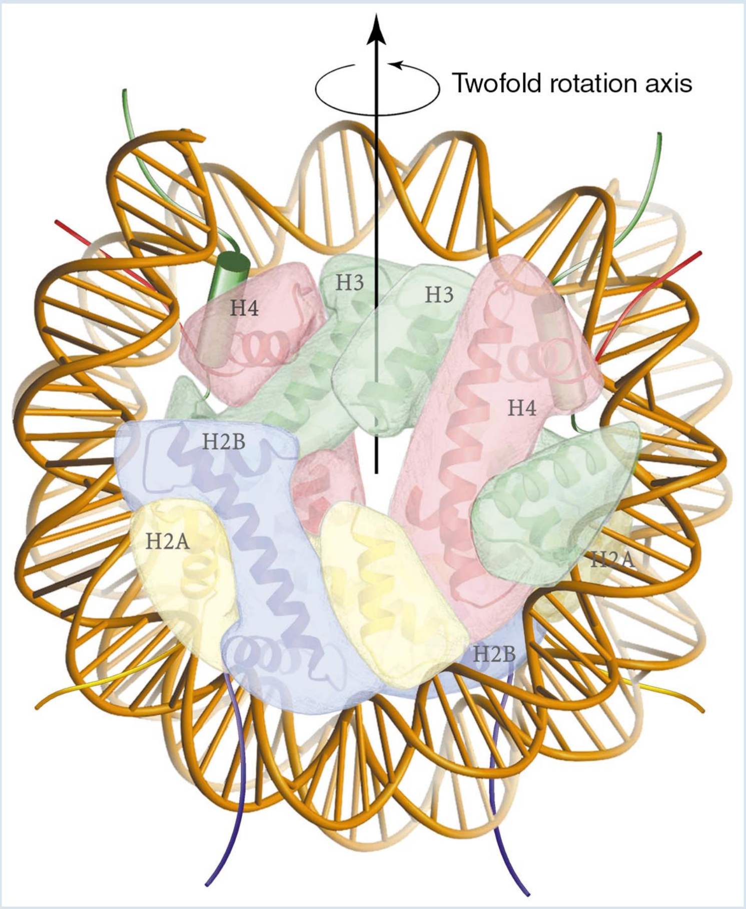
There are many examples of molecular assemblies where careful consideration of the functional states has led recently to remarkably informative structures. These include structures of the Cre protein bound to Holliday junction intermediates[29], the T-cell receptor bound to MHC–peptide complexes[30,31], DNA or RNA polymerases bound to template–primer complexes[32], various transcriptional complexes[33] and the K^+ channel[34]. 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 payback 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 biology[35]. This ribosomal subunit has a molecular mass of approximately 1.5×10^6 Daltons, and contains ~3000 nucleotides of RNA and 33 different proteins. The large subunit yields highly ordered crystals, despite the lack of internal symmetry. This finding – a landmark in structural biology – suggests that the atomic arrangements of this most complex 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 visualized[37], the integral membrane protein cytochrome *c* oxidase[38,39], F_1-ATPase[40] and the nucleosome core particle[41].

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 cores and the DNA is nonspecific. 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 (~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 its regulation[41].



TCB•TIBS•TIG

FIGURE 3. The structure of the nucleosome core, as determined by Richmond and colleagues[41]. The histone proteins form a spiral-shaped octameric assembly around which DNA is coiled. The histone octamer consists of two copies each of four different histone proteins – H2A, H2B, H3 and H4. These proteins contain tails that are shown protruding from the nucleosome. The tails are likely to be important in stabilizing the arrangement of nucleosomes in higher-order structures. Copyright 1999, Lore Leighton, used with permission.

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 first look at a new protein structure provides a thrill that could be compared to that felt by explorers in the late 19th century upon discovery of new biological species. The surprises that ensue from visualizing new protein structures have made structural biology very creative. New ideas and mechanisms have sprung from the process of discovering the unexpected features in protein structures.

