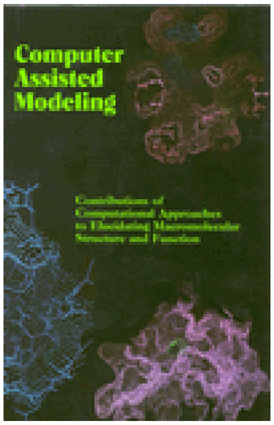


Computer Assisted Modeling: Contributions of Computational Approaches to Elucidating Macromolecular Structure and Function

Committee on Computer-Assisted Modeling, National Research Council

ISBN: 0-309-56829-3, 186 pages, 6 x 9, (1987)

COMPUTER-ASSISTED MODELING

Contributions of Computational Approaches to Elucidating Macromolecular Structure and Function

Committee on Computer-Assisted Modeling
Board on Basic Biology
Commission on Life Sciences
National Research Council

National Academy Press
Washington, D.C. 1987

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Samuel O. Thier is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

This study by the Board on Basic Biology was conducted under Contract No. DE-FGO1-86ER60457 with the U.S. Department of Energy.

Copies of this report can be obtained from the National Academy Press , 2101 Constitution Avenue, N.W. ,Washington, DC 20418 , for $3.00 per copy prepaid. Supplies are limited.

Printed in the United States of America

Cover: Connolly-Richards Solvent-accessible surfaces of trypsin, thymidylate synthase, and carbonic anhydrase. Coordinates from the Protein Data Bank, Brookhaven National Laboratory, and R. Stroud, University of California/San Francisco. Photographs taken by R. Desjarlais and B. Shoichet, University of California/San Francisco, using the facilities of the Computer Graphics Laboratory.

COMMITTEE ON COMPUTER-ASSISTED MODELING

IRWIN D. KUNTZ (Chairman), University of California, San Francisco
DAVID R. DAVIES, National Institutes of Health, Bethesda, Maryland
RICHARD E. DICKERSON, University of California, Los Angeles
RUSSELL F. DOOLITTLE, University of California, San Diego
RICHARD J. FELDMANN, National Institutes of Health, Bethesda, Maryland
JAN HERMANS, University of North Carolina, Chapel Hill
YVONNE C. MARTIN, Abbott Laboratories, Abbott Park, Illinois
JAMES H. PRESTEGARD, Yale University, New Haven, Connecticut
PETER J. ROSSKY, University of Texas, Austin
HAROLD A. SCHERAGA, Cornell University, Ithaca, New York
DENNIS H. SMITH, Molecular Design Ltd., San Leandro, California
CHARLES C. SWEELEY, Michigan State University, East Lansing
IGNACIO TINOCO, University of California, Berkeley
WALTER G. ROSEN, *Senior Program Officer*
SUSAN WALTON, *Editor*
LINDA MILLER POORE, *Senior Secretary*

BOARD ON BASIC BIOLOGY

COMMISSION ON LIFE SCIENCES

Preface

The committee, asked to provide an assessment of computer-assisted modeling of molecular structure, has highlighted the signal successes and the significant limitations for a broad panoply of technologies and has projected plausible paths of development over the next decade.

As with any assessment of such scope, differing opinions about present or future prospects were expressed. The conclusions and recommendations, however, represent a consensus of our views of the present status of computational efforts in this field.

The committee's task was made easier by colleagues who generously provided us with the benefit of their expertise. We wish particularly to thank Peter Goodford, William Jorgensen, and Andrew McCammon.

The committee is indebted to the excellent National Research Council staff whose work greatly expedited the production of this report. Special thanks are due to Linda Poore, Susan Walton, and, particularly, to Walter Rosen for unfailing help and guidance.

Irwin D. Kuntz
Chairman

Contents

1.

Executive Summary

In much of biology, the search for understanding the relation between structure and function is now taking place at the macromolecular level. Proteins, nucleic acids, and polysaccharides are macromolecules—polymers formed from families of simpler subunits. Because of their size and complexity, the polymers are capable of both inter- and intramolecular interactions. These interactions confer upon the polymers distinctive three-dimensional shapes. These tertiary configurations, in turn, determine the function of the macromolecule.

A molecular view of biological function has already led to significant advances. The conceptual breakthrough that led to our present mastery of genetic control of protein synthesis was the discovery that the nucleotide sequence in nucleic acids codes for the amino acid sequence of the protein being synthesized. The Rosetta Stone of molecular genetics was the elucidation of the actual code, whereby the sequence of trimers (codons) in the nucleic acid can be translated to the sequence of amino acids in the protein.

Amino acid sequences in proteins can be determined directly through chemical analysis or indirectly by determining the base sequence in the parent DNA. These two methods have been used to describe the primary structures (amino acid sequences) of substantial numbers of proteins. But to understand the function of a protein, we need to know more than its primary structure. Except, perhaps, for some structural elements, proteins are not found in nature as simple chains of amino acids. In their biologically active form, they are folded upon themselves, forming complex three-dimensional structures. Their shape determines their biological activity.

Currently, determining the three-dimensional structure of macromolecules depends on the interplay of various experimental and theoretical approaches, particularly x-ray diffraction and two-dimensional nuclear magnetic resonance (NMR), all of which involve the use of computers. At their most fundamental level, computers simply function to solve mathematical equations at very high speeds. Their speed makes it possible to accomplish in fractions of a second tasks that would take a human orders of magnitude longer, even with the assistance of mechanical calculators. Computers are also used to generate graphic representations of the three.dimensional structures of molecules, thereby aiding comprehension and largely eliminating the need to construct physical models—another laborious and time-consuming process. These graphic representations provide an important stimulus to the development of new ideas about the way macromolecules function.

Computers have become so inextricably involved in empirical studies of three-dimensional macromolecular structure that mathematical modeling, or theory, and experimental approaches are interrelated aspects of a single enterprise. The experimental methods, such as x-ray crystallography, NMR spectroscopy, and mass spectrometry provide the data with which to construct a mathematical model that can account for the electron density distributions, bond angles, bond energies, and other observed structural properties of the molecule. Conversely, the mathematical model must generate a structure that agrees with the experimental data. The interplay between the two is continual; theoretical models are modified repeatedly to improve their fit with experimental data, and theoretical results help in the interpretation and planning of experiments.

The potential practical applications of these techniques are myriad. When can the payoff be expected? What are the anticipated needs in terms of hardware, software, and human resources? To what extent should the effort be centralized? Is present funding adequate or should it be increased?

The task of the committee was to examine these questions as they relate to two major realms of theoretical molecular biology. The first is the prediction of tertiary structure from primary structure and other physical/chemical data. The second is the prediction of biological activity from tertiary structure and related data and theory.

PROTEINS

Primary Structure

More than 5,000 protein amino acid sequences have been reported, most of which were inferred from the DNA sequences that encode them. Although the collection is redundant (same protein from different species) and definitely biased (many human and few plant sequences, for example), several patterns stand out. The most prominent is that the number of different types of protein is not endless. It is clear that most proteins belong to identifiable families, easily recognized by their amino acid sequences alone. But surprisingly, the same families of protein primary structures are showing up in proteins in quite different settings.

At this point it is not possible to determine, with accuracy, a three-dimensional structure of a protein using only the amino acid sequence. However, recognition of patterns in structure-sequence correlations holds great promise in this area.

Predicting Secondary Structure from Amino Acid Sequences

The methods used to predict secondary structure from amino acid sequences have been (1) calculation of the energies of the major conformers for a given sequence, (2) statistical analysis of known structures, and (3) modeling. Although these methods have had some success and should continue to improve with an increasing data base, they have limited accuracy.

It is now computationally feasible to calculate energies for conformations of short peptides with or without solvent. The near-term developments in this area are most likely to be incremental improvements. Computer speed will continue to increase substantially. Data bases will continue to grow at least linearly. Experiments on the structural consequences of modifying amino acids are beginning to be reported in significant numbers. More powerful statistical and modeling efforts are under development.

The situation is less positive for larger peptides and proteins. After 10 years, however, we might well see major improvements in our ability to correlate sequences and secondary structure. Today's goal of correctly predicting every major feature in a new sequence is a plausible target for that time. However, we will need a major conceptual or computational breakthrough before it will be possible to specify accurately the secondary structural configuration of each amino acid in a protein.

Deriving Three-dimensional Structure and Function from Homology

Methods for identifying homologies and for determining sequence alignments and homologies are powerful tools to relate the structures and, potentially, the functions of two or more biopolymers. The analyses of patterns in sequences is a problem in symbolic, not numeric, computation. Languages that support symbol manipulation and pattern matching primitives such as C and LISP are often the languages of choice.

Many of the techniques used to infer higher order structural information in patterns—such as secondary structural analysis—are empirical. Sequence-based methods alone do not take full advantage of the information available in the primary structures of biopolymers. For example, different nucleotide patterns may code for the same protein sequence; different protein sequences may share very similar function.

Predicting Nucleic Acid Structures from Sequence

Computer programs used to predict RNA conformation from sequence have a limited goal and limited success. The goal is to calculate secondary structure only—to specify which bases are paired. The procedure uses experimental thermodynamic data on double-strand formation in synthetic RNA oligonucleotides. A dynamic programming algorithm considers all possible base pairs in the RNA and calculates the free energies of the corresponding structures. The free energy of a structure is assumed to be the sum of free energies of its constituent substructures (single-stranded regions, double-stranded regions, bulges, hairpin loops, and interior loops). The lowest free-energy structure is the predicted secondary structure.

Computer programs of the future will require much more detailed knowledge of thermodynamics of local regions of an RNA molecule. The extensive thermodynamic data needed to predict secondary structure correctly will most likely come from computer interpolation and extrapolation of limited data measured on synthetic oligonucleotides in a few solvents.

Proposals of methods to fold possible secondary structures into three-dimensional structures and calculate their energies are in very early stages. Prediction of secondary structures in RNA is at about the same stage as it is in proteins. However, prediction of tertiary structure in RNA is far behind similar prediction in proteins.

Rapid and efficient progress in this area will require:

- effective methods for crystallizing RNA oligonucleotides and naturally occurring RNA molecules other than transfer RNA;
- NMR methods that can provide conformations for RNA molecules that contain from 10 to 100 nucleotides;
- computer programs that can reproduce and extrapolate the experimental results.

The higher charge densities in nucleic acids (one per nucleotide) require special care in the correct treatment of solvent and ionic effects in the computer programs.

Tertiary Structure from X-ray Crystallography

Today, several hundred proteins have been analyzed by x-ray diffraction and their three-dimensional structures catalogued, and their number is growing substantially. This knowledge of molecular structure, together with the amino acid and gene sequence data, enable us to study the mechanisms of action of these proteins at the molecular level. Two-dimensional NMR techniques are a valuable complement to x-ray diffraction for relatively small molecules (molecular weight less than 10,000), but for the foreseeable future, crystal structure analysis will be the principal experimental source of structural data for enzymes, nucleic acid binding proteins, antibodies, and other proteins involved in the immune response or intercellular communication.

Determining the three-dimensional structure of a biological macromolecule involves several clearly defined steps. First, crystals of suitable size and diffraction properties must be prepared. Next, x-ray diffraction data must be collected for these crystals and also, typically, for some heavy atom derivatives of the crystals. These data can then be assembled by a computational process that yields an electron density map. This map must now be fitted with a polypeptide chain of the appropriate amino acid sequence. Because the map is of less-than-atomic resolution and because it also contains errors in the phase determination, considerable skill is required to obtain the best fit. The resulting protein model must then be refined to remove as many as possible of the errors present in the map as well as those introduced by the fitting process. Computers play an essential role in most of these steps.

The availability of new instrumentation that allows crystallographers to produce in a few days data that previously took weeks or months of labor-intensive work is revolutionizing protein crystallography at an opportune time. In recent years, developments in genetic engineering have made it possible to produce large quantities of rare proteins and to use site-directed mutagenesis to answer structural questions.

X-ray diffraction experiments have provided structures for double-stranded DNA, protein-DNA complexes, and DNA-small molecule compounds in crystals. Computer modeling is needed to extrapolate these results to more biological environments and to other complexes. An obvious application of such insight is to design more specific and more effective antibiotics. In general, we would like to be able to design molecules that can start or stop the expression of any gene in any DNA. The key to achieving this is computer-assisted calculation used in close collaboration with experimental observation.

Nuclear Magnetic Resonance

NMR is another important source for structural data, and its use is developing very rapidly. NMR results can be compared directly with theoretical (mathematical) modeling and with structures derived from x-ray crystallography. NMR has been applied to macromolecules in aqueous media and to a limited extent in other environments, such as those that approximate biological membranes. It is at its best when used to explore changes in preferred structure in response to environmental or structural perturbations.

In this sense, NMR can be important in extrapolating to other environments structural data obtained by other techniques. It also is well suited to exploring changes in structure when comparing homologous series of macromolecules, such as a series produced by site-specific mutagenesis.

More recently, technological advances have extended the applicability of NMR to solids, oriented phases, and even to total structure determination of molecules in solution. The latter use has attracted the most attention and currently has the greatest potential to affect computer-assisted modeling efforts.

The major limitation on the use of NMR methods to determine structure is the restriction on molecular weight. Current applications require proteins of MW 10,000 and less. A second limitation stems from the restricted range of measurable distances, less than 4Å. A third limitation arises because of underlying assumptions about the rigidity of macromolecules. All these limitations are likely to be overcome in time, but doing so will require advances in NMR methodology and computational capacity.

Improved computational and molecular modeling facilities could promote the use of NMR to determine structure in several ways. Processing and analyzing structural data for macro-molecules of MW 10,000 is far more time-consuming than is ac quiring the data. Each phase of this operation could be improved. Data are normally collected as a two-dimensional time domain set, and processing involves Fourier transformation to a frequency do main set. These processes are now handled by array processors associated with instrument computers, with a moderate investment in time (one hour per process). However, alternative methods of processing, including linear decomposition and maximum entropy methods, may be better in terms of the signal-to-noise ratio and may be more compatible with automating the analysis. Such methods take far more computer time and may become practical only on supercomputers.

Some efforts are underway to use semiautomated pattern recognition and expert system strategies to determine three-dimensional structure, but these will require substantial investments in programming and computer hardware.

An investment in programming is obviously warranted. This investment could be used best if we acknowledge that, in the future, we may need to accommodate types of data not used today. Some data may come from other structural methods applicable to solids and oriented specimens. Other data may come from entirely different methods, such as tunneling microscopy. Thus, programs need not be directed specifically for use with NMR data, but should, if possible, accommodate structural data from a variety of sources.

In summary, current NMR methods of determining structure are applicable to a variety of biologically important molecules that are less than MW 10,000. Data production in this size range will be greatly enhanced by better computational facilities, high field spectrometers, and modeling programs that aim for compatibility with experimental constraints of the form provided by NMR. NMR data should be meshed with data gathered through other methods of determining structure. The range of molecules accessible by these methods is likely to increase by a factor of two over the next five years. The rate of data production is likely to increase even more quickly as we improve structure determination protocols and as high-field spectrometers become more generally available.

TERTIARY STRUCTURE FROM THEORY

Energy Optimization

According to the thermodynamic hypothesis, the amino acid sequence of a protein determines its three-dimensional structure in a given medium as the thermodynamically most stable structure.

To identify this structure requires some kind of optimization strategy, which, in turn, requires procedures to generate arbitrary three-dimensional conformations of a polypeptide chain, compute the free energy of the system for each conformation, and then alter the conformation so that it ultimately corresponds to the global minimum of the free energy.

Although algorithms are available for minimizing an energy function of many variables, there are no efficient ones to use for passing from one local minimum, over a potential energy barrier, to the next local minimum—and ultimately to the global minimum in a many-dimensional surface. Thus, minimization leads to the nearest local minimum, where the procedure is trapped. This trapping in a local, rather than the global, minimum is referred to as the “multiple-minima problem.” A variety of procedures are being developed to overcome this problem, including approximations that initially place the system in a broad potential energy well in which the more sharply defined global minimum lies.

Although supercomputers will more adequately cover conformational space, workers in this field will need more time on these machines to achieve greater efficiency. Parallel processing offers a breakthrough, but will require that more software be developed to take advantage of the new hardware. With new hardware and software, it should be possible to surmount the major hurdle created by the multiple-minima problem. However, bottlenecks may develop as attempts are made to apply procedures that work on 20-residue segments to proteins containing 100 to 200 residues.

Homology

Proteins can be categorized by families. Evidence for this comes from protein sequence homology and from the architectural similarity in tertiary structures of homologous proteins as established by x-ray and NMR methods. A family of proteins can be modeled if several conditions are fulfilled. First and most important, the structure of at least one member of the family must be known. Second, the protein to be modeled must be sufficiently homologous to the known protein. Many proteins have been modeled over the past five years, and the general consensus is that if two proteins share at least 30 percent similarity, it is reasonable to use computer graphics and energy modeling to propose the unknown structure from the known.

Molecular Dynamics

Molecular dynamics simulations apply Newton's equations of motion to the atoms of one or several molecules. Newton's equations relate three independent quantities: time, conformation (three-dimensional atomic coordinates), and potential energy.

Molecular dynamics simulation allows us to estimate theoretical mean atomic positions and deviations from the mean; rates of motion and conformation change; and ensemble averages, including thermodynamic functions such as energy, enthalpy, specific heat, and free energy. Although simple in concept, molecular dynamics simulations were not practical until the advent of high-speed computers. The time is approaching when molecular dynamics calculations will produce useful predictions of the structure, dynamics, and thermodynamics of proteins, nucleic acids, and complexes of these macromolecules with one another and other molecules.

The simulation requires two initial pieces of information: a starting conformation and a potential energy function or forcefield. For a protein, the starting conformation must be firmly based on experimental observation. The forcefield is often identical to that used in molecular mechanics.

The forcefield is a very simple empirical approximation of the underlying physics, which properly should be expressed in terms of quantum mechanics, but is totally unmanageable in that form. Parameters of the forcefields currently in use have been proposed on the basis of various experimental data and, to some extent, on theoretical considerations.

Recently developed forcefields for water-water and water-protein interactions permit the simulation of the dynamics of proteins in solution. This capacity is a prerequisite for modeling events at the protein surface, including most interactions of proteins with other molecules.

The simulations of molecular dynamics of proteins require careful adjustment of starting configurations and simulation parameters. The limiting factor is always the available computing power. For example, calculation of molecular dynamics simulations of the motion of a protein over a 10^{-9}-second time interval takes roughly a month of computer time on a CRAY. Making additional computer time available to those working in the field will help in the development/application of more detailed forcefields, produce longer simulations, encourage the simulation of larger systems that pose new physical and biological questions, and promote the application of new, more time-consuming dynamics methods to be used to ask different questions about the system.

Molecular dynamics simulations show considerable promise of being able to more accurately depict the structures that are proposed on the basis of incomplete information, particularly from two-dimensional NMR. Such reiterations are thought to be the best method of investigating the atomic details of macromolecular motion.

