Section I

DATA COLLECTION, ANALYSIS, AND VISUALIZATION

1

DEFINING BIOINFORMATICS AND STRUCTURAL BIOINFORMATICS

Russ B. Altman and Jonathan M. Dugan

WHAT IS BIOINFORMATICS?

The precise definition of bioinformatics is a matter of debate. Some define it narrowly as the development of databases to store and manipulate genomic information. Others define it broadly as encompassing all of computational biology. Based on its current use in the scientific literature, bioinformatics can be defined as the study of two information flows in molecular biology (Altman, 1998). The first information flow is based on the central dogma of molecular biology: DNA sequences are transcribed into mRNA sequences; mRNA sequences are translated into protein sequences; and protein sequences fold into three-dimensional structures that have functions. These functions are selected, in a Darwinian sense, by the environment of the organism, which drives the evolution of the DNA sequence within a population. The first class of bioinformatics applications, then, can address the transfer of information at any stage in the central dogma, including the organization and control of genes in the DNA sequence, the identification of transcriptional units in DNA, the prediction of protein structure from sequence, and the analysis of molecular function. These applications include the emergence of system-wide analyses of biological phenomenon, now called systems biology. Systems biology aims to achieve quantitative understanding not only of the individual players in a biological system but also of the properties of the system itself that emerge from the interaction of all its parts. This field also includes the new field of metagenomics, where we study entire ecosystems of interacting organisms. In the same way that systems biology studies how the molecular entities in a cell combine to make the cell work, metagenomics studies how the individual organisms within an ecological system combine to create that ecology. The initial forays into metagenomics are based on high-throughput sequencing not of individual species (that generally cannot be isolated) but of the mixture of species that create an ecosystem.

Structural Bioinformatics, Second Edition Edited by Jenny Gu and Philip E. Bourne

The second information flow is based on the scientific method: we create hypotheses regarding biological activity, design experiments to test these hypotheses, evaluate the resulting data for compatibility with the hypotheses, and extend or modify the hypotheses in response to the data. The second class of bioinformatics applications addresses the transfer of information within this protocol, including systems that generate hypotheses, design experiments, store and organize the data from these experiments in databases, test the compatibility of the data with models, and modify hypotheses. The emergence and emphasis on systems-level modeling and interactions in both systems—biology and metagenomics—create major new challenges for our field.

The explosion of interest in bioinformatics has been driven by the emergence of experimental techniques that generate data in a high-throughput fashion—such as high-throughput DNA sequencing, mass spectrometry or microarray expression analysis (Miranker, 2000; Altman and Raychaudhuri, 2001; The Genome International Sequencing Consortium, 2001; Venter et al., 2001). Bioinformatics depends on the availability of large data sets that are too complex to allow manual analysis. The rapid increase in the number of three-dimensional macromolecular structures available in databases such as the Protein Data Bank (PDB,[1] Chapter 11; Berman et al., 2000) has driven the emergence of a subdiscipline of bioinformatics: *structural bioinformatics*. Structural bioinformatics is the subdiscipline of bioinformatics that focuses on the representation, storage, retrieval, analysis, and display of structural information at the atomic and subcellular spatial scales.

Structural bioinformatics, like many other subdisciplines within bioinformatics,[2] is characterized by two goals: the creation of general purpose methods for manipulating information about biological macromolecules and the application of these methods to solve problems in biology and create new knowledge. These two goals are intricately linked because part of the validation of new methods involves their successful use in solving real problems. At the same time, the current challenges in biology demand the development of new methods that can handle the volume of data now available and the complexity of models that scientists must create to explain these data.

Structural Bioinformatics Has Been Catalyzed by Large Amounts of Data

Biology has attracted computational scientists over the past 30 years in two distinct ways. First, the increasing availability of sequence data has been a magnet for those with an interest in string analysis, algorithms, and probabilistic models (Gusfield, 1997; Durbin et al., 1998). The major accomplishments have been the development of algorithms for pair-wise sequence alignment, multiple alignment, the definition and discovery of sequence motifs, and the use of probabilistic models, such as hidden Markov models to find genes (Burge and Karlin, 1997), align sequences (Hughey and Krogh, 1996), and summarize protein families (Bateman et al., 2000). Second, the increasing availability of structural data has been a magnet for those with an interest in computational geometry, computer graphics, and algorithms for analyzing crystallographic data (Chapter 4) and NMR data (Chapter 5) to create credible molecular models. Structural bioinformatics has its roots in this second group. The development of molecular graphics was one of the first applications of computer

[1] http://www.rcsb.org.