The highly programmatic and automated operations of large structural genomics consortia are likely to reduce the pleasure that individuals take in the process. However, before blaming structural genomics for the impending loss of a romantic period in structural biology, it is important to take a hard look at the current state of protein-structure determination. Someone who is working on the structure of a protein today is probably operating in direct competition with several other groups. Even when the race is won, the thrill of looking at something really new is rapidly disappearing because of the fundamental redundancy in protein structure. For questions in structural biology that hinge on the determination of the structures of relatively simple proteins, the romantic period is already over. It is probably wiser to let programmes in structural genomics answer these questions from now on.

Is this the end for creativity in protein-structure determination? It might have been – but for the realization that the structures of large, biologically significant molecular assemblies are now being determined with increasing frequency, giving hope that the wonder of viewing unexpected macromolecular structures will not diminish in the near future. There is the expectation that an increased understanding of complex cellular machineries will come from focused structural attacks on a diversity of problems. That such efforts will eventually be successful for a significant number of systems is indicated by the broad range of results that are already at hand. In parallel, structural genomics will provide us with experimental structures and useful models for most protein domains from all organisms. There is hope that the two seemingly orthogonal directions in structural biology will ultimately be integrated into a complete structural and mechanistic characterization of individual domains, proteins, as well as their assemblies.

References

1 Orengo, C.A. *et al.* (1994) Protein superfamilies and domain superfolds. *Nature* 372, 631–634

2 Holm, L. and Sander, C. (1996) Mapping the protein universe. *Science* 273, 595–602

3 Brenner, S.E. *et al.* (1997) Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* 7, 369–376

4 Hubbard, T.J.P. *et al.* (1999) SCOP: A structural classification of proteins database. *Nucleic Acids Res.* 27, 254–256

5 Holm, L. and Sander, C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* 27, 244–247

6 Orengo, C.A. *et al.* (1999) The CATH database provides insights into protein-structure–function relationship. *Nucleic Acids Res.* 27, 275–279

7 Terwilliger, T.C. *et al.* (1998) Class-directed structure determination: Foundation for a Protein Structure Initiative. *Protein Sci.* 7, 1851–1856

8 Šali, A. (1998) 100,000 protein structures for the biologist. *Nat. Struct. Biol.* 5, 1029–1032

9 Zarembinski, T.I. *et al.* (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl. Acad. Sci. U. S. A.* 95, 15189–15193

10 Burley, S.K. *et al.* (1999) Structural genomics: beyond the Human Genome Project. *Nat. Genet.* 23, 151–157

11 Cort, J.R. *et al.* (1999) A phylogenetic approach to target selection for structural genomics: solution structure of YciH. *Nucleic Acids Res.* 27, 4018–4027

12 Fischer, D. and Eisenberg, D. (1996) Fold recognition using sequence-derived predictions. *Protein Sci.* 5, 947–955

13 Sánchez, R. and Šali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 13597–13602

14 Teichmann, S.A. *et al.* (1998) Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. U. S. A.* 22, 14658–14663

15 Huynen, M. *et al.* (1998) Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* 280, 323–326

16 Rychlewski, L. *et al.* (1998) Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold. Des.* 3, 229–238

17 Jones, D.T. (1999) Genthreader: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287, 797–815

18 Guex, N. *et al.* (1999) Protein modelling for all. *Trends Biochem. Sci.* 24, 364–367

19 Wüthrich, K. (1998) The second decade – into the third millennium. *Nat. Struct. Biol.* 5, 492–495

20 Hendrickson, W.A. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron data. *Science* 254, 51–58

21 Walsh, M.A. *et al.* (1999) Taking MAD to the extreme: ultrafast protein structure determination. *Acta Crystallogr.* D55, 1168–1173

22 Blundell, T.L. *et al.* (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326, 347–352

23 Martí-Renom, M.A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* (in press)

24 Ring, C.S. *et al.* (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. U. S. A.* 90, 3583–3587

25 Sánchez, R. and Šali, A. ModBase: A database of comparative protein structure models. *Bioinformatics* (in press)

26 Conte, L.L. *et al.* (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* 285, 2177–2198

27 Nadassy, K. *et al.* (1999) Structural features of protein–nucleic acid recognition sites. *Biochemistry* 38, 1999–2017