Thermodynamics

Physicists have known, in principle, how to calculate equilibrium thermodynamic properties from molecular dynamics calculations for considerable time. Only very recently have these techniques been applied to proteins, but their use has already shifted the emphasis of the molecular dynamics simulation field to calculations of free-energy differences. Several factors explain the great current interest in this application. The most important are the magnitude and precision of available experimental data for a variety of equilibrium-involving biological macromolecules and the unexpectedly excellent theoretical estimates that were and still are produced by the simulations.

Progress in free-energy simulations, although potentially very rapid, is severely limited by available computer time. To realize the possibilities already identified will require a radical increase of computer access for molecular dynamics studies. An immediate 10-fold increase does not seem an extravagant objective if we duplicate existing hardware that is already programmed and inexpensive.

In contrast to the folding problem, the problem of computer modeling of the dynamics of protein interactions can be tackled in a series of small, increasingly complex steps, each of which solves a discrete problem of immediate biochemical interest, yet also adds to our insight and experience with the broader picture.

Beyond the need for adequate computer time, two other needs must be met. One is the need for better forcefields, particularly for nucleic acids and carbohydrates; the second is the need for improved molecular dynamic techniques designed to overcome some of the intrinsic imperfections of existing forcefields. One possible impediment to this is the apparent trend toward the development and commercialization of proprietary forcefields much like the trend toward proprietary software. This trend seems counter to the best interests of science because it limits access and precludes rigorous testing of results.

Solvent Effects

Biomolecular systems function in vivo in environments of aqueous solutions or in a membrane. The aqueous environment includes solvent as well as a substantial component that consists of various ions. Because of the potentially relatively strong interactions of these components with each other and with the macromolecular species, this environment can contribute substantially to the observed state of a macromolecule in solution. The membrane environment differs in that the charged groups and electrically neutral regions are spatially separated. A quantitative treatment of biopolymer structure and function cannot be expected to succeed unless we pay attention to the molecular role of the environment.

The ability to adequately test predictions made from theoretical calculations is an element of overriding importance in the future of this aspect of modeling. This can occur at two levels: first and most important in comparing theory and experiment and second in comparing results obtained through convenient but approximate theory with those that follow from accurate theoretical treatment. The first level is essential for accuracy and the second for the future development of viable theoretical methods for increasingly complex systems. Therefore, we should continue to encourage*both* experiment and theory for*both* macromolecular and smaller model compounds.

The evident rapid progress in the ability to describe the environmental aspects of biopolymer systems justifies our optimism that this element of biomolecular modeling will not impede development of useful predictive methods. For the most challenging aspects, however, we are at least several years away from being able to accurately mimic environmental effects of solution.

ANALYSIS AND DESIGN OF DRUGS

Central questions regarding function include (1) which aspects of protein structure and dynamics are responsible for the often enormous enhancements of rates of reaction of enzymes over the rates in water and (2) how the signal of the binding of a ligand or quantum of light is transduced into a physiological response such as an increase in blood pressure or a change in mental functioning. The questions are at best incompletely answered.

In principle, the information needed to predict the function of a biological macromolecule is encoded in its three-dimensional structure. The problem is how to decode the rules that govern the relationship between structure and function. The tools of molecular dynamics promise to offer insight into this problem, but resources must be made available.

The prospect of computer-assisted drug design based on the three-dimensional structure of the target biomolecule has recently received much attention in the scientific literature. Many medicinal chemists believe that their field is poised to undergo a revolution as dramatic as that of the 1950s and 1960s that transformed organic chemistry from a descriptive to a predictive science.

This revolution presupposes that we can (or soon will be able to) predict the functions of macromolecules from their structure. In particular, this would require predicting whether a protein can recognize and bind a ligand and predicting the structure of the optimum ligand. Beyond that, however, we would need to be able to predict how a protein recognizes and interacts with other macromolecules to alter its own and their functions. These insights will be a direct consequence of the theoretical studies discussed in this report.

The design of a new drug from theoretical principles must somehow incorporate the possible interaction of the proposed ligand with all other macromolecules of the body. Quantitative Structure/Activity Relations (QSAR) is a logical complement to the more structure-based computer methodologies. QSAR is based on computing the activity of molecules from the properties and activities of substituents. This tool can be used to model the potential whole-animal activity of new ligands and perhaps to search for unanticipated interactions with other macromolecules.

The recent interest in computer-assisted drug design arose because at last the scientist has available the elements of each of the important tools needed for such an activity. Two types of computer hardware are necessary: high speed color graphics and affordable but powerful minicomputers dedicated to modeling. Data on the three-dimensional structure of proteins are becoming available at an increasing rate as we improve our understanding of some of the relationships between structure and function of proteins. Finally, software is also available for displaying the molecules and modeling the energetics and thermodynamics of the binding. Equally important, specialized graphics tools for molecular design have been developed. Some of these arose from the related activity of “docking” a known ligand into a protein.

We must understand the relation between structure and function if we are to design agents that alter function by changing structure. Ultimately, we expect to be able to predict the biotransformations of small molecules from the structures of the enzymes involved, but we cannot do so now.

COMPLEX CARBOHYDRATES

Complex carbohydrates occur everywhere in animals, plants, and bacteria. The enzymes involved in glycoconjugate biosynthesis are glycosyltransferases that catalyze the transfer of sugar residues from the sugar nucleotides to the nonreducing end of a growing carbohydrate chain. The distinction between this process and protein synthesis is key; the latter occurs on a template of messenger RNA and is therefore determined by the genetic code for a single structural gene. In sharp contrast, glycoconjugate synthesis is accomplished by adding sugar units in a stepwise manner, with a different enzyme used for each step. The current state of knowledge does not indicate that a single DNA sequence determines the primary structure of the complex carbohydrate.

It is not yet possible to predict the primary structures of complex carbohydrates from DNA sequences, and the three-dimensional structures of glycoproteins, glycosphingolipids, and other complex carbohydrate-containing molecules can never be completely predicted without analyzing the two-dimensional structures of the carbohydrates. A complete understanding of the interactions between carbohydrates and proteins (enzymes, lectins, antibodies, and cell surface receptors) will depend on the generation of accurate three-dimensional structures of both kinds of molecules.

Of the three major classes of complex biological molecules, we have the least information at the atomic level about the three-dimensional structure of carbohydrates. Because no large carbohydrates have been crystallized, we have no data on relevant crystal structure, other than data on simple monomers to trimers, upon which to model classical or semiempirical quantum mechanical calculations. Adequate computer time, including access to appropriate parallel processors is an important consideration in support of this research.

Configurations that consist of more than one macromolecule may interact as a whole in biological phenomena such as catalysis by many enzymes, binding at a cell surface, and signal transduction across cell membranes. Hybrid systems involving complex carbohydrates, proteins and nucleic acids are important in protein and nucleic acid synthesis, repair, and regulation. We believe it will be possible in the near future to use structural methods to characterize at least parts of these systems. Computer modeling of such supramolecular structures will be necessary if we are to gain a deeper understanding of how biological materials are organized to carry out complex functional tasks.

ROLE OF COMPUTERS

Computers clearly play an essential role at virtually every stage and in virtually every process, theoretical and experimental, in determining the three-dimensional structures and biological activity of macromolecules. Both personal and laboratory computers and large, mainframe computers are used.

Computer-assisted mathematical (theoretical) modeling is augmented by computer-generated three-dimensional graphic representations of molecules. Modeling from theory is generally tightly coupled to experimental approaches such as x-ray crystallography, NMR, and monomer sequencing. The interplay of theory and experiment results in the increasing refinement of the theory-based models, which in turn can be used to predict the behavior and properties of the actual molecules.

Availability of computer software and access to computer time and computer-based data banks are often the factors that limit the rate of progress in structural biology. The conclusion is inevitable that progress toward fuller understanding of macromolecular structure and function will be accelerated if computer-based activities receive greater support. The computer facilities of the national laboratories have the attributes required to lead in providing computer-based support for both theoretical and empirical approaches to understanding macromolecular structure and function.

Progress in structural biology is likely to produce rapid advances in biotechnology, drug design, toxicology, and medicine, along with advances in understanding of basic biological processes, including heredity and development.

Our recommendations for dealing with these issues are detailed inChapter 10and summarized in the following section.

CONCLUSIONS

The conclusions we draw from our examination of the facts related to the questions contained in our charge may be summarized as follows:

Important advances in the understanding of macromolecular structure and function have been and will continue to be gained from the application of techniques using computers. To maximize the speed of progress toward a better understanding of protein foldings, macromolecular interactions and functions, impediments and limitations to the effective study of these matters must be identified and dealt with. The areas requiring attention include the need for readily available data banks of protein and nucleic acid sequences as well as model-derived structures; improved capability of and access to supercomputers; provisions of educational opportunities in the area; and use of the most appropriate physical and intellectual resources for the performance of research.

Our recommendations for dealing with these issues are detailed inChapter 10and summarized in the following section.

RECOMMENDATIONS

- A radical new policy on data banking of protein and nucleic acid sequences is required. A permanent National Sequence Data Bank should be put in place as soon as possible. A standing advisory committee of users should be appointed by a consortium drawn from the National Institutes of Health (NIH), National Science Foundation (NSF), and Department of Energy (DOE). Whether the new facility should be allied with a national laboratory, or with the National Library of Medicine, or should be a completely new academic or commercial enterprise remains to be determined.
- Support for the archiving of coordinate and model-derived structures should continue. Inclusion of data from new methods of structural analysis should be encouraged.
- We recommend in the strongest terms expanding the supercomputer initiative, funding of computer networks, improving access by the scientific community to the existing supercomputer centers at the national laboratories, upgrading those centers, and providing individual research grants for purchasing new computers. DOE should work closely with the NSF and NIH to provide

the broadest and most versatile computer network system on a national level.

- Educational opportunities in structural biology and molecular modeling should be improved. Several mechanisms are available, such as expanding graduate programs through new training grants; increased graduate fellowship and postdoctoral fellow programs; workshops, including formal hands-on training programs in molecular dynamics and molecular graphics; and working meetings of independent investigators to address critical limiting aspects of a particular problem.
- Innovative and interdisciplinary research proposals in both theoretical and experimental aspects of structural biology should be directly encouraged through the use of existing funding mechanisms.
- We see a special role for the national laboratories. The national laboratories should compete for the National Sequence Data Bank. The national laboratories and DOE have leadership status in the national computer network. They should increase efforts to make supercomputers available to the scientific community. Research efforts are going forward in molecular calculations and structural biology, with major programs at a few locations. Strengthening these efforts will assist the department's Office of Health and Environmental Research to assess the potential health and environmental effects of chemicals involved in energy processes.

2.

Introduction

Life depends on the orderly flow of information among biological macromolecules. Information is primarily stored as a linear code in nucleic acids. In general, there is a one-to-one correspondence between the amino acid sequence specified by the DNA code and that actually found in the protein. Processing, however, of the DNA itself, the transcribed RNA, or the translated protein frequently obscures this correspondence and makes it mandatory to verify the amino acid sequence directly. The amino acid sequence establishes what is called the "primary structure" of a protein. To be biologically active, the amino acids must be folded into a convoluted three-dimensional structure. There is some debate over the extent to which the final tertiary structure of a protein is governed solely by its primary sequence. The most widespread view follows from experiments by Anfinsen et al. (1961) who showed that bovine pancreatic ribonuclease regained activity from a denatured state without the involvement of any other macromolecule.

Today, it is generally accepted that all proteins of low molecular weight can refold spontaneously. The question remains, then, of whether large proteins use ribosomes or other cellular materials to fold by a kinetically driven mechanism. Thus, two important questions must be considered. First, how is the protein tertiary structure encoded in the linear sequence of amino acids? Second, is there a set of signals that controls the kinetic process of protein folding? Also, it is becoming increasingly obvious that the cell employs additional codes to signal for protein transport, protein destruction, carbohydrate structure, cell-cell recognition, hormone function, and the like. One place to begin deciphering all these messages that are so directly tied to the function of living cells is to gather information about the molecular structures of the proteins and nucleic acids.

Pauling's elegant insights into the importance of hydrogen bonding were the earliest quantitative ideas about the structure of proteins (Pauling et al., 1951). The models that emerged of helices, sheets, and the structure of collagen still form the basis of our understanding of fibrous proteins. Fibrous proteins are crucial elements in cellular architecture, but the chemistry of cells—the molecular synthesis and metabolism that are hallmarks of life—is controlled by elaborate enzyme systems whose detailed structural principles remain poorly understood. These proteins are often grouped together under the name of “globular proteins” because of their roughly spherical shape. In spite of the hundreds of x-ray studies of crystals of globular proteins (Perutz, 1965; Perutz et al., 1965) and important efforts at classification (Levitt and Chothia, 1976; Rossmann and Argos, 1977; Richardson, 1981), the most useful model of globular proteins remains the fundamental analysis put forward by Kauzmann in 1959.

As Kauzmann (1959) noted on thermodynamic grounds, the water-soluble globular proteins are formed from a hydrophobic core and a hydrophilic exterior. In the ensuing years, this idea has been generalized in two ways. First, it is now apparent that globular proteins of more than 100 to 300 residues are constructed from independent domains, each of which is built according to the Kauzmann hypothesis (Wetlaufer, 1973). Second, membrane spanning proteins appear to reflect similar thermodynamic principles. The hydrophobic environment of the core of the membrane causes the protein architecture to invert, producing a hydrophobic surface and*a charged or polar interior*, often in the form of a membrane-spanning channel (Henderson, 1979; Stroud and Finer-Moore, 1985).

Despite the success of the hydrophobic core model, it has not led to a theory detailed enough to offer an atomic description of protein structure. The basic difficulty is the amazing complexity of the atomic arrangements in these macromolecules. The three-dimensional structure of globular proteins is governed by a balance of the often contradictory requirements of optimum hydrogen bonding, burial of hydrophobic sidechains, and overall close packing. Thus, their geometry is governed to a significant degree by tertiary interactions, rather than by the simple hydrogen bonding that dominates the fibrous proteins and nucleic acid helices.

This complexity leaves us with what has frequently been called the "protein folding problem": how to calculate the tertiary structure of a protein to a useful degree of accuracy from the amino acid sequence. This report will address protein structure calculations at some length. But we recognize that there are much broader concerns dealing with the structures of nucleic acids and carbohydrate species. For example, for some years after the modeling efforts of Watson and Crick (1953), investigators took for granted the structure of nucleic acids. Recently, there has been renewed interest in the range of structures presented by double helical DNA (e.g. A,B,Z)(Drew and Dickerson, 1981); by RNA secondary structures; by higher-order packing of these molecules into nucleosomes, chromosomes, and ribosomes; and by the specific interactions of nucleic acids and proteins. As another example, the importance of postbiosynthetic modifications of all biological macromolecules is gaining increasing attention. Although the exact roles of phosphorylation, acetylation, methylation, and glycosylation are still being worked out, there is no doubt that these modifications frequently constitute the biologically active form of proteins and nucleic acids in living cells. Further, one must recognize the critical influence of environment on three-dimensional structure. The protein surroundings may be as simple as an aqueous electrolyte solution or as complex as macromolecular assemblies.

Another matter of concern is the actual pathway by which the protein folding takes place. In the cell, this pathway may be influenced or regulated by the proximity of the ribosome itself, by chain cleavage, or by the timing of the addition of carbohydrates or other chemical species. Even in the test tube, refolding appears to be a complicated process.

Beyond these questions, the interplay of structure and function suggests that we should try to understand the properties of proteins and how to manipulate them through the amino acid sequence. We should discourage from the start the view that a single structure exists for each protein. The conformational choices are numerous, even in the crystal state (Smith et al., 1986a). Further, most globular proteins are designed to provide organized internal motions (allosteric effects) as part of their functioning. Thus, one expects major conformational flexibility in many of these molecules. Some of this flexibility can be seen as thermally induced fluctuations that can be easily accounted for with normal mode analysis. Other aspects of the conformational freedom are much more complex and can be examined with the tools of molecular dynamics and statistical mechanics. In this report we will try to assess our understanding of all these issues.

It is almost impossible to overstate the importance of the protein folding problem to the elucidation of structure-function relations. As a purely intellectual exercise, it clearly displays the fascinating complexity of a first-class scientific puzzle. But the spotlight of attention is directed at protein structural predictions for other reasons. The most impelling is the incredible growth of knowledge in molecular genetics and protein engineering. New sequences are being reported worldwide at a rate of 1 every 10 minutes, while protein crystal structures are determined at a rate of 1 per month. Thus, sequences are being generated between two and three orders of magnitude faster than structures can be determined. Even with dramatic improvements in the technology of crystallography and magnetic resonance, the backlog of sequence data will grow rapidly. Further, the need to make rational plans for modification of protein properties is a major issue for all aspects of the biotechnology industry. A thorough understanding of the growth and development of living organisms depends crucially on a molecular structural and functional description. This understanding would certainly lead to a revolution in health care and a much firmer grip on the problems of chemical toxicity.

How valuable is it to know the structure of a protein? Clearly, models of structures have led to the design of pharmaceutical agents (Goodford, 1984 and Hol, 1986) and the engineering of specific properties such as improved stability (Ultsch et al., 1985). We are at the very beginning of such activities. They will surely become very important with time. A more difficult question is, what can be done with incomplete or lower resolution structures? Their present value is in organizing experimental data and planning new experiments (Cohen et al., 1986b). They can also be refined against new x-ray or nuclear magnetic resonance (NMR) data (e.g. Fitzwater and Scheraga, 1982; Brunger et al., 1986a).

It is not yet clear if these approximate structures, by themselves, can be refined sufficiently to compete with either crystallographic or NMR experiments for all uses. However, for the design of new pharmaceutical agents they are invaluable aids if an experimental structure is not available (e.g. Plattner et al., 1986).

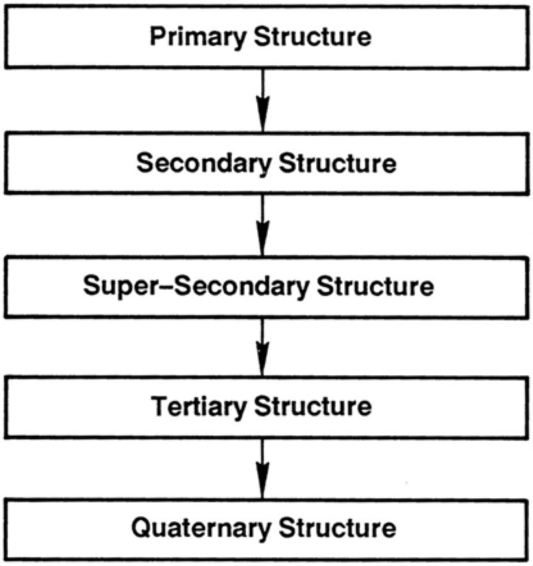


FIGURE 2-1 Hierarchy of structural descriptions of biological macromolecules.

In the broadest terms, the convergence of results from experiment and theory yields useful models of protein and nucleic acid structure and function. Given the very large number of sequences expected in the next two decades, we estimate that thousands of these will yield useful structural calculations.