[2] The International Society for Computational Biology (ISCB, http://www.iscb.org/) is the professional organization for bioinformatics; many developments in structural bioinformatics are reported in the journals and conferences associated with this society.

graphics (Langridge and Gomatos, 1963). The elucidation of the structure of DNA in the mid-1950s and the publication of the first protein crystal structures in the early 1960s created a demand for computerized methods for examining these complex molecules. At the same time, the need for computational algorithms to deconvolute X-ray crystallographic data and fit the resulting electron densities to the more manageable ball-and-stick models created a cadre of structural biologists who were very well versed in computational technologies. The challenges of interpreting NMR-derived distance constraints into three-dimensional structures further introduced computational technologies to biological structure. As the number of three-dimensional structures increased, the need to create methods for storing and disseminating this data led to the creation of the PDB, one of the earliest scientific databases.[1] In the past 10 years, we have seen a third wave of interest in biological problems from a group that was not engaged by the availability of 1D sequence data or 3D structural data. This third wave has arisen in response to the increased availability of RNA expression data and has captured the interest of computational scientists with an interest in statistical analysis and machine learning, particularly in clustering methodologies and classification techniques. The problems posed by these data are different from those seen in both sequence and structural analysis data. The recent introduction of high-throughput DNA sequencing technologies that produce short-length (25–50) snippets of DNA sequence is re-energizing the sequence analysis community with new challenges.

Structural bioinformatics is now in a renaissance with the success of the genome sequencing projects, the emergence of high-throughput methods for expression analysis, and identification of compounds via mass spectrometry. There are now organized efforts in structural genomics (Chapter 40) to collect and analyze macromolecular structures in a high-throughput manner (Teichmann, Chothia, and Gerstein, 1999; Teichmann, Murzin, and Chothia, 2001). These efforts include challenges in the selection of molecules to study, the robotic preparation and manipulation of samples to find crystallization conditions, the analysis of X-ray diffraction data, and the annotation of these structures as they are stored in databases (Section II). In addition, there have been advancements in the capabilities of NMR structure determination, which previously could only study proteins in a limited range of sizes. The solution of the malate synthase G complex from *E. coli* with 731 residues has pushed the frontier for NMR spectroscopy and suggests that NMR is having its own renaissance (Tugarinov et al., 2005). The PDB now has a critical mass of structures that allow (indeed require!) statistical analysis to learn the rules of how active and binding sites are constructed which allow us to develop knowledge-based methods for the prediction of structure and function. Finally, the emergence of this structural information, when linked to the increasing amount of genomic information and expression data, provides opportunities for linking structural information to other data sources to understand how cellular pathways and processes work at a molecular level.

Toward a High-Resolution Understanding of Biology. The great promise of structural bioinformatics is predicated on the belief that the availability of high-resolution structural information about biological systems will allow us to precisely reason about the function of these systems and the effects of modifications or perturbations. The genetic analyses can only associate genetic sequences with their functional consequences, whereas the structural biological analyses offer the additional promise of ultimate insight into the mechanisms of these consequences, and therefore a more profound understanding of how biological function follows from the structure. The promise for structural bioinformatics lies in four areas: (1) creating an infrastructure for building up structural models from

component parts, (2) gaining the ability to understand the design principles of proteins, so that new functionalities can be created, (3) learning how to design drugs efficiently based on structural knowledge of their target, and (4) catalyzing the development of simulation models that can give insight into function based on structural simulations. Each of these areas has already seen success, and the structural genomics projects promise to create data sets sufficient to catalyze accelerated progress in all these areas.