28 Sicheri, F. and Kuriyan, J. (1997) Structures of Src-family tyrosine kinases. *Curr. Opin. Struct. Biol.* 7, 777–785

29 Gopaul, D.N. *et al.* (1998) Structure of the Holliday junction intermediate in Cre-loxP site-specific recombination. *EMBO J.* 17, 4175–4187

30 Garboczi, D.N. *et al.* (1996) Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* 384, 134–141

31 Garcia, K.C. *et al.* (1996) An $\alpha\beta$ T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. *Science* 274, 209–219

32 Doublie, S. *et al.* (1999) An open and closed case for all polymerases. *Struct. Fold. Des.* 7, R31–R35

33 Patikoglou, G. and Burley, S.K. (1997) Eukaryotic transcription factor–DNA complexes. *Annu. Rev. Biophys. Biomol. Struct.* 26, 289–325

34 Doyle, D.A. *et al.* (1998) The structure of the potassium channel: molecular basis of conduction and selectivity. *Science* 280, 69–77

35 von Bohlen, K. *et al.* (1991) Characterization and preliminary attempts for derivatization of crystals of large ribosomal subunits from *Haloarcula marismortui* diffracting to 3 Å resolution. *J. Mol. Biol.* 222, 11–15

36 Ban, N. *et al.* (1999) Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. *Nature* 400, 841–847

37 Gouet, P. *et al.* (1999) The highly ordered double-stranded RNA genome of bluetongue virus revealed by crystallography. *Cell* 97, 481–490

38 Iwata, S. *et al.* (1995) Structure at 2.8 Å resolution of cytochrome *c* oxidase from *Paracoccus denitrificans*. *Nature* 376, 660–669

39 Tsukihara, T. *et al.* (1995) Structures of metal sites of oxidized bovine heart cytochrome *c* oxidase at 2.8 Å. *Science* 269, 1069–1074

40 Abrahams, J.P. *et al.* (1994) Structure at 2.8 Å resolution of F_1-ATPase from bovine heart mitochondria. *Nature* 370, 621–628

41 Luger, K. *et al.* (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389, 251–260

42 Schindler, T. *et al.* (1999) Crystal structure of Hck in complex with a Src family-selective tyrosine kinase inhibitor. *Mol. Cell* 3, 639–648

43 Xu, W. *et al.* (1999) Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol. Cell* 3, 629–638

44 Yamaguchi, H. and Hendrickson, W.A. (1996) Structural basis for activation of the human lymphocyte kinase Lck upon tyrosine phosphorylation. *Nature* 384, 484–489

Acknowledgements

We thank S.K. Burley, B. Honig and R. MacKinnon for discussions. We are also grateful to M. Martí-Renom for preparing Fig. 1. A. Šali is an Alfred P. Sloan Research Fellow.

Opportunities at the interface of chemistry and biology

Andrew B. Martin and Peter G. Schultz

Andrew B. Martin
andym@scripps.edu

Peter G. Schultz
Peter.schultz@nifg.Novartis.com

The Scripps Research Institute,
10550 North Torrey Pines Road,
La Jolla, CA 92037,
USA; and NIFG,
3115 Merryfield Row, San Diego,
CA 92121, USA.

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.

The tools of chemistry, most notably chemical synthesis and spectroscopy, have had a remarkable impact on biology – from the structural elucidation of the double helix to the chemical synthesis of peptides and oligonucleotides. At the same time, modern molecular biology has made it possible not only to manipulate protein and nucleic acid structure but also the genetic composition of living organisms. The ability to use these tools in combination opens an unprecedented opportunity in the coming millennium, both for understanding complex biological systems at a molecular level as well as for the generation of molecules with novel biological, chemical and physical properties. This article illustrates both the opportunities and the challenges that lie at this interface of chemistry and biology by describing a number of examples, including the use of combinatorial methods to generate novel biological,

TCB, Vol. 9, N° 12 (0962-8924) PII: S0962-8924(99)01682-7
TIBS, Vol. 24, N° 12 (0968-0004) PII: S0968-0004(99)01498-X
TIG, Vol. 15, N° 12 (0168-9525) PII: S0168-9525(99)01907-1