The body of this report is a review of computer-assisted modeling of macromolecular structure and function. We have interpreted modeling in broad terms as the use of computers in molecular calculations of all kinds. The organization of the report is illustrated inFigure 2-1, which shows how information flows from primary sequence to secondary and tertiary structure models. The role of computers in experimental methods is summarized in the section on tertiary structure. Issues of modifying function and designing new materials are considered next, followed by a discussion of trends in computer hardware.

3.

Primary Structure of Proteins and Nucleic Acids

PROTEIN SEQUENCES AND DATA BASES

As of mid-1987, more than 5,000 protein amino acid sequences had been reported, most of which were inferred from the DNA sequences that encode them. Although the collection is redundant (same protein from different species) and definitely biased (many human and few plant sequences, for example), several patterns nevertheless stand out. Foremost among these is that the number of different types of protein is finite. It is becoming increasingly clear that most of the proteins determined thus far belong to identifiable families that are easily recognized by amino acid sequences alone. Indeed, the chances are now better than even that a newly determined amino acid sequence from a eukaryotic organism will be found to resemble a previously entered sequence.

Some of these families were anticipated (Table 3-1) on the basis of similarities in function and size. Thus, globins were all known to bind heme and to have very similar properties. We knew about large numbers of kinases, serine proteases, and thiol proteases, for example, and scores of protease inhibitors. It is not surprising, either, that many dehydrogenases and reductases have related sequences or that all the ATPases belong to a homologous set.

TABLE 3-1. Some well-established protein familiesa

Enzymes:		
Serine proteases	Dehydrogenases	Carboxypeptidases
Thiol proteases	Reductases	Transcarbamoylases
Acid proteases	Kinases ATPases	Phosphorylases
Non-Enzymes:		
Globins	Collagens	Immunoglobulins
Cytochromes	Keratins	Polypeptide hormones
Histones	Crystallins	Glycopeptide hormones
Protease inhibitors	Lipid-binding proteins	Interferons
Toxins	Transferrins	T cell receptors
MHC antigens		

[a] In each group some members are known to have similar amino acid sequences.

What is surprising is that the same kinds of protein structures are appearing in proteins in quite different settings. Polypeptide hormone precursors have been found that are related to protease inhibitors, for example, and structural proteins of the lens of the eye have been found to be related to regulatory agents called “heat shock” proteins (Ignolia and Craig, 1982). Sometimes the connections seem astonishing at first, but, upon reflection, they are very reasonable. Thus, the recently determined sequences of the beta-adrenergic receptor, which binds adrenaline and its derivatives, were found to be similar to that of rhodopsin, the eye pigment protein that responds to stimulation by light. This was extraordinary (Dixon et al., 1986), but not inexplicable, since the activating signals—adrenaline in the one case and light in the other—both can provoke excitatory actions. Subsequently, Kubo et al. (1986) found that a third protein, the muscarinic acetylcholine receptor, also belongs to this family.

Many proteins, then, came into being through a process of “duplicate and modify.” Gene duplications (partial or complete, as will be discussed further) lead to extra gene copies that suffer base substitutions in the usual way; the substitutions, in turn, lead to modified proteins. For most proteins, these replacements are established sufficiently slowly that, even after a billion years or more of evolution along diverging lines of descent, it is possible to recognize common origins.

More to the point for this report, an abundance of data show that three-dimensional structures of proteins are better conserved during the course of evolution than are their amino acid sequences. In this regard, a detailed crystallographic study of a series of serine proteases led to the conclusion that recognizably similar three-dimensional structures may endure as much as 10 times longer than do distinguishably similar amino acid sequences, a natural consequence of the diverse ways in which amino acids may be arranged to yield equivalent structures (James et al., 1978). As a result, it is reasonable to assume that similar amino acid sequences give rise to similar three-dimensional structures. This is obviously an important point because it is considerably easier to determine amino acid sequences (albeit using DNA sequencing) than to determine crystal structures.

The Current Crystal Structure Census

In the next decade, the sequences of 50,000 proteins are likely to be determined. In the majority of cases, it should be possible to assign them to existing families. The question then arises: For what fraction of those known families do we have crystal structures?

The Brookhaven Protein Data Bank, which keeps data for all known protein structures, lists about 300 entries. Like the sequence data banks, however, the Brookhaven collection is heavily redundant and biased. The entries include many variations of the same proteins crystallized in different settings (14 entries for egg-white lysozyme alone) and from different species. Actually, only about 100 different protein structures have been determined, and of these, many belong to the same families, as do the dehydrogenases or the serine proteases. Equally important, many known protein families—the interferons, for example—have not yet had a single crystal structure determined.

The situation is changing rapidly, however, and the prospects appear very good for determining a truly representative set of crystal structures. Innovations in recombinant DNA technology now allow the production of proteins in quantities sufficient for crystallization; previously, many of these proteins were available only in trace amounts. Beyond that, new techniques for crystallizing membrane proteins have opened an entirely new dimension (Michel, 1982). In addition, modern techniques for rapid data collection have revolutionized the entire field and hastened the process immensely (Xuong et al., 1978). Finally, improved techniques for structure solution and refinement have also accelerated this process.

Modeling on the Basis of Sequence Alone

At present, it is not possible to generate, with any hope of accuracy, a three-dimensional structure of a protein using only the amino acid sequence. Opinions differ widely as to whether a general solution to the “folding problem” is near and recent developments in the field are discussed elsewhere in this report. Certainly Cohen et al. (1986b) have shown that in special situations, much can be predicted about a protein on the basis of its sequence, but the routine application of an all-inclusive procedure is not yet in sight.

It is possible, of course, to make computer-assisted predictions about secondary structure (Chou and Fasman, 1974), and although these methods have a limited accuracy (Kabsch and Sander, 1983; Nishikawa, 1983), they can nevertheless provide useful information when applied judiciously. Predictions about protein structure can also be made with computer programs that assess hydropathy (Kyte and Doolittle, 1982). These have been especially successful in predicting the membrane-spanning segments of membrane-associated proteins.

Existing Data Bases

The major data bases for sequences are the Protein Identification Resource (PIR) at the National Biomedical Research Foundation, Washington, D.C.; GenBank, operated by Bolt, Beranek, and Newman in Cambridge, Massachusetts; and EMBLData Bank in Heidelberg, Germany (Table 3-2). GenBank and EMBL store only DNA sequences, although recently, they have begun to make available derived amino acid sequences. GenBank and EMBL exchange data frequently as a way to enhance the data sets. Currently, these two data bases together contain more than 15 million bases of nucleic acid sequences.

TABLE 3-2. Some sequence data banks and searching facilities

Protein Identification Resource
Georgetown University Medical Center
National Biomedical Research Foundation
3900 Reservoir Road, N.W.
Washington, D.C. 20007 U.S.A.

GenBank: Genetic Sequence Database
Computer & Information Science Div.
BBN Laboratoreis, Inc.
10 Moulton Street
Cambridge, MA 02238 U.S.A.

EMBL Data Library
Graham Cameron, Data Library Manager
Postfach 10 2209 Meyerhofstrasse 1
6900 Heidelberg, Germany

University of Wisconsin Genetics Computer Group (UWGCP)
John Devereux
University of Wisconsin Biotechnology Center
1710 University Avenue
Madison, WI 53705 U.S.A.

Unite d'Informatique Scientifique
Jean-Michel Claverie
Institute Pasteur
Paris, France

Bionet: National Computer Resource for Molecular Biology
Intelligenetics, Inc.
124 University Avenue
Palo Alto, CA 94301 U.S.A.

PRF Amino Acid Sequence Collection
Yasuniko Seto
Peptide Institute, Protein Res. Found.
476 Ina
Minon, Osaka 562 Japan

[a] Although GenBank and the EMBL Data Library primarily store DNA sequences, translated versions of the data are available.

All of the three major banks currently have backlogs of data awaiting entry. In the case of the PIR, for example, most of the protein sequence data are still typed in from the published literature. GenBank and EMBL are trying to make arrangements with some key journals (*Nucleic Acid Research*is one) that will allow data to be submitted to the data bases in various computer modes: diskettes, tapes, and direct transmission.

The logistic and policy problems associated with such data bases are enormous and many committees and societies are trying to establish an acceptable policy that will speed things up. At the same time, all indications are that the generation of sequence data will increase exponentially in the next few years. Clearly, we need a new, permanent, centralized data repository. Ideally, this should be international; practically, it may be more readily attained at the national level. This should be an institution at least as large as the National Library of Medicine. It could be located anywhere, although certainly consideration should be given to Los Alamos, which has already been heavily involved in sequence banking with GenBank. Models for this center, which must be a constantly updated base and not merely a repository, include the National Bureau of Standards or the Coast and Geodetic Survey.

In addition to these major sequence collections, some smaller enterprises are operating, both in the United States and elsewhere (Table 3-2). However, all of these appear to rely heavily on the PIR-GenBank-EMBL collections as their data cores. Finally, some beginning efforts are underway to create a carbohydrate structure bank.

PATTERN BASED COMPARISONS

Methods for determining sequence alignments and homologies represent a powerful set of tools for relating the structures and, potentially, the functions of two or more biopolymers. However, these homologies alone do not take full advantage of the information content of primary structures of biopolymers. For example, different nucleotide patterns may code for the same protein sequence; different protein sequences may share a very similar function. Often the differences at the level of primary sequence are substantial, and similarity of function among sequences is lost when the sequences are compared by homology. Yet, if we can somehow relate the sequence of a new protein to that of another protein whose structure and function are known, we can begin to determine the structure of the new protein. How might we take advantage of a more general view of homology, and view sequences as patterns in order to extract more structural information from them?

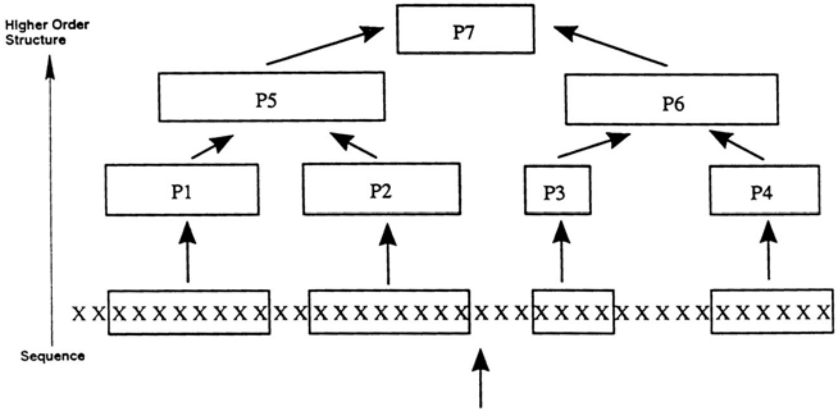


FIGURE 3-1 Hierarchical relationship among patterns in biological sequences.

If we consider the concept of a sequence as a string of characters, we can view the concept of a pattern as a string (or collection) of partial sequences. This view of primary amino acid sequences derives from our belief in a hierarchy of structural descriptions of proteins (See Figure 2-1).

If the only structural information we have about a protein is its sequence or primary structure, we seek patterns in its sequence that will guide us to a better understanding of the protein. Patterns can be derived by examining the secondary and tertiary structures of known proteins, and relating spatial information back to patterns in the primary structure. In essence, we attempt to map what is known about spatial configurations into the one-dimensional world of sequences. This mapping itself can be done hierarchically, closely related to the hierarchy shown above, as shown from the bottom up inFigure 3-1.

The scheme symbolized inFigure 3-1represents the fact that a sequence can be analyzed to determine patterns of partial sequences, illustrated by the boxed elements. The individual patterns*P1 -P4*, for example, secondary structural elements, may themselves be part of a larger patterns, such as*P5*, *P6*, which may only be recognized after the abstraction to*P1 -P4* is accomplished.*P5* and*P6*, for example, structural domains, may then be recognized as part of an even larger pattern,*P7*, for example, the tertiary structure of a protein.

The hierarchy ofFigure 3-1also represents a computational model for deriving secondary and tertiary structural information from the patterns themselves. This computational model is described below.

Methodology

The concept of patterns is useful for expanding our view of the primary structure of proteins only if we find some method for labeling the individual amino acids or partial sequences with properties other than identity. For example, we can choose labelings related to physical, chemical, or functional properties of the amino acid. If we choose properties related to secondary or higher order structure, we can encode higher order properties in the sequence itself. Labelings can be assigned to individual amino acid residues or to groups of residues based on calculations over a partial sequence. Typical properties included are:

- charges: plus, minus or neutral
- pK: acidic, basic, neutral; or specific value
- hydropathicity: a hydrophobicity or hydrophilicity value for one residue or calculated over a set of residues
- chemical similarity: several possible definitions
- tendency for replacement over evolutionary time
- secondary structure calculated over a group of residues (see below)

Some labelings are valuable for examining sequence homologies. For example, one can examine similarities in the chemical properties of a sequence by performing the homology search among sequences labeled with characters or flags that represent assignment to various chemical classes. If the evaluation function for rating the degree of homology is adjusted appropriately based on the variety of the labels, or the*alphabet*, it is possible to find meaningful homologies among sequences with substantially different amino acid sequences. One example of the success of this approach is a method for finding protein sequence homologies based on the tendency for replacement over evolutionary time of one amino acid by another (Dayhoff, 1978; Lipman and Pearson, 1985).

Software and Hardware Considerations

The computer software and hardware needed to carry out the pattern-matching operations on sequences contrast in interesting ways with those required for the intense numerical calculations described in other sections of this report. First, the analyses of patterns in sequences is a problem in symbolic, not numeric, computation. For efficient program development and application, languages that support symbol manipulation and pattern-matching primitives are desirable; C and LISP are often the languages of choice. Second, the algorithms used are the result of years of development of dynamic programming techniques, recursion, and other methods that allow flexible pattern detection and matching. Mismatches, density matches, gaps, and other irregularities are all characteristics of “fuzzy” patterns that must be accommodated. Third, many of the techniques used to infer higher order structural information in patterns, such as secondary structural analysis and the heuristic techniques described below, are highly empirical. These methods often have a theoretical foundation or justification, but the actual inferences may be based on statistics over data bases of known structures or on judgmental rules based on experience and analysis of patterns in known systems.

The computer power needed for these symbolic computations is widely available on several general purpose machines. However, as is true for numerical computations, progress in symbol manipulation and pattern matching could be hampered by insufficient hardware. It is becoming more and more difficult to explore complex patterns in large data bases because of the computer time required. Thus, just as array processors and high performance graphic devices are available to support numerical calculations and present their results, special purpose pattern-matching hardware has been developed to emulate the necessary symbolic computations. This hardware contains the algorithms required for symbol processing. It is often several orders of magnitude faster than the corresponding software.

4.

Secondary Structure of Proteins and Nucleic Acids

PROTEINS

The secondary structural features of proteins can be grouped into three broad classes: helical features, extended strands, and turns or loops. The most commonly seen helices are the so-called alpha helices, which were first described by Pauling and Corey (Pauling et al., 1951). Minor forms include the 3-10 and pi helices. Beta structures are formed from hydrogen bonding between the backbone amides of extended polypeptide strands. Although such features are properly considered tertiary structure, they are often discussed as secondary structure. They are named according to whether the strand pairs run parallel or antiparallel to each other (based on a vector drawn from the N to C termini of the feature) and whether the sheets are folded or rolled into a barrel (Richardson, 1981). Turns are much less regular (Rose et al., 1985). They are characterized according to local geometric features.

Prediction of secondary structure has been of interest since the first protein structures were determined. Accurate secondary structure prediction is one direct approach to the development of a tertiary prediction algorithm.

The methods used to predict secondary structure from amino acid sequence have been (1) calculation of the energies of the major conformers for a given sequence; (2) statistical analysis of known structures; and (3) modeling. It is now computationally feasible to calculate energies for conformations of short peptides with or without solvent. However, this is not yet a definite method of predicting secondary structure in proteins because the energy differences among conformers are relatively small compared to the interactions between the peptide and the rest of the protein and because of neglect of the solvent entropy terms.

Statistical methods began with the efforts of Chou and Fasman (1974), who characterized the preference of each amino acid found in each type of secondary structure. Other early efforts focused on turns in proteins (Kuntz, 1972; Lewis et al., 1971). Robson (Robson and Osguthorpe, 1979; Robson, 1986) followed with more sophisticated approaches. These approaches give the general impression that the statistical methods are easy to use but have significant random and systematic noise that limits their accuracy. For example, they ignore long-range effects (Kabsch and Sander, 1983) and prosthetic groups.

Modeling efforts have grown from the early observations of Schiffer and Edmundson (1967) that alpha helices in globular proteins often contain hydrophobic and hydrophilic faces in agreement with the ideas of Kauzmann (1959). Many investigators followed this line of thought. Thus, helical propensity has been identified with helical nets, from a Fourier analysis of the hydrophobicity (Eisenberg et al., 1984; Finer-Moore and Stroud, 1984), or from pattern-matching (Cohen et al., 1986b). Beta structures have been treated in similar ways, although less successfully. Turns are often associated with regions of hydrophilicity.

Some labelings are valuable for detecting higher order structural information. One of the most common methods used to explore this information is secondary structure analysis. This analysis provides information on possible patterns of coils, sheets, and helices in a protein. Other information can be extracted from the sequence by recognizing that certain combinations of amino acids indicate turns or other structural features. Alternative representations can be used to determine patterns comprising amphiphilic beta-sheets or alpha- or pi-helices (Kaiser and Kézdy, 1984). Often this information can be gathered by empirical, rule-based systems, described below. However it is determined, this information can be used to build up a hierarchy of patterns. This hierarchy is indicated symbolically inFigure 4-1, which is based on the general scheme shown inFigure 3-1.

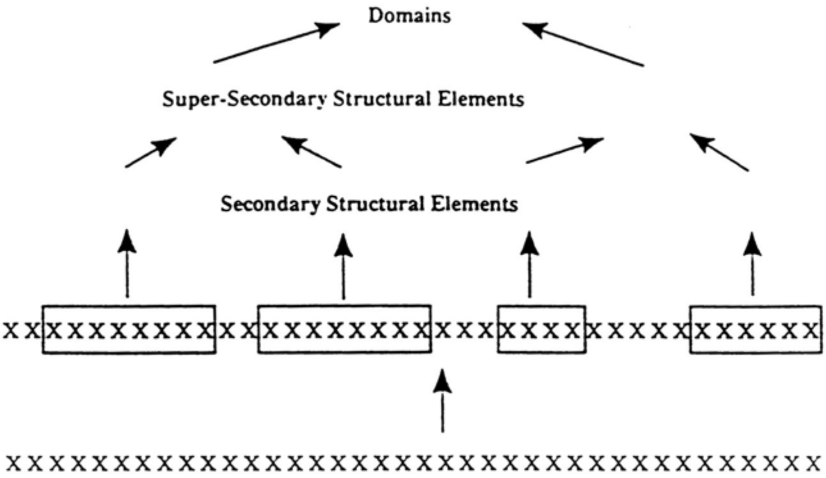

Domains
Super-Secondary Structural Elements
Secondary Structural Elements
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

FIGURE 4-1 Hierarchical relationship of a protein sequence to higher order patterns of organization, culminating in the tertiary structure of the protein itself.