With respect to creating an infrastructure for modeling larger structural ensembles, we are already seeing the emergence of a new generation of structures larger by an order of magnitude than the structures submitted to the PDB a few years ago. Some achievements in recent years include (1) the elucidation of the structure of the bacterial ribosome (with more than 250,000 atoms) (Ban et al., 2000; Clemons Jr et al., 2001; Yusupov et al., 2001), (2) the publication of the RNA polymerase structure (with about 500,000 atoms) (Cramer et al., 2000), and (3) the increased ability to solve the structure of membrane proteins (transporters and receptors, in particular) that have proven technically difficult in the past. Each of these allows us to examine the principles of how a large number of component protein and nucleic acid structures can assemble to create macromolecular machines. With these successes, we can now target numerous other cellular ensembles for structural studies.

The design principles of proteins are now in reach both because we have a large ''training set'' of example proteins to study and because methods for structure prediction are beginning to allow us to identify structures that are unlikely to be stable. There have been preliminary successes in the design of four-helix bundle proteins (DeGrado, Regan, and Ho, 1987) and in the engineering of TIM barrels (Silverman, Balakrishnan, and Harbury, 2001). There has been interesting work in ''reverse folding'' in which a set of amino acid side chains is collected to stabilize a desired protein backbone conformation (Koehl and Levitt, 1999).

Rational drug design has not been the primary way for discovering major therapeutics (Chapters 27, 34 and 35). However, recent successes in this area give reason to expect that drug discovery projects will increasingly be structure based. One of the most famous examples of rational drug design was the creation of HIV protease inhibitors based on the known three-dimensional crystal structure (Kempf, 1994; Vacca, 1994). Methods for matching combinatorial libraries of chemicals against protein binding sites have matured and are in routine use at most pharmaceutical companies.

The simulation of biological macromolecular dynamics dates almost as far back as the elucidation of the first protein structure (Doniach and Eastman, 1999). These simulations are based on the integration of classical equations of motion and computation of electrostatic forces between atoms in a molecule. Methods for simulation now routinely include water molecules and are able to remain stable (the molecule does not fall apart) and reproduce experimental measurements with some fidelity. The simulation of larger ensembles and structural variants (such as based on known genetic variations in sequence) should lead to a more profound understanding of how structural properties produce functional behavior. The NIH has recognized the importance of simulation and created a national center devoted to physics-based simulation of biological structure (SIMBIOS, http://simbios.stanford.edu/).

Special Challenges in Computing with Structural Data

Structural bioinformatics must overcome some special challenges that are either not present or not dominant in other types of bioinformatics domains (such as the analysis of sequence

or microarray data). It is important to remember these challenges when assessing the opportunities in the field. They include the following:

- Structural data are not linear and therefore not easily amenable to algorithms based on strings. In addition to this obvious nonlinearity, there are nonlinear relationships between atoms (the forces are not linear). This means that most computations on structure need to either make approximations or be very expensive.
- The search space for most structural problems is continuous. Structures are represented generally by atomic Cartesian coordinates (or internal angular coordinates) that are continuous variables. Thus, there are infinite search spaces for algorithms attempting to assign atomic coordinate values. Many simplifications can be applied (such as lattice models for 3D structure; Hinds and Levitt, 1994), but these are attempts to manage the inherent continuous nature of these problems.
- There is a fundamental connection between molecular structure and physics. While this statement seems obvious and trivial, it means that when reduced representations, such as pseudoatoms (Wuthrich, Billeter, and Braun, 1983) or lattice models are applied, they become more difficult to relate to the underlying physics that governs the interactions. The need to keep structural calculations physically reasonable is an important constraint.
- Reasoning about structure requires visualization. As mentioned above, the creation of computer graphics was driven, in part, by the need of structural biologists to look at molecules (Chapter 9). This is both a benefit and a detriment; structure is well defined, and well-designed visualizations can provide insight into structural problems. However, graphical displays have a human user as a target and are not easily parsed or understood by computers, and thus represent something of a computational ''dead end.'' The need to have expressive data structures underlying these visualizations allows the information to be understood and analyzed by computer programs and thus opens the possibility of further downstream analysis.
- Structural data, like all biological data, can be noisy and imperfect. Despite some amazing successes in the elucidation of very high-resolution structures, the precision of our knowledge about many structures is likely to be limited by their flexibility, dynamics, or experimental noise (Chapters 14, 15, 37, and 38). Understanding the protein structural disorder may be critical for understanding the protein's function. Thus, we must be comfortable in reasoning about structures for which we only have partial knowledge.
- Protein and nucleic acid structures are generally conserved more than their associated sequence. Thus, sequences will accumulate mutations over time that may make identification of their similarities more difficult, while their structures may remain essentially identical. This is a challenge because sequence information is still much more abundant than structural information, and so for many molecules it is the sequence information that is readily available. The need to identify distant sequential similarities to gain structural insights can be a major challenge.
- Structural genomics will likely produce a large number of structures at the level of the domain—relatively well-defined modules that associate to form larger ensembles. The principles by which these domains associate and cooperatively function pose a major challenge to structural biology (Chapters 17, 18, 20, and 26).

- Finally, we must recognize that there is a major gap in our knowledge of a large fraction of proteins that are not globular and water soluble. In particular, membrane-bound and fibrous proteins are simply not well understood and structures have not been available in the numbers required to allow routine statistical and informatics approaches to their study. The importance of this shortcoming cannot be over emphasized, since these classes of proteins are among the most important ones for understanding a large number of cellular processes of great interest, including signal transduction, cytoskeletal dynamics, and cellular localizations and compartmentalization. Recently, some fascinating structures of membrane-bound transporter proteins, such as a zinc transporter (Lu and Fu, 2007), have improved our understanding of membrane protein structure (Chapter 36).

TECHNICAL CHALLENGES WITHIN STRUCTURAL BIOINFORMATICS

The scientific challenges within structural bioinformatics fall into two rough categories: the creation of methods to support structural biology and structural genomics and the creation of methods to elucidate new biological knowledge. This distinction is not absolute, but is useful for dividing much structural bioinformatics work. The support of experimental structural biology is currently an area of particular interest with the emergence of efforts in high-throughput structural genomics. Informatics approaches are required for many aspects of this enterprise, and can be briefly reviewed here:

- *Target Selection*: Structural genomics efforts with finite resources must select proteins to study carefully. Informatics methods are used to compare the database of existing structures and known sequences with potential targets to identify those that are most likely to add to our structural knowledge base. This selection can be informed by the expected novelty of the structure, and even its importance as reflected in the published literature (Linial and Yona, 2000). A critical part of target selection is the identification of domains within large proteins. Domains are often easier to study initially in isolation, and then in complexes. The definition of domains from sequence data alone is a challenging problem (Chapter 20).
- *Tracking Experimental Crystallization Trials*: One of the major bottlenecks in structural genomics is the discovery of crystallization conditions that work for proteins of interest. In addition to the obvious need of storing and tracking information on proteins, the conditions attempted, and the results, there is also an opportunity to apply machine learning methods to these data to extract rules that may help increase the yield of crystals based on previous experience (Hennessy et al., 2000). Until recently, the results of failed crystallization experiments were not generally available, thus making it difficult to apply automated machine learning methods to these data sets.
- *Analysis of Crystallographic Data*: A long-standing area of computation within structural biology is the algorithm for deconvoluting the X-ray diffraction pattern, which involves computing an inverse Fourier transform with partial information (i.e., with missing phase information). There is interest in *ab initio* methods for automating these computations, and success in this area reduces the number of heavy atom derivatives that must be created for structures of interest

(Gilmore, Dong, and Bricogne, 1998). Multiwavelength anomalous diffraction (MAD) (Hendrickson, 1991) is now the preferred method for solving the crystallographic phase problem. Over one-half of all structures are determined by MAD, a development in keeping with the availability of tunable synchrotron sources. Similarly, once the electron density is computed, there is a challenge in fitting the density to a standard ball-and-stick model of the atoms. While this has been done manually (with graphical computer assistance), there is interest in finding methods for using image processing techniques to automatically identify connected densities and match them to the known shape of protein backbone and side chain elements (Barr and Feigenbaum, 1982). Recent progress has been made on automated electron density map fitting and refinement (Chapter 4).