According to this scheme, patterns of partial sequences can be recognized as representing secondary structural elements. Combinations of these elements are recognized as higher order patterns that correspond to supersecondary structures. These, in turn, may be recognized as patterns that make up a domain. If any portion of such a hierarchy is determined using only a sequence, this represents a major step toward constructing an actual three-dimensional structure for a protein. More frequently, this procedure is used with other computational techniques for examining combinations of patterns that may lead to recognizable structural motifs in three dimensions.

The general method used to study these problems involves only a few basic steps. First, one or more techniques are used to make the first assignment of structural features to patterns in a sequence—for example, hypothesized secondary structural elements associated with specific partial sequences. Second, the pattern or patterns obtained are compared to those derived from known structures or hypothesized by the investigator. Third, if a match is found, the investigator proceeds to the next step of searching for patterns of the next-highest order, which may be composed of several different combinations of smaller patterns identified in the previous step. If one can derive sufficient structural information, hypothetical three-dimensional structures can be proposed. Several examples have appeared recently where this approach has been successfully applied to pattern-based elucidation of structural features of proteins of known and unknown structure (Abarbanel, 1984, 1986; Cohen et al., 1986; Taylor and Thornton, 1983).

Current Status and Future Prospects

The assignment of each amino acid in a protein sequence to a particular secondary structure class has rarely been more than 70 percent accurate and is often worse. Some of the newer approaches increase accuracy by reducing the scope of the problem. For example, Cohen et al. (1986a) describe procedures for the prediction of turns in subgroups of proteins. By tailoring algorithms to take advantage of the characteristics of, for example, all-alpha domains, the accuracy is improved to about 90 percent.

In the near term, developments in this area are most likely to be incremental improvements. Computer speed will surely continue to increase substantially. Data bases will continue to grow at least linearly. Experiments on the structural consequences of modifying amino acids are beginning to be reported in significant numbers (Ultsch et al., 1985; Alber et al., 1987). More powerful statistical and modeling efforts are under development. What is more important, these approaches can be combined in useful ways. Within five years, several laboratories should have set up unified programs that allow comp]ex inquiries of structural data bases. Automatic learning programs that extract secondary structure features will also have been intensively studied.

Within 10 years, we may well see major improvements in our ability to correlate sequences and secondary structure. The current goal of correctly predicting every major feature in a new sequence is a plausible target. Accurate specification of the secondary structural environment of each amino acid in a protein is probably not attainable without a major conceptual or computational breakthrough.

NUCLEIC ACIDS

Predicting RNA Structure

RNA molecules are crucial to all stages of protein synthesis. Messenger RNA carries the code that specifies the amino acid sequence of the protein; the transfer RNA molecules translate the code word by word into protein; and the ribosomal RNAs in the ribosome provide part of the machinery to do the synthesis. Many animal and plant viruses that cause tremendous damage to human health and economic well-being are RNA viruses. Human disease RNA viruses include those for colds and influenza, AIDS, some cancers, and hepatitis. It would be very useful to be able to use only the sequence to predict the folded three-dimensional structure of any RNA in any environment. This structure determines how stable an RNA molecule will be in a biological cell, because the ability of the enzymes that hydrolyze RNA (exo and endo nucleases) to degrade a particular RNA is very sensitive to RNA conformation. Also, each RNA molecule requires the correct conformation in order to function biologically. The conformation, in turn, will depend on the environment as characterized by the type and concentration of ions, the presence of specific interacting molecules and other variables.

The computer programs now used to predict RNA conformation from sequence have limited goals and limited success (see, for example, Zuker and Steigler, 1981). They were designed to calculate secondary structure only—to specify which bases are paired. The computerized procedure uses experimental thermodynamic data on double-strand formation in synthetic RNA oligonucleotides (Freier et al., 1986). A dynamic programming algorithm considers all possible base pairs in the RNA—a sequence of N nucleotides has N (N-1)/2 possible base pairs—and calculates the free energies of the corresponding structures. The free energy of a structure is assumed to be the sum of the free energies of its constituent substructures (Tinoco et al., 1971), including single-stranded regions, double-stranded regions, bulges, hairpin loops, and interior loops. The lowest free energy structure is the predicted secondary structure. The computer programs allow one to specify that any two bases are paired to each other or that any base is unpaired. Thus, other experimental data based on enzymatic digestion experiments, chemical reactivity, or phylogenetic comparisons can be introduced.

Clearly, the predicted results are only as good as the experimental thermodynamic and other data used. For example, although all transfer RNAs are thought to fold as clover leaves, present computer programs only calculate about 90 percent clover leaves. Also, the thermodynamic experiments have all been done in one standard solvent, so knowledge about other solvents is needed. Finally, a limited number of oligonucleotides have been studied; they provide only a very small sample of the structural elements present in natural RNA molecules. A much better understanding of the thermodynamics of possible substructures in an RNA is needed before an accurate and complete prediction of secondary structure is possible.

The existing computer methods for calculating secondary structure in RNA are useful as aids in designing experiments to determine the actual secondary structure. For example, a program can provide the calculated lowest free energy structure, as well as other significantly different low free energy structures (Williams and Tinoco, 1986). Experiments are done to test some of the predicted substructures, and their results are incorporated into the calculation of the next prediction. Successive approximations thus lead to more correctly determined secondary structures (Cech et al., 1983).

Computer methods of the future must be based on much more detailed knowledge of the thermodynamics of local regions of an RNA molecule. The extensive thermodynamic data needed for correct prediction of secondary structure will most likely come from computer interpolation and extrapolation of limited data measured on synthetic oligonucleotides in a few solvents. Once we understand better the forces and energies involved in the interactions of nucleotides with solvent, ions, and each other, it will become easier to calculate secondary structures for large RNA molecules. Algorithms exist for calculating low free energy structures as a sum of substructure energies, so their use with RNA of up to about 600 nucleotides requires only the appropriate data. However, for a rigorous search for structures, central processor (CPU) time and memory requirements increase as the cube or fourth power of the number of nucleotides. We estimate that a molecule of 3,000 nucleotides (typical ribosomal RNAs or small viruses) would require about 40 hours of CPU time on a Cray XMP and 2 gigabytes of memory. Parallel processing or equivalent improvements in hardware then become necessary.

Prediction of tertiary structure in RNA—the three-dimension al structure of the RNA—is much more difficult. Levitt (1969) made an early attempt at predicting the structure of transfer RNA, but an algorithm to find the lowest free energy tertiary structure still does not exist. Proposals of methods to fold possible secondary structures into three-dimensional structures and calculate their energies are in very early stages of development. Prediction of secondary structure in RNA is about at the same stage as it is in proteins, but prediction of tertiary structure is far behind. Novel methods are needed for RNA that take into account either implicitly or explicitly:

- the long range electrostatic repulsion of phosphates shielded by counter ions;
- the detailed conformation of hairpin, bulge, and interior loops;
- the hydrogen bonding between loops and single-stranded regions (pseudo-knots);
- non-Watson-Crick base pairs and triple base interactions; and
- all the usual London van-der-Waals interactions, including solvent.

Kollman and his associates at the University of California, San Francisco have made a beginning in this direction with the program AMBER (Bash et al., 1987a). This program performs the molecular mechanics calculations of energies with parameters optimized for nucleic acids. Other programs calculate differences in free energies caused by changes in conformations.

The most useful computer modeling process would provide real-time calculation of free energies as the folding of a macromolecule in a solvent was shown on a computer graphics screen. Achieving this will require great improvement in hardware and software. Close collaboration with experimentalists will be needed to ensure meaningful calculated results.

Rapid and efficient progress in this field will require:

- Effective methods for crystallizing RNA oligonucleotides and naturally occurring RNA molecules. To date, transfer RNA is the only RNA molecule whose x-ray structure has been determined.

- Nuclear magnetic resonance methods that can provide conformations for RNA molecules that contain from 10 to 100 nucleotides.
- Computer programs that can reproduce experimental results. The high charge densities in nucleic acids (one per nucleotide) require special care in the correct treatment of solvent and ionic effects. Once calculations can be done that provide known structures, we can place some confidence on extrapolation to new structures.

We have been stressing free energies (thermodynamics), but kinetics may be just as important; equilibrium is never attained in a living system. RNA is folded as it is synthesized, so kinetic barriers may prevent it from reaching a global minimum. For some RNA molecules, such as ribosomal RNA, the dynamic movement from one conformation to another may be an important part of their function.

We would like to be able to calculate and verify structures of RNA molecules and their interactions with a wide variety of molecules. In a ribosome, for example, the ribosomal RNA interacts with messenger RNA and transfer RNAs, as well as all the proteins involved in protein synthesis. It would be very useful to know how a change in any variable would affect the efficiency and fidelity of protein synthesis. We want to be able to design efficient messenger RNAs to produce any protein desired. We need to develop models of the process of protein manufacture so that we can then improve the productivity, cut the cost, and ensure high quality output of proteins. Computer modeling and calculations should provide the sequences of the ribosomal RNAs, the transfer RNAs, and the messenger RNAs that would be optimal for the production of a particular protein. We are very far from this ideal.

We need mathematical and dynamic structural models of how an RNA virus replicates, how reverse transcriptase copies the RNA into DNA, and how the RNA is packaged into its protein coat. With this knowledge, we will be much closer to finding ways to prevent or cure diseases caused by RNA viruses, which include colds, influenza, AIDS, and hepatitis.

RNA is now known to have catalytic activity (Zaug and Cech, 1986). To date, it has been demonstrated that RNA catalysis is involved in RNA processing and in glycogen synthesis. This catalytic activity of RNA is needed for the replication cycle of some viroids and virusoids (small infective RNA particles) and the processing of some ribosomal RNAs. Fundamental advances in understanding this catalytic activity require knowledge of the location and structure of the active site of RNA enzymes. Computer calculations in conjunction with mutation experiments may allow us to progress most rapidly.

Predicting DNA Structure

Although DNA and RNA are very similar, in practice, the important problems relating to their sequences and structures are usually different. DNA stores all the genetic information that determines the organism and its characteristics. Its sequence, conformation, and interactions with proteins, small molecules, and other DNA molecules determine how and when the genetic information is expressed. The ability to understand and ultimately to control the genetic expression will make it possible to control genetic diseases, bacterial diseases, and DNA viral diseases. Thus,the important questions for DNA are:

- What is the detailed conformation of any sequence of double-stranded DNA and how does it depend on the environment?
- How does the conformation interact with other molecules?

To answer these questions, we will need to understand nucleic acid structure, protein structure, and their interactions in complex environments. It will take a great deal of effort to achieve this understanding, but the rewards for society will be very high.

X-ray diffraction experiments have determined structures for double-stranded DNA, protein-DNA complexes, and DNA-small molecule compounds in crystals. Computer modeling is needed to extrapolate these results to predict what would occur in different biological environments and to other complexes. One obvious application is in the design of more specific and more effective antibiotics. In general, we would like to be able to design molecules that can start or stop the expression of any gene in any DNA.

5.

Tertiary Structure of Proteins and Nucleic Acids: Experimental

X-RAY DIFFRACTION OF BIOLOGICAL MACROMOLECULES

In 1934 Bernal and Crowfoot demonstrated that a crystalline protein could give rise to a well-ordered x-ray diffraction pattern, thus setting the stage for modern analysis of the structure of proteins. Progress was gradual at first, interrupted by World War II, but in 1953 Green, Ingram, and Perutz took another essential step when they accomplished the first heavy atom analysis of a hemoglobin crystal (Green et al., 1954). The culmination of these years of work came in 1959 when Kendrew and his colleagues (1960) reported the analysis of myoglobin at 2 Å resolution, revealing for the first time the underlying structure of a globular protein. They noted the complexity and lack of regularity of the molecule—major features that continue to impress us today as general features of protein structure. The alpha-helices and beta sheets of Pauling and Corey (Pauling et al., 1951) form striking regions of regularity, but are joined together in very complex ways.

Another major event of protein structure analysis occurred the same year when Cullis et al. (1959) described the structure of hemoglobin at 6 Å and demonstrated that the folding of the globin chain is similar to that in myoglobin, despite relatively low sequence homology between the two. This observation of a family pattern to the three-dimensional structure of globins has been followed by the identification of many other families.

Today, several hundred proteins have been analyzed by x-ray diffraction and their three-dimensional structures catalogued. This number continues to grow at an ever-increasing rate and, together with the amino acid and gene sequence data, forms the principal basis for understanding the mechanisms of action of these proteins at the molecular level.

The development of two-dimensional nuclear magnetic resonance (NMR) techniques already is a valuable complement to x-ray diffraction for relatively small molecules (less than 10,000 molecular weight) but for the foreseeable future, crystal structure analysis will be the principal experimental source of structural data for enzymes, nucleic acid binding proteins, antibodies, and other proteins of the immune system, receptors, and indeed, for all proteins that can be effectively crystallized.

Determining the three-dimensional structure of a biological macromolecule by crystallography involves a number of clearly defined steps. First, crystals of suitable size and diffraction properties must be prepared. Next, x-ray diffraction data must be collected for these crystals and also, typically, for a number of heavy atom derivatives of the crystals. These data can then be assembled to obtain an electron density map using a computational process that resembles the action of the lens in a microscope. This map must then be fitted with a polypeptide chain of the appropriate amino acid sequence. Because the map is of less-than-atomic resolution and also contains errors in the phase determination, the investigator must have considerable skill to obtain the best fit. The resulting protein model must then be refined to remove the errors present in the map as much as possible as well as those errors introduced by the fitting process.

Computers play an essential role in most of these steps. Even the analyses of the first protein crystal structure, myoglobin, could not have been accomplished without the use of the EDSACII in Cambridge (Kendrew, 1960). At present, modern crystallography depends completely on heavy computer use, and this dependence will certainly increase steadily in the future. In the four mathematical procedures required to solve a structure using protein crystallography—data processing, phase determination, map fitting, and refinement—new methods are continually appearing that depend on ready access to considerable computer power.

First, we will consider the first step, data collection. This is now changing significantly, as most laboratories in the United States convert from the use of film or diffractometer to the use of area detectors. These machines can increase the speed of data collection within the laboratory by as much as two orders of magnitude. The output from the area detectors is generally processed directly, resulting in the speedy production of finished intensity data. The ability to produce directly, in a few days, data that previously were collected in weeks or months of labor-intensive work is revolutionizing the field at an opportune time, when developments in genetic engineering have made it possible to use site-directed mutagenesis to answer many structural questions. These questions, however, require separate data sets for each mutant.

When measuring x-ray intensities, whether using photography or area detectors, the presentation of the data on a computer graphics screen can make it much easier to analyze the diffraction pattern. Area detectors coupled with graphics facilities can now be used to align a crystal almost in real time. It is not clear how much the use of computer graphics in data collection will continue to increase in the future, since its impact will probably be reduced by the increasing power of the data processing software.

A major discovery of protein crystallography is that most proteins belong to families with closely related three-dimensional structures. Examples include the hemoglobins, serine proteases, aspartic proteinases, and the immunoglobulin domain structure. Consequently, we can now use the known structure to obtain phase information about an unknown structure, a method known as molecular replacement (Rossmann, 1972). There have been numerous examples of the successful application of molecular replacement to determine crystal structures. The structure of a bacterial serine protease inhibitor was used to analyze the structure of the protease bound with an inhibitor (James et al., 1978). Another example is the use of the known structures of lysozyme and the two-domain modules of an immunoglobulin Fab to analyze the structure of a monoclonal antibody bound to lysozyme (Sheriff et al., in press). The use of this molecular replacement method will become even more widespread as more members of a protein family are investigated.

Molecular replacement techniques have also been applied in the use of redundancy to obtain phase information (Rossmann, 1972; Harrison et al., 1978). A spectacular illustration of this occurred in the recent analyses of the picornaviruses for polio and the common cold (Rossmann et al., 1985; Hogle et al., 1985).

These methods require heavy computational analysis for their success. For example, Rossmann and Argos (1977) concluded that they could determine the rhinovirus structure only with extensive use of the supercomputer at Purdue University and could not have carried out the analysis and phase extension without such a facility.

COMPUTER-ASSISTED MODELING IN DNA STRUCTURE ANALYSIS

In the current method of fitting the electron density map, computer graphics are essential. Although several programs can fit an approximate model of a protein to the map without human intervention, most crystallographers do not use them, preferring to use instead computer graphics to fit. Here, the development of color and stereo graphics systems has been an important advance. Computer graphics constitutes such a colorful and seductive tool that it virtually compels the nonscientific observer to believe in whatever phenomenon is being displayed. This inherent fascination with graphics display extends to some working scientists as well. One always must ask whether [to paraphrase an old joke about statistics] a scientist is using computer graphics like a drunk uses a lamp post: more for support than for illumination. But in the area of macromolecular structural analysis, particularly with DNA and its complexes with antitumor drugs and control proteins such as repressors, computer graphics will illuminate by enabling the investigator to carry out the structure analysis efficiently and to see aspects of the structure of the molecule that he could perceive only with difficulty or would overlook entirely using more traditional methods.

In the early 1950s, before any protein structure had been determined by x-ray methods, the British crystallographer J. D. Bernal once remarked that, even if we were to obtain an electron density map of a protein, we would never be able to understand it until we could build a map big enough to walk through and point out features around and above us as we walked (personal communication, around 1955). Kendrew had this dictum in mind when he constructed the first electron density map of any protein, myoglobin, by attaching color-coded spring clips up the lengths of steel rods mounted in a regular grid on heavy plywood baseboard. (A portion of this first map still exists on display in the Kensington Science Museum in London.) Richards (1968) provided the next step in the display of large and complex electron density maps of macromolecules, with a half-silvered mirror arrangement that came to be known as a Richards Box. The device superimposed a direct image of mylar sheets of electron density on a reflected image of the wire model being constructed (Figure 5-1). In the 1960s and 1970s, virtually all macromolecular structure groups had at least one Richards Box to use in interpreting electron density maps and building macromolecular models into them.

This entire half-silvered mirror box technology has been replaced by new methods of computer graphics. Detailed chain fitting is carried out on the graphics screen, fitting stick bond skeletons into “chicken-wire” three-dimensional contoured volumes. Diamond (1966) wrote the first such display program, BILDER. This routine has largely been superseded by the easier-to-use FRODO routines of Jones (1985). More recently many even more flexible packages have been offered, both for large mainframe computers and minicomputers.

The software most widely used is the program FRODO developed by Alwyn Jones, who is at the University of Uppsala, Sweden. The use of FRODO on a computer graphics system has now almost entirely replaced the construction of mechanical models, and greatly improved the accuracy in the modeling and, in particular, the speed and precision of the more predictive kinds of modeling, such as fitting substrates to the surfaces of enzyme molecules.