- *The Analysis of NMR Data*: NMR experiments provide complementary data to the crystallographic analyses. NMR experiments produce two (or higher) dimensional spectra for which each individual peak must be assigned to an atomic interaction. The automated analysis and assignment of atoms in these spectra is a difficult search problem, but the one in which progress has been made to accelerate the analysis of structure (Zimmerman and Montelione, 1995). Given a set of atomic proximities from NMR, we need methods to ''embed'' these distance measures into three-dimensional structures that satisfy these constraints. Distance geometry (Moré and Wu, 1999), restrained molecular dynamics (Bassolino-Klimas et al., 1996), and other nonlinear optimization methods have been developed for this purpose (Altman, 1993; Williams, Dugan, and Altman, 2001).
- *Assessment and Evaluation of Structures*: Given the results of a crystallographic or NMR structure determination effort, we must check the structures to be sure that they meet certain quality standards. Algorithms have been developed for assessing the basic chemistry of structural models and also for identifying active and binding sites in these structures (Laskowski et al., 1993; Feng, Westbrook, and Berman, 1998; Vaguine, Richelle, and Wodak, 1999). Computational methods are still needed for automatically annotating 3D structures with functional information, based on an understanding of how molecular properties aggregate in three dimensions to produce function (such as binding, catalysis, motion, and signal transduction) (Wei, Huang, and Altman, 1999, Chapter 5).
- *Storing Molecular Structures in Databases*: The storage of the results of structural genomics efforts is an important task, requiring data structures and organizations that facilitate the most common queries. Ideally, databases of structure will store not only the resulting model but also the raw data upon which it is based. The PDB (Chapter 11) is the major repository for three-dimensional structural information of proteins; the Nucleic Acids Database (NDB, Chapter 12) serves this function for nucleic acids. There is also an effort to store the raw data associated with crystallography in the PDB/NDB and the raw data associated with NMR in the BioMagResBank (BMRB).[3]
- *Correlating Molecular Structural Information with Structural and Functional Information Gained from Other Types of Experimentation*: In the end, we perform structural studies in order to get an insight into how the molecules work. Structural studies with crystallography and NMR are two methods that can be used to probe structure–function relationships. The integration of the results of these methods

[3] http://www.bmrb.wisc.edu

with other structural and functional data allows us to build comprehensive models of mechanism, specificity, and dynamics. A major bottleneck for using informatics methods for this integration is the lack of repositories of structural and functional data that can be accessed by computer programs doing systematic analyses. One exception is the noncrystallographic structural data on the 30S and 50S ribosomal subunits stored in the RiboWEB (http://riboweb.stanford.edu/), a knowledge base of ribosomal structural components that stores more than 8000 noncrystallographic structural and functional observations about the bacterial ribosome. It stores its information in structured ''information templates'' that are easily parsed by computer programs, thus making possible automated comparison and evaluation of structural models. For example, RiboWEB has been used to assess the compatibility of the published ribosomal crystal structures with over 1000 proximity measurements from cross-linking, chemical protection, and labeling experiments (collected during the past 25 years). Incompatibilities between these data and the crystal structures may suggest artifactual data or (more usefully) may suggest areas of important dynamic motion for the ribosome (Whirl-Carrillo et al., 2002).

Understanding the Structural Basis for Biological Phenomenon

Given the structural information created by efforts in X-ray crystallography and NMR, there is a wide range of analytic and scientific challenges to informatics. It is not possible to cover the full scope of activities, but they can be reviewed briefly to show the richness of opportunities in the analysis of structural data.

- *Visualization*: The creation of images of molecular structure remains a primary activity within structural biology (Chapter 9). The complexity of these molecules seems to demand novel display methods that are able to combine structural information with other information sources (such as electrostatic fields, the location of functional sites, and areas of structural or genetic variability). The issues for informatics include the creation of flexible software infrastructures for extending display capabilities and the use of novel methods for rapidly rendering complex molecular structures (Huang et al.,1996; Sanner et al., 1999).
- *Classification*: The database of known structures is already sufficiently large, making it necessary to cluster similar structures together to form families of proteins. These families are often aggregated into superfamilies, and indeed entire structural hierarchies have been created. The structural classification of proteins (SCOP; Chapter 17) is an example of a semiautomated classification of all protein structures (Murzin et al., 1995), and there have been numerous efforts to create automated classification—usually based on the pair-wise comparison of all structures to create a matrix of distances (Chapter 18; Holm and Sander, 1996; Orengo et al., 1997).
- *Prediction*: Despite the growth of the structural databases, the number of known three-dimensional structures has lagged far behind the availability of sequence information. Thus, the prediction of three-dimensional structure remains an area of keen interest. The Critical Assessment for Structure Prediction (CASP;[4] Chapter 28) meetings have provided a biennial forum for the comparison of methods

[4] http://predictioncenter.llnl.gov

for structure prediction. The main categories of prediction have been homology modeling (based on high sequence homology to a known structure; Chapter 30; Sánchez and Sali, 1997), threading (based on homology (Chapter 31); Bryant and Altschul, 1995), and *ab initio* prediction (based on no detectable homology; Chapter 32; Osguthorpe, 2000). The diversity of methods invented and evaluated is quite inspiring, and the resulting lessons about how proteins are put together have been significant.

- *Simulation*: The results of crystallographic studies (and to some extent, NMR studies) are primarily static structural models. However, the properties of these molecules that are of the greatest interest are often the results of their dynamic motions. The definition of energy functions that govern the folding of proteins and their subsequent stable dynamics has been an area of great interest since the first structure was determined. Unfortunately, the timescales on which macromolecular dynamics must be sampled (fractions of picoseconds) are much shorter than the timescale on which biologically important phenomena occur (from microseconds to seconds). Nevertheless, the availability of increasingly powerful computers and clever approximation and search methods is enabling molecular simulations of sufficient length and accuracy to emerge, making contributions to our understanding of protein function.[5] The associated computation of electrostatic fields of macromolecular structures (Chapter 24) has emerged as an important component of understanding molecular function (Sheinerman, Norel, and Honig, 1992).

It should be emphasized that although there has been primary focus on protein structures, with respect to the challenges outlined above, there is increasing interest in the same issues for RNA structure. The last decade has shown that the role of RNA molecules in the cell goes far beyond being a passive information carrier as messenger RNA. A large number of structured RNA molecules are involved in gene regulation (through RNA inhibition and other mechanisms), whose 3D structure is critical for understanding their function. The overall challenges for RNA structure are similar to proteins, but the details differ—RNA structure is dominated by electrostatics and not hydrophobic interactions, the secondary structure is easier to predict but offers a more limited repertoire for structural uses, and the molecules are more prone to finding stable misfolded states. Nonetheless, our understanding of structural biology will necessarily include the structure of RNA and RNA–protein complexes (Chapter 3, 12, and 33).

INTEGRATING STRUCTURAL DATA WITH OTHER DATA SOURCES

Structural bioinformatics has existed, in one form or another, since the determination of the first myoglobin structure. One could argue that the roots go back to the time when small molecular structure determination was introduced. In any case, the challenges for the field are clearly abundant and significant. As we look into the coming decade, it appears that a primary challenge in structural bioinformatics will be the integration of structural information with other biological information, to yield a higher resolution understanding

[5] The IBM BlueGene project (http://www.research.ibm.com/bluegene) is focused on the creation of a very large supercomputer, with the theoretical capability of simulating the folding of a small protein in about 1 year. The computer is being designed to have 10^{15} floating-point operations per second.

of biological function. The success of genome sequencing projects has created information about all the structures that are present in individual organisms, as well as both shared and unique features of these organisms. Even with the success of structural genomics projects, bioinformatics techniques will most likely be used to create homology models of most of these genomic components. The resulting structures will be studied with respect to how they interact and perform their functions. Similarly, the emergence of high-throughput expression measurements provides an opportunity to understand how the assembly of macromolecular structures is regulated (including the key structural machinery associated with transcription, translation, and degradation). Mass spectroscopic methods that allow the identification of structural modifications and variations (such as genetic mutation or post-translational modifications) will need to be integrated with structural models to understand how they alter functional characteristics. Cross-linking data, particularly *in vivo*, will provide valuable information about the physical association between macromolecules and ligands and the dynamics of molecular ensembles, thus helping us to create a structural portrait of a cell in three dimensions at near-atomic resolution (Tsutsui and Wintrode, 2007). Finally, cellular localization data will allow us to place three-dimensional molecular structures into compartments within the cell, as we build more complex models of how cells are organized structurally to optimize their function. This exciting activity will mark the next phase of structural bioinformatics—when the organization and physical structure of entire cells are understood and represented in computational models that provide insight into how thousands of structures within a cell work together to create the functions associated with life.