In these programs, atoms can be placed within the observed density with great accuracy. Because the coordinates are in the computer as soon as the atoms are located, acceptable bond lengths and angles can be built into the trial model from the outset. Once a trial model has been built into the displayed electron density, least squares refinement against the x-ray can be carried out using one of several available programs that compensate for a too-small ratio of data points to refined parameters at less-than-atomic resolution. The programs impose constraints on

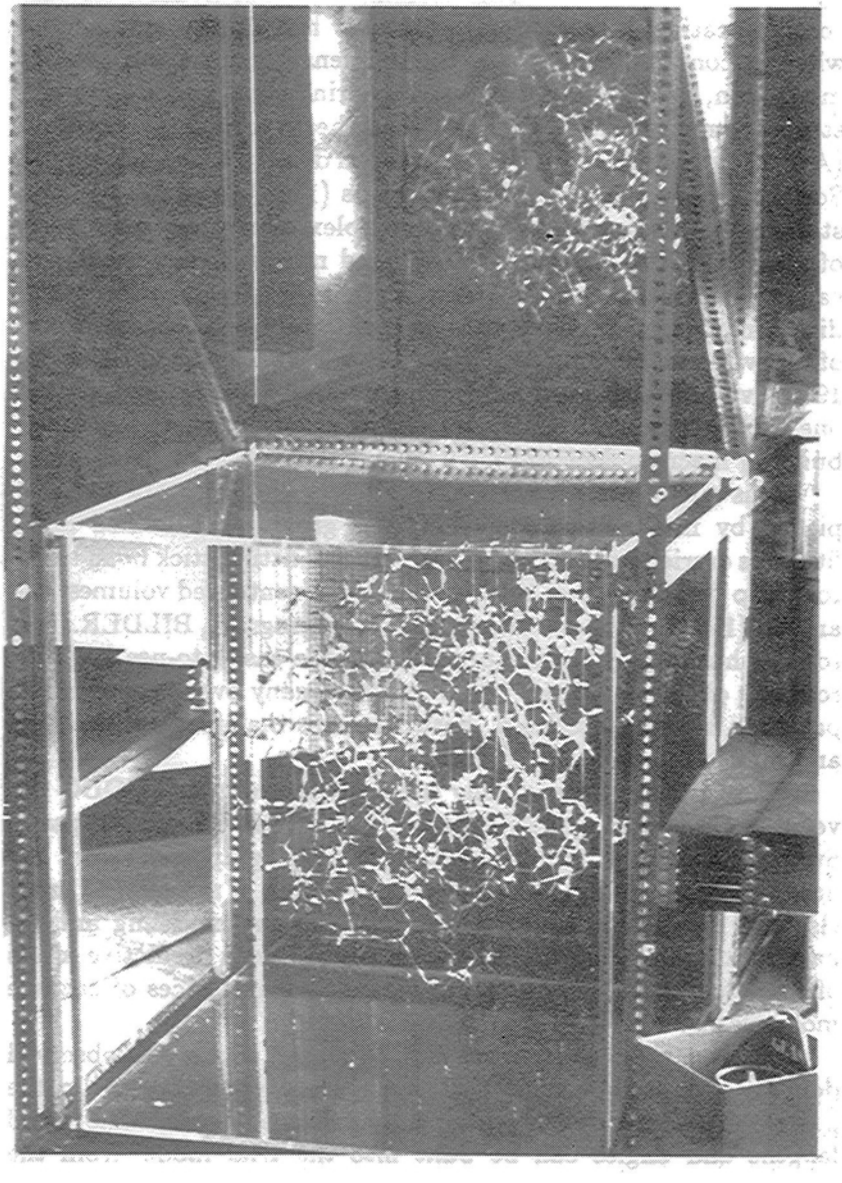


FIGURE 5-1 Illustrations of the method used before the advent of computer graphics to interpret an electron density map of a macromolecule in a Richards box. Section through the map of the protein cytochrome*c* are mounted on plexiglass sheets that can be slid over the face of a light box at the upper rear. The lower left front shows a wire model of the cytochrome*c* molecule, constructed in a framework with transparent top. The half-silvered mirror in the frame at 45 degrees to the main frame supporting the box superimposes a direct view of the contoured map and a reflected view of the wire model below. This approach was “high tech” around 1968.

bond lengths, bond angles, and (if desired) bond torsion angles to keep them within known acceptable limits.

In refining protein and nucleic acid structures, the use of new methods is increasing investigators' dependence on heavy computing. Current methods of refinement include the use of restrained least squares programs originally developed by Hendrickson and Konnert (1980), Sussman (1985), and by Jack and Levitt (1978) among others. These programs can be used successfully on relatively slow computers, particularly when the Fast Fourier Transform algorithm is used to accelerate them. At present, the restrained least squares refinement procedure must be interrupted frequently to compare the fit of the model to the electron density map, remodel to remove stereochemically unacceptable local minima, insert solvent molecules, and for other procedures. This process of model building is the step that is limiting the rate of the refinement procedure.

It takes many days, even weeks, for a comprehensive, residue-by-residue examination of an average-sized protein molecule of, for example, 40,000 molecular weight. The procedure for identifying solvent molecules on the protein surface is also tedious and time-consuming. Any method that would reduce the number of interventions during a refinement or would require less human intervention during the modeling would speed up the refinement process. In this respect, Brunger et al. (1987) have recently shown that by combining protein dynamics simulation with the refinement process, one can avoid many of the minor minima that sometimes occur in the usual refinement procedures. However, the successful application of the dynamics calculations would require considerably more computer power than is currently available to most crystallographic laboratories.

At any point during refinement, computer graphics enable one to examine the trial structure alone or superimposed on the electron density in one of several options: a simple Fo electron density map, a $(F_o - F_c)$ difference map, or the $(2F_o - F_c)$ map that is in fact the superimposition of the two previous functions. When refinement is complete, the resulting coordinates are immediately available within the computer for use in drawing figures that illustrate the final structure. This ability to work within the computer during model fitting, refinement, and display of results saves an enormous amount of time over the older technique of constructing physical models.

Difference Patterson Vector Maps to Locate Heavy Atoms

The standard way of determining the phases of the x-ray diffraction pattern of a macromolecular structure is to prepare one or more heavy atom derivatives of the macromolecule that differ from the parent compound only by the addition of a heavy atom—metal or halogen—at defined locations in the molecule. For DNA, this can be done either by synthesizing the DNA oligomer using 5-bromocytosine instead of cytosine at one point in the sequence or by diffusing heavy atom complexes into the DNA after crystallization. Data are then collected on the parent DNA crystals and on each of the available heavy atom derivatives. Heavy atom positions are found by interpreting a difference Patterson vector map, which in principle has features that locate all of the vectors between heavy atoms in the crystal unit cell.

Difference Patterson vector maps traditionally are examined by using minimaps stacked on plexiglass sheets that are placed on a light box. Computer graphics display discussed above, however, offers several advantages over this old technique. Once a trial position of a heavy atom is established, one must calculate all of the heavy atom-heavy atom vectors possible and see how many of them correspond to features in the Patterson map. If the crystal has appreciable space group symmetry, this is a tedious and painstaking job. But it is a trivial task for a computer. The operator need only type in or locate a trial heavy atom position, and*all* the resulting interatomic vectors can be displayed instantly on the graphic image of the Patterson map. This allows rapid interpretation of the vector map and is especially important if the space group has high symmetry or there are multiple heavy atom sites.

Computer graphics*per se* is of relatively little assistance in the subsequent process of refinement of heavy atoms and single or multiple isomorphous replacement phase analysis. But once a rough electron density map of the DNA helix is calculated, graphics again proves extremely useful.

Construction of a Trial Structure into an Electron Density Map

The advantages of computer graphics in fitting a structure into a displayed electron density map have already been mentioned: greater accuracy in fitting the map, the ability to build in realistic bond lengths and angles from the outset, and the immediate availability of the coordinates in the computer.

Display of Intermediate Maps for Checking Errors in the Structure

No restrained least squares refinement process is automatic, and people who have assumed too much in this regard have made serious errors. The investigator must monitor the progress frequently during least squares refinement by examining difference maps ($2F_o - F_c$) to look for wrongly positioned groups or missing features. This is tedious when one uses contoured minimaps, but can be much less so on a computer graphics display.

The value of computer display of difference maps is well illustrated when one is examining the structure of a DNA-drug complex. In solving the structure of a B-DNA dodecamer of sequence C-G-C-G-A-A-T-T-C-G-C-G with the antitumor antibiotic netropsin, Kopka first positioned the DNA in the crystal unit cell, using the results from the DNA alone, and refined the DNA until no further improvement was possible (Kopka et al., 1985a, 1985b, 1985c). She then calculated a difference map of coefficients ($F_o - F_c$), where Fo represented the observed intensity data from crystals of the DNA-drug complex, and Fc represented the trial structure calculated from the DNA alone. The result is a “chicken wire” contoured image of the drug molecule (Figure 5-2). The known chemical structure of netropsin could be fitted easily and accurately into the graphics display, and refinement then continued to completion of the DNA and drug together. As a control, Kopka also drew a conventional minimap at the same point in refinement, but this was nearly uninterpretable because of the awkward orientation of the sectioning of the map and the difficulty of building an idealized drug molecule into a map of stacked plexiglass sheets. In this particular application, the conventional minimap was tedious, but the graphics display was very simple to interpret.

Location of Solvent Molecules and Ions Around a Macromolecule

The images of solvent molecules around a macromolecule cannot all be found from unrefined electron density maps. The quality of detail of the entire map improves as the fitting of the DNA is sharpened. Locating solvent molecules is a repetitive process that involves adding a restricted number of solvent peaks in the immediate neighborhood of the DNA, refinement, and examination of the improved map for new images of solvent. This process is shown inFigure 5-3 for the B-DNA dodecamer C-G-C-G-A-A-T-T-C-G-C-G. This particular analysis was carried out with minimaps because the computer graphics capability did not exist at the time, but recent DNA structure analyses use the more efficient graphics display.

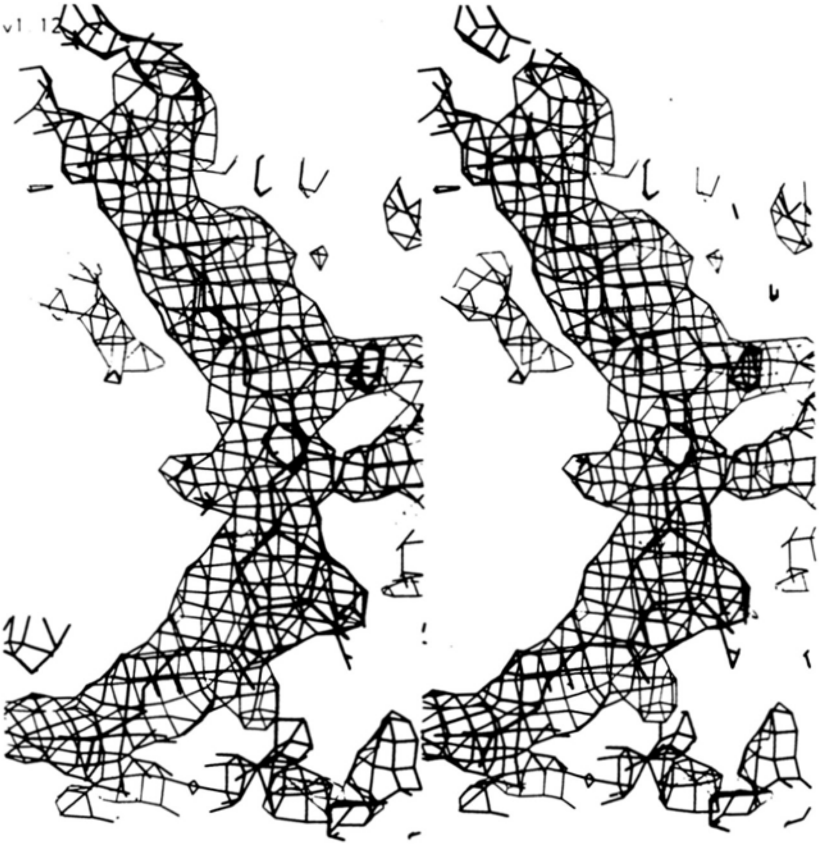


FIGURE 5-2 Stereo pair drawing, photographed off the face of an Evans and Sutherland Multipicture System graphics station, of the difference electron density of the antitumor drug netropsin in its complex with B-DNA. The screen image was photographed onto Ektachrome slide film, and this film then was used as a negative in making a positive print. The framework is a representation in three dimensions of one contour level in the electron density of the drug, and the graphics operator has built a skeleton of the netropsin molecule within this contour cage. This is the first point at which information about the drug was built into the analysis and was the point of departure for further least-squares refinement of the DNA-drug complex. Source: Kopka et al., 1985c.

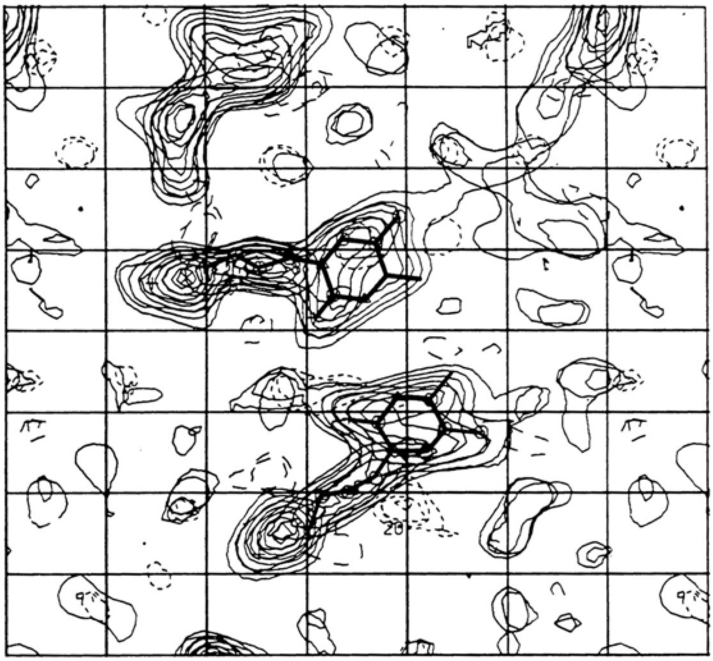


FIGURE 5-3 Illustration of the iterative process of locating solvent molecules around a DNA double helix. The quality of the solvent images improves with refinement and improvement of the phases used in the electron density map calculation. One must avoid adding "solvent" molecules too hastily, for fear of introducing erroneous peaks that then will persist and confuse during later refinement. Such a search for solvent requires the inspection of many successive electron density maps as refinement proceeds, and computer graphics are of enormous help in speeding up this process. (a) Vicinity of thymines No. 7 and 19 of the B-DNA dodecamer C-G-C-G-A-A-T-T-C-G-C-G, prior to the addition of any solvent molecules to the phasing. Residual error, R = 27 percent. (b) Later stage of refinement, R = 21 percent, three solvent peaks visible in these sections. (c) Still later stage, R = 20 percent, five solvent peaks visible. (d) Final map, R = 18 percent, ten solvent peaks indicated. Source: Drew and Dickerson, 1981.

Display of Completed Macromolecular Structure

The most familiar application of computer graphics to macromolecular structure is the display of the final results. But even here, the flexibility of computer graphics enables one to go beyond simple drawing of views of the molecule and see features that ordinarily would be overlooked. A case in point is provided by the stereo pair inFigure 5-4, which shows the drug molecule netropsin complexed with a 12 base pair B-DNA double helix. The image was oriented to sight directly down the minor groove, with the object of demonstrating that netropsin sits in the middle of the groove rather than to one side. But an unintended secondary dividend emerged. In this view, the bases of each individual strand of the double helix are seen to be stacked atop one another, almost as though the other strand of the helix did not exist. This efficiency of intrachain base stacking means that when the two strands are wound around one another to build a double helix, the bases of each base pair are not coplanar; they are given what is defined as a positive propeller twist about the long axis connecting them.

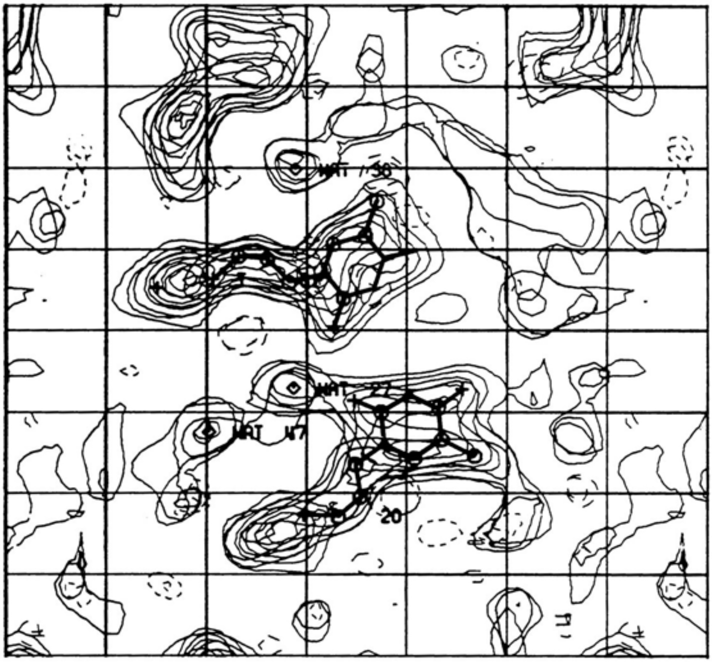


FIGURE 5.3b continued

The lessened resistance to propeller twisting in AT base pairs, which results because AT base pairs have two connecting hydrogen bonds, as compared to three for GC pairs, means that the minor groove of B-DNA is closed down more in AT regions than in GC. This makes a flat multiring drug molecule such as netropsin fit more snugly into the narrow AT region and is part of the explanation for the previously established binding of netropsin only to AT base pairs. Hence, in this example, an unusual view of the DNA that was intended to illustrate one point revealed an unexpected new association. Examples of this serendipity with computer graphics are found over and over again in macromolecular structure analysis.