REFERENCES

Altman RB (1993): Probabilistic structure calculations: a three-dimensional tRNA structure from sequence correlation data. In: *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology,* July 6–9, 1993, Bethesda, MD. Menlo Park, CA: The AAAI Press.

Altman RB (1998): A curriculum for bioinformatics: the time is ripe. *Bioinformatics* 14:549–550.

Altman RB, Raychaudhuri S (2001): Whole-genome expression analysis: challenges beyond clustering. *Curr Opin Struct Biol* 11:340–347.

Ban N, Nissen P, Hansen J, Moore P, Steitz T (2000): The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289:878–879.

Barr A, Feigenbaum E (1982): Crysalis. In: *The Handbook of Artificial Intelligence*. Stanford, CA: HeurisTech Press, pp 124–133.

Bassolino-Klimas D, Tejero R, Krystek SR, Metzler WJ, Montelione GT, Bruccoleri RE (1996): Simulated annealing with restrained molecular dynamics using a flexible restraint potential: theory and evaluation with simulated NMR constraints. *Protein Sci* 5:593–603.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL (2000): The Pfam protein families database. *Nucleic Acids Res* 28:263–266.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000): The protein data bank. *Nucleic Acids Res* 28:235–242.

Bryant SH, Altschul SF (1995): Statistics of sequence–structure threading. *Curr Opin Struct Biol* 5:236–244.

Burge C, Karlin S (1997): Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268:78–94.

Clemons WM Jr, Brodersen DE, McCutcheon JP, May JLC, Carter AP, Morgan-Warren RJ, Wimberly BT, Ramakrishnan V (2001): Crystal structure of the 30S ribosomal subunit from *Thermus thermophilus*: purification, crystallization and structure determination. *J Mol Biol* 310:827–843.

Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, Thompson NE, Burgess RR, Edwards AE, David PR, Kornberg RD (2000): Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* 288:640–649.

DeGrado W, Regan L, Ho S (1987): The design of a four-helix bundle protein. *Cold Spring Harbor Symp Quant Biol* 52:521–526.

Doniach S, Eastman P (1999): Protein dynamics simulations from nanoseconds to microseconds. *Curr Opin Struct Biol* 9:157–163.

Durbin R, Krogh A, Mitchison G, Eddy S (1998): *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Feng Z, Westbrook J, Berman HM (1998): NUCheck. Rutgers Publication NDB-407. New Brunswick, NJ: Rutgers University.

Gilmore CJ, Dong W, Bricogne G (1998): A multisolution method of phase determination by combined maximisation of entropy and likelihood. VI. The use of error-correcting codes as a source of phase permutation and their application to the phase problem in powder, electron and macromolecular crystallography. *Acta Crystallogr A* 55:70–83.

Gusfield D (1997): *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.

Hendrickson WA (1991): Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* New York, NY 254:51–58.

Hennessy D, Buchanan B, Subramanian D, Wilkosz PA, Rosenberg JM (2000): Statistical methods for the objective design of screening procedures for macromolecular crystallization. *Acta Crystallogr D* 56:817–827.

Hinds D, Levitt M (1994): Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol* 243:668–682.

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

Huang CC, Couch GS, Pettersen EF, Ferrin TE (1996): Chimera: an extensible molecular modeling application constructed using standard components. In: Hunter L, Klein TE, editors. *Pacific Symposium on Biocomputing,* January 3–6, 1996, USA, Hawaii. Singapore: World Scientific Publishing, p 724.

Hughey R, Krogh A (1996): Hidden Markov models for sequence analysis: extension and analysis of the basic method. *CABIOS* 12:95–107.