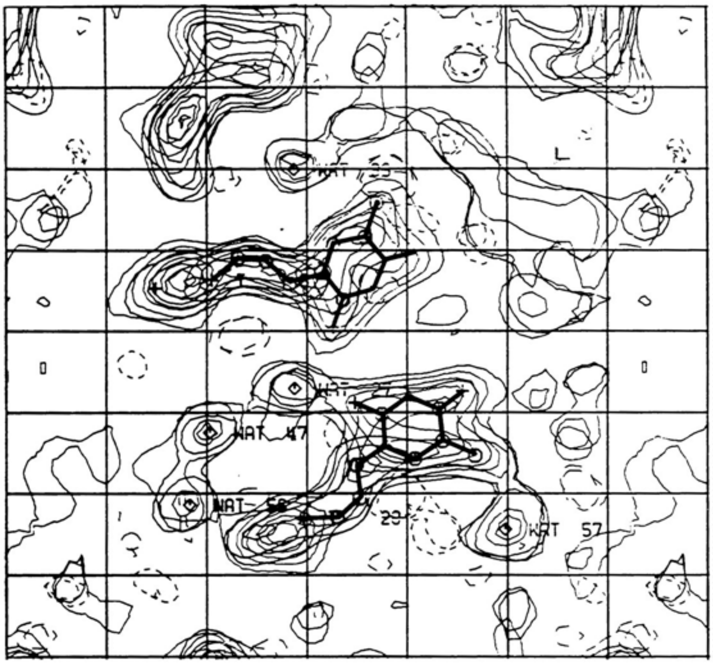


FIGURE 5.3c continued

Studies of Docking and Macromolecular Interactions

Once the structure of a macromolecule is known, it can be compared with those of related macromolecules or with other molecules with which it forms complexes. Computer graphics permits one to move molecules relative to one another and to introduce minor changes in a manner that would be tedious or impossible with physical models. For example,Figure 5-5shows the fitting of a netropsin molecule against the floor of the minor groove seen in profile. Without computer graphics, it would be virtually impossible even to represent the contour of the floor of the minor groove, yet the drug binds to this surface. The figure illustrates the structurally significant finding that the ends of the molecule are more closely associated with the DNA than is the central amide. One can change base pairs in the DNA from AT to 2-aminoadeninethymine and show the steric clash that then ensues with the drug. One also can modify the drug itself and see what effects this is likely to have on its binding to DNA.

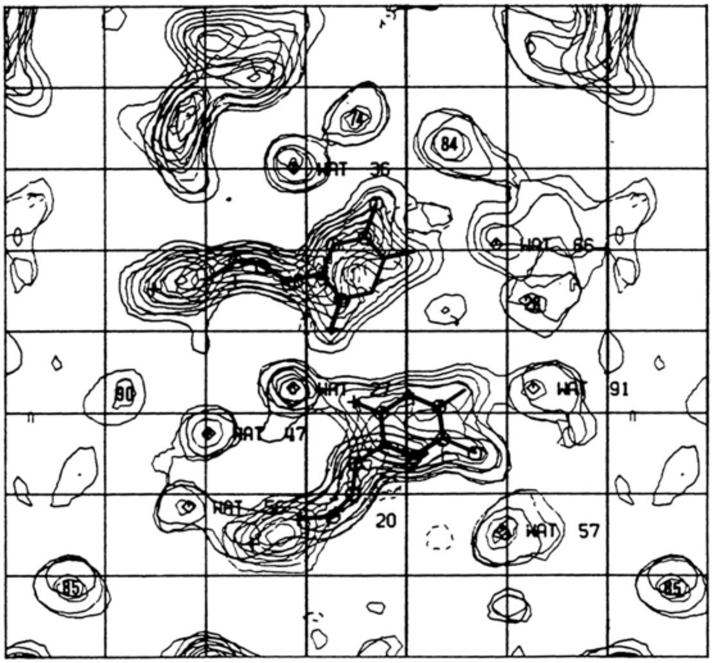


FIGURE 5.3d continued

This ability to examine intermolecular “docking” is equally important when the two molecules are known, but their complex is not. Trials of various docking geometries, with calculations of relevant energies, can suggest new modes of interaction, as well as the experiments needed to test them.

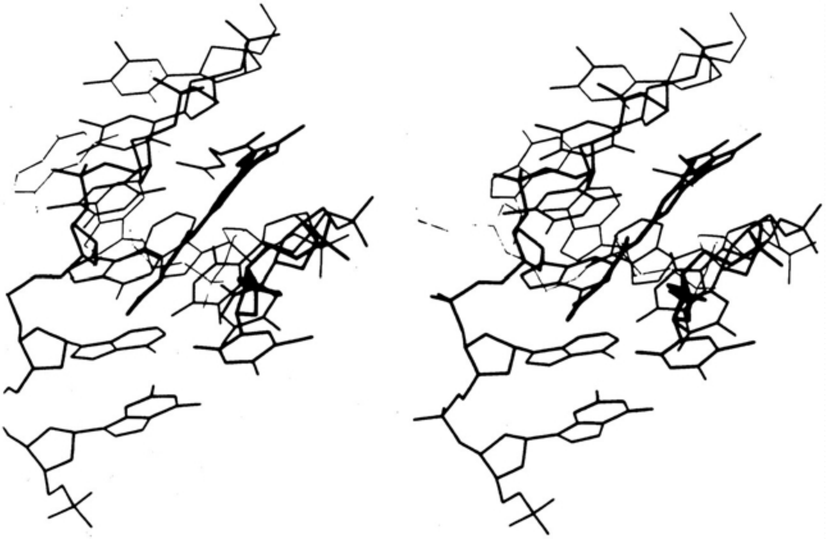


FIGURE 5-4 Oblique computer graphics stereo diagram of the complex of netropsin with C-G-C-G-A-A-T-T-C-G-C-G, in a view sighting directly down the minor groove with the drug molecule slotted into it in a crescent curving away from the viewer. This view also illustrates the strong stacking of bases down each strand of a Beta-DNA double helix, in a more striking representation than is obtained from a more conventional view of the Beta helix. Positive print from Ektachrome photo taken directly from graphics terminal. Source: Kopka et al., 1985a.

In summary, computer graphics is far more than an attractive way of drawing pictures of macromolecules. It is a very powerful tool that greatly aids in understanding the interactions between DNA and other macromolecules and so leads to insights that otherwise would be overlooked. Even without computer graphics, considerations of only energy minimization have led to a predicted structure of an enzyme-substrate complex (Pincus and Scheraga, 1979) that was subsequently verified by experiment (Smith-Gill et al., 1984).

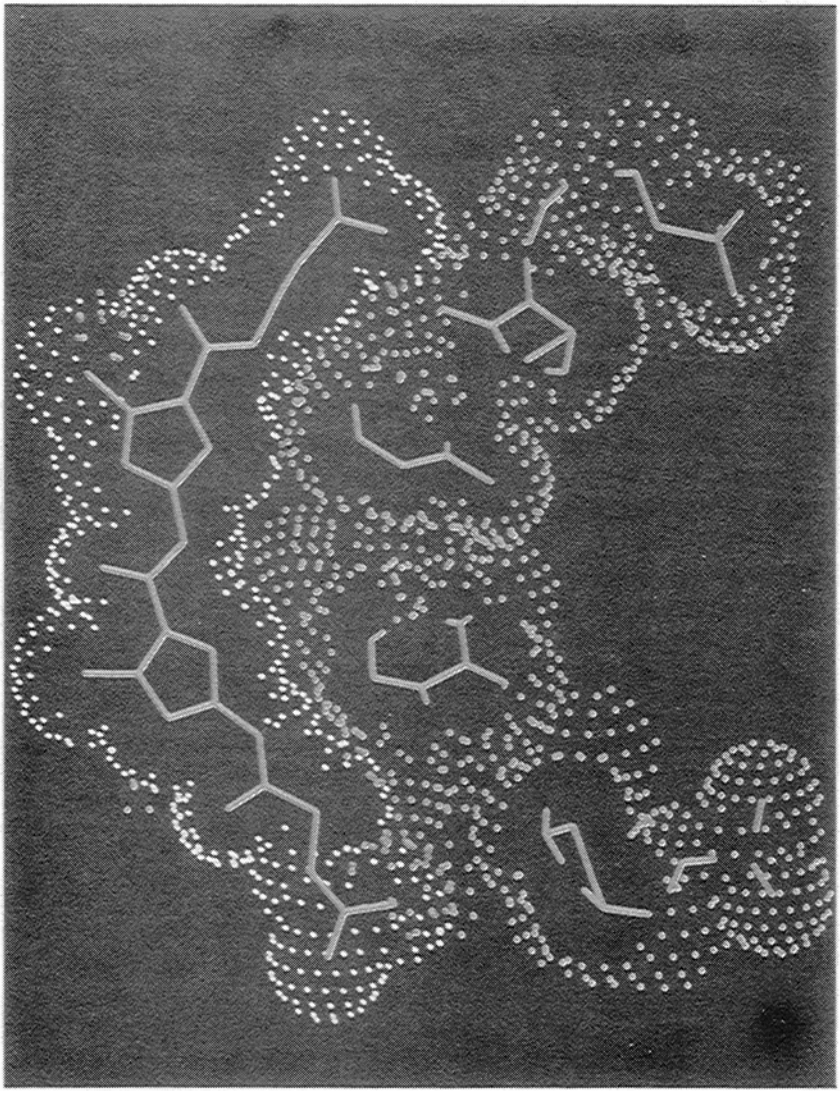


FIGURE 5-5 Van der Waals surface representations of the netrospin molecule (top) fitted along the floor of the minor groove of Beta-DNA. This highly unconventional view is possible only with computer graphics. It follows the plane of the drug molecule, and hence cuts the DNA base pairs only in oblique sections. The DNA major groove is seen at bottom nearly in profile, flanked left and right by phosphates. This view illustrates the less intimate contact with the floor of the minor groove at the center of the drug molecule than at the two ends. Source: Kopka et al., 1985c.

USING NUCLEAR MAGNETIC RESONANCE TO DETERMINE TERTIARY STRUCTURES

Nuclear magnetic resonance (NMR) is an important source for structural data, particularly when one considers its potential for meshing with theoretical modeling and other sources of structural data, such as x-ray crystallography. Its use has been growing explosively. NMR has been applied in aqueous media and, to a limited extent, in environments that approximate those of biological membranes (Braun et al., 1981). It is most useful when it is necessary to explore changes in preferred structure in response to environmental or structural perturbations. In this sense, NMR can play an important role in extrapolating structural data obtained by other techniques to alternate environments. It is also well suited to the exploration of changes in structure when comparing homologous series of macromolecules, such as a series produced by site-specific mutagenesis (Markley, 1987).

More recently, technological advances have extended the range of applicability to solids, oriented phases, and even total structure determination of molecules in solution. The latter extension has attracted the most attention and has the greatest potential impact on computer-assisted modeling efforts. Examples of total structure determination by NMR are most numerous among relatively small molecules: peptides, oligosaccharides, and nucleotide oligomers. However, several groups have determined peptide or protein structure for molecules in the 5 to 10 kDa molecular weight range (Arseniev et al., 1984; Braun et al., 1986; Havel and Wüthrich, 1985; Kaptein et al., 1985; Kline et al., 1986; Williamson et al., 1985). A brief discussion of the protein examples provides insight into the potential contribution NMR may make to the prediction of macromolecular structure and function in years to come.

Methodology

In most cases, the basis of structure determination by NMR methods is the interatomic distance dependence of the nuclear Overhauser effect (NOE). This is a nuclear spin relaxation phenomenon that depends on through space dipolar interaction between magnetic moments centered on different nuclei (usually protons). The NOE shows an inverse sixth power dependence on distance. In principle, Havel et al. (1979) found that the conversion of enough NOE measurements to distance constraints between pairs of protons in macromolecules would be completely equivalent to the specification of structure through a set of Cartesian coordinates. The ubiquitous occurrence of protons as the hydrogen nucleus in chemical structures insures an abundance of NOE data. These data have been gathered and used for a long time in studies of small molecules, but until recently, the sheer abundance of data for macromolecular systems has prevented unequivocal interpretation in terms of macromolecule structure.

A protein of 10 kDa will have approximately 1,000 protons, each giving rise to one or more resonances in a proton NMR spectrum. Resolution in conventional spectral acquisitions is in adequate to identify each resonance, let alone assign it to a primary structure site or acquire NOE data on a significant fraction of the possible 5×10^5 proton pairs. The situation with other types of macromolecules, nucleic acids or oligosaccharides, is less formidable in terms of numbers of protons. However, it is complicated because residues in these structures are less chemically diverse and the resulting resonances are less dispersed in NMR spectra and so it is difficult to assign a particular peak to a particular proton.

The resolution and assignment problem has been solved with the advent of higher field magnets (currently 14 Tesla) that provide increased chemical shift resolution, as well as by the devel opment of two-dimensional acquisition techniques that provide multidimensional resolution and greatly improve efficiency of data acquisition (Ernst et al., 1987). Wüthrich (1986) summarizes the methods for proteins and nucleic acids in his work.

Assignment of a resonance to a proton at a particular point in the primary structure site relies on a combination of experiments that display through bond scalar connectivities of resonances (COSY, or coupling correlated spectroscopy, for example) and through space dipolar connectivities of resonances (NOESY, or Nuclear Overhauser Effect spectroscopy). COSY is important in finding scalar coupling patterns that correspond to sets of spins that are characteristic of a particular type amino acid. For example, only alanine has three equivalent methyl protons coupled to an alpha-proton. The NOESY experiment is important in linking resonances assigned to a given amino acid to resonances of a neighboring amino acid. In cases where the amino acid sequence is known, the linkage of two to three amino acids is often enough to create a segment that occurs only once in the sequence. In principle, this procedure provides a complete sequence specific assignment of NMR resonances.

After the assignment of resonances, the determination of secondary and tertiary structures of proteins proceeds largely on the basis of NOE information. Qualitative analysis of the intensity of the crosspeaks that connect proton resonances involved in an NOE is often enough to characterize secondary structure. For example, in an alpha-helix, the amide protons of adjacent residues are 2.8 Å apart as compared to 4.6 Å in a beta-sheet (Wüthrich, 1986). Because NOE has a steep inverse distance dependence, this leads to strong amide-amide crosspeaks in an alpha-helix but crosspeaks that are a factor of 20 less intense in a beta-sheet (usually unobservable). It is possible to assign residue conformations to known types of turns as well as to more extended secondary structural elements. Since this assignment is sequence specific, it gives valuable information for verifying structure predictions and possibly assessing folding preferences even without a tertiary structure determination.

Determining tertiary structure is more difficult. To determine a structure, a limited amount of longer range distance constraint information must be systematically integrated with other constraints. The most direct approach probably employs a distance geometry search for structures that have interproton distances between the experimentally determined upper and lower bounds (Braun and Gō, 1985; Havel et al., 1979). Altman and Jardetzky (1986) recently developed an expert system that has also succeeded in producing a space-filling representation of protein structure.

It is, however, proving desirable to increase the degree to which theoretical constraints play a role in determining structures. This increase has been achieved using both molecular dynamics and molecular mechanics-based programs (Brunger et al., 1986b; Kaptein et al., 1985). It is desirable to integrate NMR data with theoretical predictions for two reasons. First, NMR data often leave certain regions of the macromolecule poorly defined; for example, there may be little NOE data on certain sidechain conformations in proteins. Theoretical predictions can help specify the conformation of these regions. Second, use of some experimental data seems a viable approach to selecting among multiple minima in the complex energy surfaces that must be searched by theoretical modeling programs.

Viability of Approach

As with any new approach to structure determination, NMR-based methods are being tested for viability. The most direct method of evaluation is to compare structures determined independently by the new methodology with an established technique such as x-ray crystallography. Unfortunately, when Braun and coworkers (1986) first determined a structure without the aid of an existing crystal structure (metallothionein, a 7,000 Da protein) the structure varied dramatically from the x-ray structure when it did appear. The reasons for the differences, although still unresolved, are more likely the result of actual structural differences in the samples examined than of a flaw in the methodology. More recent comparisons show excellent agreement of structures determined from real or simulated NMR data with structures determined by x-ray crystallography; examples include an alphaamylase inhibitor, a 8,000 Da protein (Kline et al., 1986) and crambin, a 5,000 Da protein (Laue et al., 1985).

Resolution is difficult to define in NMR structures because some aspects, such as the conformation of the backbone, are determined very precisely, while others involving sidechain conformations are poorly defined. It is even possible that the lack of adequate numbers of distance constraints will leave some regions completely undetermined. In principle, distances can be determined precisely, within 0.01 Å, but this is only accomplished when relaxation processes are very well defined and interproton distances are short. Recent estimates for general resolution based on root-mean-square (*rms*) deviations of heavy atoms in multiple structure solutions of proteins obtained from NMR data suggest resolution to be approximately 3 Å (Williamson et al., 1985). Although these average deviations are larger than those usually seen in x-ray data, NMR methods can be used in a variety of media and yield very precise distance information on selected distances. Both of these advantages compensate for the lower precision.

The effort required to produce a structure is difficult to assess in this early stage of development. The first few tertiary structures probably required several person-years of effort. However, this is dropping rapidly. Secondary structures have always been produced with far less effort. To produce a tertiary structure of a protein in the 10 kDa range, one month of spectrometer time and six months of analysis is a reasonable estimate. Sample requirements are modest; 50 mg of a soluble 10 kDa protein. Smaller molecules require a proportionately smaller sample or a quadratically smaller investment in time. It is important to realize that this is an emerging technology, compared to better established structural methods. Large opportunities are available to improve efficiency and viability.

Limitations and Prospects for the Future

The major limitation to the above methods appears to be one of accessible macromolecule size. Current applications in which near-complete assignments are made and structures are determined appear to be restricted to proteins of 10 kDa and less. This is due in part to inadequate resolution when thousands of connecting peaks are involved. It is also due to loss of sensitivity in one of the key spectral assignment data sets (COSY sets). COSY crosspeak intensity is extremely dependent on the ratio of scalar coupling constants to linewidth. The linewidths increase rapidly with molecular weight, leading to loss of signal. It is significant that the principal source of distance information, the NOESY experiment, does not have the same degree of sensitivity loss as one increases macromolecule size. If other assignment strategies emerge and resolution of chemical shift can be improved, it will be possible to use NMR methods on larger macromolecules.

We believe that improved assignment strategies will emerge. Already significant work has been done replacing normal amino acids with isotopically substituted amino acids in order to assign peaks arising from particular amino acid types (Kainosho and Tsuji, 1982; LeMaster and Richards, 1985). This strategy by-passes some of the dependence of assignment on NOESY spectra and can be applied to proteins of more than 12 kDa (Kainosho and Tsuji, 1982; LeMaster and Richards, 1985). Replacement of normal amino acids with amino acids that contain 15N and 13C and the use of indirect detection methods to improve sensitivity also allow use of the increased chemical shift dispersion displayed in spectra of other nuclei. Proteins of 19 kDa (McIntosh et al., 1987) and 23 kDa (Kainosho et al., 1987) are under study. In these cases, the isotopic labelings were easily performed because the proteins were obtained from microorganisms. These developments, along with the probable advance in resolution from higher field magnets, make it likely that general structure determination methods may be applied to proteins in the 20 kDa class within the next five years. Application to proteins as large as 60 kDa, where questions are focused on specific sites, already are possible. A rather limited quantity of residue-specific information may be required to improve dramatically the quality of theoretical predictions.

A second limitation, to current methodology beyond accessible macromolecule size, stems from the restricted range of measurable interproton distances, <4 Å. Although this range is adequate to determine structures of short segments, tertiary structures of larger systems are frequently the result of successive application of many short range constraints. This introduces the possibility that significant errors will be propagated. Here again, optimism is justified. The use of paramagnetic labels leads to perturbations of spectra interpretable in terms of distances over separations of more than 10 Å (Kosen et al., 1986).