Kempf D (1994): Design of symmetry-based, peptidomimetic inhibitors of human immunodeficiency virus protease. *Methods Enzymol* 241:334–354.

Koehl P, Levitt M (1999): Structure-based conformational preferences of amino acids. *Proc Natl Acad Sci USA* 96:12524–12529.

Langridge R, Gomatos PJ (1963): The structure of RNA. *Science* 141:694–698.

Laskowski RA, McArthur MW, Moss DS, Thornton JM (1993): PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 265:283–291.

Linial M, Yona G (2000): Methodologies for target selection in structural genomics. *Prog Biophys Mol Biol* 73:297–320.

Lu M, Fu D (2007): Structure of the zinc transporter YiiP. *Science* New York, NY 317: 1746–1748.

Miranker AD (2000): Protein complexes and analysis of their assembly by mass spectrometry. *Curr Opin Struct Biol* 10:601–606.

Moré J, Wu Z (1999): Distance geometry optimization for protein structures. *J Global Optim* 15:219–234.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995): SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997): CATH: a hierarchic classification of protein domain structures. *Structure* 5(8):1093–1108.

Osguthorpe DJ (2000): *Ab initio* protein folding. *Curr Opin Struct Biol* 10:146–152.

Sánchez R, Sali A (1997): Advances in comparative protein-structure modelling. *Curr Opin Struct Biol* 7:206–214.

Sanner MF, Duncan BS, Carrillo CJ, Olson AJ (1999): Integrating computation and visualization for biomolecular analysis: an example using PYTHON and AVS. In: Altman RB, Lauderdale K, Dunker AK, Hunter L, Klein TE, editors. *Pacific Symposium on Biocomputing*, Hawaii, USA, Singapore: World Scientific Publishing.

Sheinerman FB, Norel R, Honig B (1992): Electrostatic aspects of protein–protein interactions. *Curr Opin Struct Biol* 10:153–159.

Silverman JA, Balakrishnan R, Harbury PB (2001): Reverse engineering the (beta/alpha)8 barrel fold. *Proc Natl Acad Sci USA* 98:3092–3097.

Teichmann SA, Chothia C, Gerstein M (1999): Advances in structural genomics. *Curr Opin Struct Biol* 9:390–399.

Teichmann SA, Murzin AG, Chothia C (2001): Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol* 11:354–363.

Tugarinov V, Choy W-Y, Orekhov VY, Kay LE (2005): Solution MR-derived global fold of a monomeric 82-kDa enzyme. *Proc Natl Acad Sci USA* 102(3): 622–627.

Tsutsui Y, Wintrode PL (2007): Hydrogen/deuterium exchange-mass spectrometry: a powerful tool for probing protein structure, dynamics and interactions. Current medicinal chemistry 14: 2344–2358.

The Genome International Sequencing Consortium (2001): Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Vacca J (1994): Design of tight-binding human immunodeficiency virus type 1 protease inhibitors. *Methods Enzymol* 241:331–334.

Vaguine AA, Richelle J, Wodak J (1999): SFCheck: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr D* 55:191–205.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutto GG, Smith HO, Yandell M, et al. (2001): The sequence of the human genome. *Science* 291:1304–1351.

Wei L, Huang ES, Altman RB (1999): Are predicted structures good enough to preserve functional sites. *Structure* 7:643–650.

Whirl-Carrillo M, Gabashvili IS, Bada M, Banatao DR, Altman RB (2002): Mining biochemical information: lessons taught by the ribosome. *RNA* 8:279–289.

Williams GA, Dugan JM, Altman RB (2001): Constrained global optimization for estimating molecular structure from atomic distances. *J Comput Biol* 8:523–547.

Wuthrich K, Billeter M, Braun W (1983): Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton–proton distance constraints with nuclear magnetic resonance. *J Mol Biol* 169:949–961.

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JHD, Noller HF (2001): Crystal structure of the ribosome at 5.5 Å resolution. *Science* 292:883–896.

Zimmerman DE, Montelione GT (1995): Automated analysis of nuclear magnetic resonance assignments for proteins. *Curr Opin Struct Biol* 5:664–673.