A third limitation occurs because the conversion of NOE measurements to distances requires assumptions about macromolecule rigidity. Since some portions of macromolecules are not rigid, the distances extracted for those portions are imprecise. To some extent, the imprecision is reduced by the $1/r^6$ dependence of the NOE. An error of ± 50 percent in an NOE ratio for a 3 Å contact converts to error limits of −0.21 and +.33 Å. It is nevertheless an important limitation. At present, these potential errors are handled by assigning generous distance constraint limits, rather than by trying to specify distances precisely. In the future, it may be possible specifically to include dynamic information eliminating this necessity.

AREAS OF POTENTIAL IMPACT OF NMR

Proteins and Peptides

It is clear from the above discussion that much of the research into the structure and dynamics of macromolecules, particularly proteins, may be restricted to relatively small members of the class. This limitation is imposed both by the magnitude of the task and the loss of sensitivity as macromolecule size increases. We should consider how these restrictions may affect biological science and what special problems or opportunities may arise in considering smaller members of macromolecule families.

Limiting research over the next five years to proteins of molecular weight 10 kDa or less may seem highly restrictive. However, a survey of the current protein sequence data base maintained by the Protein Identification Resource shows a surprisingly large fraction (20 percent) to be under this size (Barker et al., 1986). This fraction may be inflated by the ease of handling shorter sequences, but even 10 percent of the current 4,000 sequenced proteins is a very large number to be studied by NMR spectroscopy. The current rate of physical characterization stands at 20 proteins per year by x-ray crystallography and 10 or more by NMR. Given that rate and the likelihood of producing sufficient quantities of many of the sequenced proteins by cloning techniques, human resources and characterization facilities are more likely to be limiting factors than is the number of proteins on which we want information.

Beyond the issue of numbers, small proteins and even smaller peptides constitute a physiologically important class. Polypeptide hormones such as atrial naturietic factor, oxytocin, vasopressin, and insulin (subunit) fall in these classes (Wallis et al., 1985). Various neurotoxins are small polypeptides, and a number of endogenous opioid peptides exist, such as the enkephalins and endorphins.

One may also ask whether studies of small protein or peptide structure might be relevant for the understanding of larger structures. Although some behavior, such as allosteric interaction, is certain to be poorly represented in small molecules, the basic structural considerations are reasonably likely to carry over, and it is likely that fundamental processes such as protein folding can be studied. A large number of proteins are composed of smaller subunits that largely maintain their structure when isolated. In some single chain proteins, structural domains that can be cleaved and functionally reconstituted without reforming a covalent linkage can be identified (Rose, 1985).

Small peptides present some potentially unique problems for conformational studies. It is not clear that the dominant conformations observed in solution, crystals, or simulations are those that are important for biological function. These molecules do interact with receptors, which may select a minor conformer or dictate a unique conformation that is a function of the properties of both molecules. Because most receptors exist in very small numbers, it has only recently become possible to produce enough for physical study. For the coming years, a major challenge is the modeling of systems in which both activator and receptor can change conformation. This challenge relates not only to the understanding of physiologically important molecules, but also to the design of pharmacologically important molecules. It is possible that a study of relatively flexible polypeptides will contribute to the development of the methods needed to meet this challenge.

Experimental measurements designed to explore minor conformers should contribute in important ways to the understanding and development of modeling methods. An experiment using NMR where this type of study appears possible offers one example. Small molecules (1-2 kDa) have relatively inefficient pathways leading to nuclear Overhauser enhancements. When these small molecules bind to macromolecular receptors, more efficient path ways are present. Distance constraints derived from the measured NOEs on molecules that exchange rapidly between bound and free states therefore pertain to the conformation of the bound molecule more than the free molecule (Clore and Gronenborn, 1983). Such experiments have many limitations of size, binding constants, and rates of exchange, but they still promise to make some inroads into a very difficult set of problems in molecular biology.

Nucleic Acids

As a class, intact nucleic acids exist as molecular entities in sizes far larger than the proteins with which we typically work in NMR. However, smaller segments appear to display structural characteristics important in biological function. For example, double helix structures formed from oligomers of 8 to 12 bases in length exhibit variations in backbone torsion angles, base twist, and base tilt angles that characterize helix forms suspected to modulate transcription activity.

A dodecamer helix has an effective molecular weight equivalent to small proteins. The same two-dimensional NMR methods applied to proteins allows us to characterize nucleic acid structures and explore in detail the factors that lead to interconversion of structural forms. The imino protons involved in the hydrogen bond connecting base pairs exchange slowly with protons in the solvent and are easily resolved in the low field region of a proton NMR spectrum. Sequential assignments are made possible by the proximity of imino protons on adjacent base pairs and the strong cross relaxation peaks that connect these resonances in NOESY spectra. Variation in distances between imino protons and sugar backbone protons—for example, H6 protons on pyrimidine bases and H2' protons on the attached deoxyribose ring—makes it possible to distinguish the type of helix on the basis of the presence or absence of cross relaxation peaks. More quantitative treatments of structure employ the same molecular mechanics, molecular dynamics, and distance geometry methods used with proteins (Hare and Reid, 1986; Nilsson et al., 1986; Suzuki et al., 1986).

Beyond simple helical structures lies a vast region of structural biology of nucleic acids that has been much less explored. Nucleic acids are important elements of ribosome structure and function. Some ribonucleic acids have even been shown to exhibit enzyme-like activity. Hybrid systems involving proteins and nucleic acids are important in synthesis, repair, and regulation of protein structure and function. In the near future, we should have the potential for structural characterization by NMR of at least parts of these systems. In building a more theoretical basis for the extrapolation of primary structure to three-dimensional structure and function, one key step is the experimental verification of the existence of fundamental structural elements and understanding of the factors that dictate their occurrence.

Carbohydrates

The carbohydrate moieties of glycolipids and glycoproteins are a third class of biologically important molecules. They are important modulators of communication between and across cell membranes. A knowledge of three-dimensional structure and flexibility is essential to understand this modulating function. Because the function of these molecules has been less widely appreciated, they are discussed in more detail later in this report.

Most oligosaccharides, if examined in isolation, are smaller than the protein and nucleic acid systems we have been discussing. Structure determination is, however, no less challenging because of the structural diversity of this class of molecules. The number of monosaccharide building blocks is intermediate in number between that of proteins and nucleic acids, but each monosaccharide can be linked through any one of several sites, with either of two anomeric configurations. That there are multiple linkage sites also opens the possibility of branching. It is difficult to predict the number of structurally distinct oligosaccharides in an organism, but it is certainly large. Even if we attend only to a single class of oligosaccharide containing-molecules, such as glycosphingolipids, a significant number (more than 150) of primary structures have been determined (Hakomori, 1986). Few of these structures have experimentally determined tertiary structures, and all evidence indicates that 150 represents a small fraction of the total number that exist. Also, unlike the protein and nucleic acid problems, carbohydrates present a primary as well as a tertiary structure problem. NMR methodology has contributed to the solution of both and, with some advances, should contribute further (Yu et al., 1986).

Analogies with protein structure determination exist. COSY spectra are important in assigning resonances to particular residue types. NOESY spectra are important in identifying linkage sites, linkage configuration, and sequence. Bottlenecks in terms of manual assignment and conversion of cross relaxation data to structures are very similar to those encountered in protein studies (Bush et al., 1986; Yu et al., 1986; Homans et al., 1987; Dabrowski et al., 1986). Progress has been impeded slightly more because computer modeling programs for oligosaccharides are somewhat less refined than are those for proteins. Oligosaccharides are also likely to be less rigidly structured and so require more attention to proper treatment of motional averaging. Nevertheless, significant advances are possible, at least in smaller molecules and more sterically constrained molecules of this class. With the establishment of a larger experimental data base using methods such as two-dimensional NMR, theoretical predictions of structure from sequence should become possible.

DEMAND ON COMPUTATIONAL FACILITIES

Improved computational and molecular modeling facilities could advance structure determination using NMR methods in several ways. Processing and analysis of NMR data for a 10 kDa macromolecule is far more time-consuming than is data acquisition. Each phase of this operation could be improved. Data are normally collected as a two-dimensional time domain set and processing involves Fourier transformation to a frequency domain set. These processes are now handled by array processors associated with instrument computers with a moderate investment in time (tens of minutes). However, alternative methods of processing, including linear decomposition and maximum entropy methods, may be more advantageous in terms of signal to noise and may be more compatible with automating analysis (Laue et al., 1985; Schussheim and Cowburn, 1987). These require far more computer time and may become practical only on supercomputers.

Automating analysis can be difficult when working directly with a frequency domain data set. The sets are 4 to 16 megawords in size, and the connectivity peaks used in assignment are complex shapes that have both frequency and phase information. Several approaches are being explored to reduce these complex peaks to a few pieces of connectivity information, but it is possible that methods such as linear decomposition, which reduce frequency domain sets to lists of peak frequency and intensity at an early stage, will provide the breakthrough needed in this area.

Once resonances have been cataloged, the next step is the assignment and extraction of spectral characteristics important for the determination of secondary and tertiary structures. This process is now mostly done by hand. Some efforts are underway to use semiautomated pattern recognition and expert system strategies, but these will require substantial investments in programming and computer hardware (Pfandler and Bodenhausen, 1986).

At present, conversion of the data to a three-dimensional structure is handled by distance geometry or one of the pseudo energy approaches. Most such programs that currently run on supercomputers require one to several hours of computer time per trial structure.

At least 5 trial structures should be generated for each data set to explore the constraint boundary conditions. As interest and capabilities for the production of data sets increase, the demand for computer time will become staggering. It is difficult to project if and how much these demands will be offset by improved algorithms. An investment in programming is obviously warranted. This investment could be put to best use by acknowledging that, in the future, it may be desirable to accommodate types of data not used today. Some data may come from other NMR methods that are applicable to solids and oriented specimens. Other data may come from entirely different methodologies, such as fluorescence spectroscopy or tunneling microscopy. Although modeling and energy refinement programs have existed for years, it is clear that attempts to accommodate NMR data have met with some obstacles. Most modeling programs presume the existence of a

Cartesian coordinate set similar to that obtained from an x-ray structure. NMR data are most compatible with interatomic distances or interactive manipulation of segments of known secondary structure.

Present methods, except in a few preliminary calculations, also assume that structure of biomolecules can be described in terms of a single rigid conformer. This is certainly not true. Relaxation of these assumptions will lead to drastically increased computational demands. Some of the problems and opportunities related to this are discussed more fully under the section on molecular dynamics.

In summary, current NMR methods to determine structure seem applicable to a variety of biologically important molecules of less than 10 kDa. Data production in this class will be made much easier by improved computational facilities, available high field spectrometers, and attention to the compatibility of modeling programs with experimental constraints provided by NMR. It is important that those working in NMR collaborate with investigators attempting to determine structure with other methods, because such links will increase the structural data base and also enhance testing and theoretical modeling. It is likely that the range of molecules accessible by NMR methods will increase by a factor of two over the next five years. The rate of data production is likely to increase even more as structure determination protocols are improved and high-field spectrometers become more generally available. Advances in high-temperature superconductors may accelerate this process.

6.

Tertiary Structure of Proteins and Nucleic Acids: Theory

ENERGY OPTIMIZATION

According to the thermodynamic hypothesis, based on Anfin sen's (Anfinsen et al., 1961) experiments on the oxidative refolding of bovine pancreatic ribonuclease A from the reduced form, the amino acid sequence determines the three-dimensional structure of a protein in a given medium as the thermodynamically most stable one. It must be emphasized that this hypothesis applies to the length of the polypeptide chain at the stage when folding takes place, and not at some subsequent processing stage. For example, this hypothesis would be applicable to the single-chain pro-insulin and (possibly) not to the processed product, the two-chain disulfide-linked insulin. Thus, it is a challenge to chemists to understand how the interatomic interactions within the polypeptide chain and the interactions between the atoms of the chain and those of the surrounding solvent lead to the thermodynamically stable structure, that is, the one for which the statistical weight of the system is a maximum.

The hypothesis that the statistical weight is a maximum immediately implies that some kind of optimization strategy is neces sary to find the most stable structure. This requires procedures to generate arbitrary conformations of a polypeptide chain, compute the statistical weight for each conformation, and then alter the conformation so that it ultimately corresponds to the global maxi mum of the statistical weight. Although the problem is formidable, current indications are that it can be solved using a sound scien tific approach without resorting to ad hoc procedures to deduce "folding rules" that do not explain the molecular Basis of such "rules."

To compute the global statistical weight, one optimizes the conformational energy of the polypeptide, incorporates the effect of the solvent, and takes account of the entropy of the system. Procedures are available for carrying out such computations, and currently available supercomputers permit the computations to Be applied to very large systems. Although these procedures will undoubtedly improve, they are now adequate for computating and obtaining results that can be checked experimentally. The major difficulty still to be overcome, although partial success has been achieved (see review by Gibson and Scheraga, 1988; also, Robson, 1986; Crippen, 1984), arises from the presence of many local min ima in the multidimensional energy surface. Although algorithms are available for minimizing an energy function of many variables, there are no efficient ones for passing from one local minimum, over a potential barrier, to the next local minimum—and ulti mately to the global minimum of the potential energy in a very high dimensional space. Thus, minimization leads to the nearest local minimum, where the procedure is trapped. This trapping in a local, rather than the global, minimum is referred to as the "multiple-minima problem." Efforts are being made to overcome this problem using a variety of procedures, including approximations that place the system in the potential well in which the global minimum lies (Gibson and Scheraga, 1988). Then, any approxi mations (introduced in the initial stages) are abandoned, and the full energy function is minimized. The use of molecular dynamics for minimization is an alternative strategy; it is considered in the next section.

The energy minimization approach and its associated compu tational program, described here for polypeptides and proteins, is equally applicable to any other type of macromolecules, such as polynucleotides and polysaccharides, as well as to interactions between the various types of macromolecule. General references to this methodology include the works of Anfinsen and Scheraga

(1975), Némethy and Scheraga (1977), Levitt (1982), Karplus and McCammon (1983), Scheraga (1984), and Richards (1986).

Generation of an Arbitrary Conformation

Analytical geometry and associated matrix algebra provide the tools to generate a polypeptide chain. To do so, internal co ordinates (dihedral angles for rotating about bonds) or Cartesian coordinates may be used as independent variables. When inter nal coordinates are used, the bond lengths and bond angles are held fixed, but chosen very carefully from x-ray structures of model systems so as to properly reflect the geometric results of side chain-backbone interactions within each amino acid residue. The validity of this approach has been demonstrated for both polypeptides and polysaccharides—systems in which strained conformations rarely arise (Scheraga, 1984). When Cartesian coordinates are used, and hence bond lengths and bond angles are allowed to vary, one faces the problem that force constants for these motions are not as well known as the geometric features of the molecule. Further, bond-angle bending modes are anharmonic, and no currently available forcefield takes anharmonicity into account. Thus, we will have to overcome inadequacies in the force constants and problems of anharmonicity before we can rely on computed bond angles. Impo sition of fixed bond lengths and bond angles is at least reconciled with observed crystal structures of small molecules. The subject of fixed versus flexible geometries has been discussed by Swenson et al. (1978).

Potential Functions

Many research groups have developed potential energy functions with which to carry out such computations on polypep tides, polysaccharides, polynucleotides, and synthetic polymers. The various potential functions have many similarities, but differ enough in detail to make them difficult to compare. This difficulty is compounded by the small number of cases for which parameters that characterize the strengths of the various interactions have been established in a self-consistent way on experimental data such as crystal structures, lattice energies, and barriers to inter nal rotation. At present, the potential functions for polypeptides, polysaccharides and synthetic polymers have been better parameterized than have those for polynucleotides, primarily because there are fewer reliable data for model nucleotide compounds. Similarly, more experimental data will be required for proper parameterization of the potential functions for the prosthetic groups attached to biological macromolecules. Until recently, the poten tial functions involving water molecules still would not account adequately for the observed radial distribution function of water, but this situation is improving (see, e.g., the work of Jorgensen et al., 1983).

It must be emphasized that the molecule responds to the total (in principle, quantum mechanical) potential function, and its partition into various empirical components (such as non-bonded, electrostatic, hydrogen-bonding, and other interactions) is at best arbitrary, although there is some physical basis for as signing such names to the various components. Consequently, one must avoid combining components from forcefields from different research groups. Each forcefield must be parameterized by itself in a self-consistent way. The total energy of a given conformation is expressed as a sum of the energies of all nonbonded pairs of atoms.

When considering biological function, such as the formation of an enzyme-substrate complex, then the pair interactions both within and between the two partners of the complex must be included in the total energy. Thus, the influence of each member of the complex on the other (the so-called induced fit phenomenon) is taken into account. Consequently, the conformations of the partners in the complex can differ from their conformations as individual species. This means that the computed conformation of an isolated hormone may not resemble the biologically active one when it is bound to a receptor (Scheraga, 1984).

Although potential functions could still be improved, those for which parameters have been determined in a self-consistent way have led to many computed structures that have subsequently been checked by experiment. For example, the computed structure of the collagen-like poly(Gly-Pro-Pro) (Miller and Scheraga, 1976) agrees with the subsequently determined crystal structure (Okuyama et al, 1976) within a root-mean square (rms) deviation of O.3 Å. Scheraga (1984) has cited many similar examples of experimental verification of computed structures. Therefore, we can be confident that the problem of adequate potential functions is not serious, and more effort should be devoted to the most difficult problem of all—the multiple-minima problem. After the multiple-minima problem is solved, it will then be worthwhile to reexamine the possibility or necessity of improving the potential functions further.

Solvation

There are several methods to include the effect of hydration in the computations. Hydration tends to force the polar groups to the surface of the molecule, putting them in contact with water, and forces the nonpolar groups to the interior, removing them from contact with water.

One method includes the water molecules explicitly, and calculates the interaction energy between the molecule and the water. The success of this approach depends on the adequacy of the po tential function describing the water-water interaction. Another approach ignores the structural features of the water molecule, and assigns a hydration shell (and an accompanying free energy of hydration) to each atom or group of atoms. As the conformation changes and hydration shells overlap, a free-energy penalty is assessed. The hydration-shell model is parameterized on experimental data on free energies of hydration (Kang et al., 1987), but current efforts are being made to obtain such free energies from Monte Carlo and molecular dynamic studies of aqueous solutions of small molecules (see, e.g., Jorgensen et al., 1983). The compu tation of free energies by these simulation techniques faces many theoretical obstacles, although this very active field of research is progressing quickly.

Entropy

A variety of methods exist to incorporate entropic effects. One entropic effect arises from the hydration; this effect is treated as described earlier. The other entropic effect arises from the conformational fluctuations of the molecule. Several procedures may be used to compute this contribution, the most direct being the evaluation of the second derivative of the potential function. This describes the curvature at the bottom of the potential energy well and hence the fluctuation in conformation about each local minimum in the potential energy surface. Local minima that are not the global minimum of the potential energy can become the conformations of higher statistical weight if there is a large enough entropy gain from large conformational fluctuations. Paine and Scheraga (1987) have encountered such effects.

Optimization Procedures

Optimization procedures are available for searching for the local minima. These include direct energy minimization, Monte Carlo, and molecular dynamics procedures. In energy minimization, the variables describing the conformation are altered system atically so as to lower the energy continuously. The Monte Carlo method makes random changes in the conformational variables and accepts the new conformation according to various protocols that compute the energies before and after the random changes in the conformation. In molecular dynamics calculations, Newton's equations of motion for the atoms of the macromolecule (subject to interatomic forces determined by the potential functions) are solved to obtain a trajectory in conformational space. Very efficient energy minimization algorithms exist, even for functions of many variables, but lead only to local minima (however, see the following section). Conventional Monte Carlo procedures can overcome local minima but tend not to cover conformational space efficiently enough. However, this difficulty is being overcome using modifications that include adaptive importance sampling and other efficiency-seeking procedures (Gibson and Scheraga, 1988). Molecular dynamics can also surmount local barriers, but the pi cosecond time scale of practical computations does not approach the millisecond time scale of actual protein folding.

Because most published papers do not provide the relevant data, it is difficult to compare the computer time needed for various optimization procedures. The required computation time will be very sensitive to how well the code is optimized, whether parallel processing is carried out, and other factors. To obtain such benchmarks, it would be necessary to run each procedure on several different computer systems—a task not yet undertaken.

Solutions to the Multiple-Minima Problem for Macromolecules

Since no mathematical procedures are available to locate the global minimum for any macromolecule (except in energy surfaces of very low dimensionality), as mentioned above we must first resort to approximate procedures to obtain structures that might lie close to that of the native macromolecule. Then, the approximations are abandoned, and full-scale energy minimization, Monte Carlo, or molecular dynamics procedure is carried out. A variety of such procedures have been developed (Gibson and Scheraga, 1988). These include:

- a "build-up" method, in which large structures are built up from ensembles of low-energy conformations of smaller ones (with energy minimization being carried out at each stage);
- optimization of electrostatic interactions;
- optimization in a space of high dimensionality where fewer intervening barriers exist (with subsequent relax ation back to three dimensions);
- Monte Carlo sampling among local minima (accompa nied by energy minimization);
- adaptive importance Monte Carlo sampling (to drive the system more efficiently to the global minimum);
- pattern recognition methods to assemble organized back bone structures as alpha-helices and beta-sheets;
- use of distance constraints either from experiment (Nu clear Overhauser Effects (NOEs), nonradiative energy transfer, or NMR on spin-labeled molecules) or from statistical analysis of x-ray structures of proteins; and
- empirical methods to predict the locations of alpha helices, beta-sheets, and beta-turns.

Numerous valid predictions of global minimum structures of peptides have been made using these methods (Gibson and Scher aga, 1988). However, they have thus far been successful only for structures that contain at most 20 residues, and current efforts (most of which require access to supercomputers) are being made to extend these methods to larger molecules—to proteins containing on the order of 100 amino acid residues.

Successes and Failures

Numerous structures have been predicted and subsequently confirmed experimentally (Scheraga, 1984). The right- or left-handed twists of the fundamental structures (alpha-helices and beta-sheets) from which proteins are built have been accounted for by energy minimization. The observed packing features of alpha-helices and beta-sheets have likewise been accounted for in energetic terms. Parameters calculated for conformational tran sitions (e.g., the helix-coil transition in water) have been verified by experiment. The computed structures of open-chain and cyclic molecules (e.g., the 20-residue membrane-bound portion of melit tin and the 10-residue gramicidin S, respectively), and those of collagen-like poly-tripeptides have also been verified by experi ment (Scheraga, 1984). Finally, the computed structure of an enzyme-substrate complex (hen egg white lysozyme and a hexas accharide substrate) (Pincus and Scheraga, 1979) has been verified by experiment (Smith-Gill et al., 1984). These and other examples should give us confidence in the validity of the potential functions and computational methodology (Gibson and Scheraga, 1988).

The failures, in the sense of not yet having solved the protein-folding problem, exist because no one has yet used optimization techniques to deduce the three-dimensional structure of even a small protein, such as the 58-residue bovine pancreatic trypsin inhibitor (BPTI). Current procedures applied to BPTI have not yet yielded a computed structure that comes closer to the x-ray structure than 2-3 Å. Several procedures that work to overcome the multiple-minima problem on small molecules become compu tationally intensive as they are used on larger molecules. However, the increasing use of supercomputers will help overcome this prob lem.

Impediments to Progress

Although supercomputers will allow larger calculations and thus cover conformational space better, workers in this field will need additional time to be allotted on these machines to do the research necessary to achieve greater efficiency. Parallel processing offers a breakthrough, and this will require new software to take advantage of the hardware enhancements. With new Hardware and software, it should be possible to surmount the major hurdle created by the multiple-minima problem. However, it is conceivable that bottlenecks may develop as we attempt to scale up procedures that work on 20-residue segments to proteins containing 100 to 200 residues. We will also need imaginative new approaches to overcome this problem.

Potential functions should be improved, especially those for polynucleotides and prosthetic groups, and for water-water inter actions, but this is not now the most serious problem. Certainly, this problem should be addressed again when the multiple-minima problem is solved for bovine pancreatic trypsin inhibitor.

Finally, new developments will be needed to bring molecular dynamics from the picosecond to the millisecond time scale.

Future Prospects

At every stage in the development of conformational energy calculations over the past 25 years, we always seemed to face in surmountable obstacles. However, the steady progress during this period indicates that many of these obstacles have been overcome. The remaining major hurdle is the multiple-minima problem (Gib son and Scheraga, 1988), but we have an array of possible solutions to it. The solutions have worked for small molecules, and current and impending developments in computer hardware and software should justify our confidence that, within 5 to 10 years, we may ex pect to understand how interatomic interactions dictate not only the final folded structure but the pathways taken by the newly formed polypeptide chain to reach the native structure.

MOLECULAR DYNAMICS

The principle behind a molecular dynamics simulation is sim ply the application of Newton's equations of motion to the atoms of one or more molecules. Newton's equations relate three in dependent quantities: time, conformation (atomic coordinates), and potential energy. As the calculation progresses and the positions and velocities of the atoms change, the system will traverse many different states; as the simulation is prolonged, the observed states together approach a perfect sample of the thermal equilib rium ensemble of all states the system will occupy. The thermal equilibrium distribution may also be sampled without considering motion, using appropriate purely statistical methods (Monte Carlo techniques). In principle, a Monte Carlo calculation might produce a representative sample using less computer time. Noguti and Gō (1985) indicate how, with knowledge of the second-derivative matrix of the potential energy, the atomic coordinates can be ef fectively used to speed up the Monte Carlo process. However, it is as yet uncertain whether this accelerated Monte Carlo procedure produces a more rapid exploration of conformation space of a protein than a molecular dynamics simulation. Thus, the molecular dynamics simulation gives us a way to make theoretical estimates of mean atomic positions and deviations from the mean; of rates of motion and conformation change; and of ensemble averages, including thermodynamic functions such as energy, enthalpy, spe cific heat, and free energy. Since free energies can be expressed as equilibrium constants and vice versa, simulations are being used to obtain theoretical estimates of differences of affinity of proteins for small molecules. Recent results show remarkably good agree ment with experimental values. Major pharmaceutical companies have already noted the usefulness of accurately predicting these differences.

Molecular dynamics simulations, although simple in concept, were not practical before the advent of high speed computers. This method of theoretical chemistry is particularly useful for the study of condensed phases and was first used to study the structure and dynamics of liquids. Later, several investigators applied existing techniques to protein molecules (Karplus and McCammon, 1983; Berendsen [cf. Hermans, 1985, Beveridge and Jorgenson, 1987]). At present, several laboratories are active in the field. More are be coming involved as the methods are applied to increasingly quan titative studies that aim to reproduce experimental observations as closely as possible in the computer model. Many investiga tors express the belief that molecular dynamics calculations will soon produce useful predictions of structure, dynamics and ther modynamics of proteins, nucleic acids, and complexes of these macromolecules with one another and with other molecules.

The simulation requires two pieces of information at the out set: a starting conformation and potential energy function or forcefield. For a protein, current technology requires that the starting conformation be firmly based on experimental observation: because many conformations exist at local minimum energy, a conformation that is very different from the correct most stable conformation evolves too slowly to reach the correct conformation in the length of a typical calculation.

The forcefield is a very simple empirical approximation to the underlying physics, which properly should be expressed in terms of quantum mechanics but is totally unmanageable in that form. Parameters of the forcefields now in use have been proposed on the basis of a variety of experimental data and to some extent on theoretical considerations. Overall, the several forcefields proposed by different groups for computation of the internal energy of proteins tended to have very similar sets of parameters. Recently developed forcefields for water-water and water-protein interactions permit the simulation of dynamics of proteins in solution, which is a prerequisite for modeling events at the protein surface, including most interactions of proteins with other molecules. (The problems of developing adequate forcefields are discussed in the following section on “Solvation.”)

Carrying out molecular dynamics simulations of proteins is very much an art of the feasible, the limiting factor always being the available computing power. One is always facing the conse quence of an inescapable physical fact: that the most rapidly fluc tuating atomic motions, bond stretching, and bond angle bending vibrations have periodicities of the order of once in every few fem toseconds (10^{-15} sec). Current simulation methodology requires that periodic motions be sampled several times per period, and each sampling requires an evaluation of the system's potential energy, requiring computer time in milliseconds on the fastest machines, Cray and Cyber 205. Clearly, simulations cannot now span a time that is on the biological time scale of microseconds to seconds. Unavoidably, molecular dynamics simulations use sim ple forcefields to span a longer time. Given more computer time, those working in the field will improve simulations in various ways: use of more detailed forcefields; longer simulations; simulation of larger systems posing new physical and biological questions; and application of new, more time-consuming, dynamics methods to ask different questions about the system. To those working in the field, the future is bright; ideas and interesting problems abound, and new computer technology continues to widen the limits of feasibility.

RESULTS

Collected papers for symposia held in 1984 and 1985 give an overview of methods and results of applications of molecular dynamics to proteins.[1]Subsequent work achieved many of the possibilities proposed in these papers, but did not deviate radically from the directions anticipated at the symposia. The following section summarizes achievements and describes possible future applications and advances. This section is divided into three parts that cover structure, dynamics and kinetics, and thermodynamics of macromolecules. The section concludes with a prognosis of developments to come.

The Determination of Macromolecular Structure

The first molecular dynamics calculations of protein molecules produced trajectories whose mean atomic positions deviated very significantly from the starting positions (by root-mean-square (rms) displacements of several Angstroms), which were known within a much smaller error from x-ray crystallography. With the development of better forcefields and the inclusion of a solvent en vironment or even a complete crystalline matrix that consisted of solvent and other protein molecules, the root-mean-square differ ence of the atomic positions decreased significantly. Nevertheless, x-ray crystallographic structures, especially after crystallographic refinement, have a precision well inside this difference. The sit uation appears to be reversed with regard to the thermal distri bution of the atomic positions about their means. Except in rare instances, x-ray crystallography produces a single parameter for each atom that represents the width of an isotropic Gaussian distribution of the atomic center. In contrast, the results of molecular dynamics simulation can be used to describe in detail anisotropic distributions of any shape, even distributions with several max ima. In the one case where results of molecular dynamics simulation have been compared with anisotropic thermal parameters from x-ray crystallography, the agreement was very good. Meth ods for introducing theoretical estimates of thermal restraints into crystallographic structure refinement are being developed.

Molecular dynamics simulations show considerable promise of being able to increase our knowledge of structures proposed on the basis of incomplete information, particularly information derived

[1] Hermans, 1985; Beveridge and Jorgensen, 1986; results described in these symposium papers are not explicitly referenced in this section. Some interesting work has been reported on nucleic acids. However, the technical difficulties of working with tese highly charged molecules much exceed the difficulties encountered in simulations of proteins; given a limited amount of resources, it is understandable that technically less formidable problems have received priority.

from two-dimensional NMR. Two-dimensional NMR produces a set of distances between hydrogen atoms, the Nuclear Overhauser Effect (NOE) distances; for any given (small) protein, many but not all of these distances can be assigned to individual atom pairs. Regular structures such as helices and beta sheets are easily identified and assigned to particular segments of the molecule. However, it is seldom possible to obtain a sufficiently large number of the longer distances that define the relative packing of the regular parts and the structure of irregular chain segments. In this situation, additional information must be brought to bear, an obvious choice for this information being the constraints imposed on the structure by its chemistry and by the requirement of adequate interchain "packing." As these requirements are those observed in a typical molecular dynamics simulation, this has led to the development of a method of molecular dynamics with added constraints, i.e., the NOE distances. By varying the importance of the constraints that determine the conformation and varying the tem perature (the total kinetic energy), the structure can be made to evolve to one with a lower conformational energy. This structure also meets the requirements imposed by the NOE measurements as well as or better than the starting model and may show new distances between hydrogens that are sufficiently short to be detected in the NOE measurement, but whose assignment had been ambiguous (see also the section on "Tertiary structures of macro molecules using NMR" in this report).

Spectroscopy/Kinetics and Molecular Dynamics

Molecular dynamics is a unique tool for simulating time dependent processes in condensed systems. The problem of eval uating the time dependence of motion of a protein is formidable. Because of a lack of symmetry, each atom introduces molecular motion at three new frequencies (normal modes), each of which may be distributed over all atoms. This both overstates and un derstates the situation: it is an overstatement because the fre quencies of many normal modes (e.g. bondstretching modes) are predictable and correspond to localized vibrations; it is an un derstatement both because each conformation of minimum energy (that contributes to the thermal ensemble) contributes its own set of normal modes, and because transitions between conformations, across energy barriers, produce additional motion. This additional motion is usually not a periodic oscillatory motion, but one gov erned by the statistics of barrier crossing.

The high-frequency oscillations of a molecular dynamics tra jectory are easily determined, but are also the least interesting type of motion. Slower motions are far more likely to be relevant to biological function. These typically involve many atoms and have large amplitudes, which are also expected of molecular motions required for the macromolecule's biological function, although not every low-frequency mode will be significant in this respect.

Important motions that have been studied by simulation in clude the hinge bending of lysozyme and the internal breathing motions necessary to transport oxygen to reach the active site (heme group) of myoglobin and hemoglobin. The hinge-bending motion is typical of that presumably required for many enzymes to accept a substrate in the active site and release the products of catalysis. Because of the low frequency of these motions, the hinge bending was not simulated by a direct molecular dynamics calculation. Instead, its frequency was estimated by combining an analysis of the potential energy required to bend the hinge region with hydrodynamic considerations. In contrast, the breathing motion of myoglobin was observed in a molecular dynamics sim ulation of 100 picoseconds (10^{-10}sec) to have a period of roughly 30 picoseconds. It would have been missed had its frequency been only twice as large.

The motion of carbon monoxide in hemoglobin following the photodissociation of carbon monoxide hemoglobin has been sim ulated with molecular dynamics (Henry et al., 1985) to compare the results with extensive spectroscopic experimental studies of the events that follow this reaction. The agreement is very good; a striking result of the simulation was the considerable local in crease of atomic thermal motion that follows the absorption of the photon and breakage of the heme-CO bond. This increase in ther mal motion has a very significant effect on the early kinetics, i.e. during the time required for the excess kinetic energy to dissipate into the protein and then into the solvent.

Case and McCammon (1986) have analyzed the dynamics of ligands in the interior of the myoglobin molecule, with emphasis on the details of the passage of the ligand molecule into and out of the heme pocket. This movement of oxygen is an example of a process requiring passage of the system over a (free) energy barrier. The breathing motions of myoglobin appear to open passages or gates that, when open, allow probes (and by implication, oxygen) to move back and forth between cavities inside the protein (Tilton et al., in press).

Very few thorough investigations have been conducted of such gated events. The best is a study of the manner in which tyrosine rings inside proteins rotate by 180°, a process for which exper imental information is available from NMR spectroscopy. This rotation is an essentially stochastic process, as opposed to the regularly occurring oscillatory motions, and can occur only when the protein assumes particular favorable local conformations, in cidentally, in the course of its internal vibrations. This is often described as a “gated” event. The kinetic process is best studied by placing the protein in a gate-open conformation and determining the relaxation, during which the otherwise rare event (i.e. ring flip) may take place with a good probability (Gosh and McCam mon, 1987). Because the trajectories are reversible, the required kinetic information can be extracted. A difficulty of these studies is that the scientist chooses what parts form the gate and how it opens. As this work has been refined with careful attention to detail and use of improved potential functions, the model's kinetic parameters have tended to approach the experimentally observed values. However, because the motion is so localized, this study of tyrosine ring flips may be the only well-developed example of its kind. We seem far from being able to analyze the kinetic path ways with molecular dynamics, let alone predict rate constants, for the biologically important conformation changes of allosteric proteins, in which many residues readjust their conformation and parts of the protein may undergo relative shifts in position of many Angstroms.

Many interesting conformational relaxation processes of proteins are too slow to be directly accessible using current techniques of molecular dynamics simulation. However, there is a range of fundamental interest (10^{-12} to 10^{-9} sec) that can be studied by both molecular dynamics and NMR spin relaxation spectroscopy. We believe there is substantial opportunity for productive comparison of NMR and the results of molecular dynamics simulation. Perturbed NMR resonances for spin one-half nuclei relax primar ily because of modulation of dipolar interactions by global and internal molecular motion. Resonances in NMR spectra can also be assigned to discrete sites and, in cases where the geometry of the dipolar interaction is well defined, time scales for motion at a particular site can be extracted. Although relaxation mechanisms can be very complex, especially for protons, useful analysis should be possible in some cases. Use of^{13}C NMR can, for example, simplify relaxation time analysis, because most relaxation interactions occur with directly bonded protons (Wagner and Bruhwiller, 1986). It is also becoming increasingly possible to introduce amino acids enriched in^{13}C at specific sites. NMR of^{13}C-enriched proteins and peptides offers substantial possibilities for extraction of experimental time scales of motion in the 10^{-12} - 10^{-9} range for verification of theoretical predictions.

When motion of groups or relaxation interactions are complex, molecular dynamics simulations may also improve the interpretation of NMR data. Here, Levy et al. (1981) have shown that it is possible to construct appropriate dipolar correlation functions from states sampled in a molecular dynamics simulation. In prin ciple, this allows calculation of NMR relaxation parameters that can be used to validate models used for interpreting NMR data. Thus, the improvement of molecular dynamics simulations and the development of experimental methods for determining structure may prove symbiotic. Carrying out this dual strategy will require substantial investment in producing an accurate description of spin relaxation, as well as coordination among those developing simulation programs.

Thermodynamics of Macromolecules

Physicists have known for a considerable time about techniques to calculate equilibrium thermodynamic properties from molecular dynamics calculations. The techniques have been ap plied to proteins very recently, but already their use has shifted the emphasis of the simulation field to calculations of free energy dif ferences. Several factors explain the great current interest in this application, the most important being the availability of many precise experimental data for a variety of equilibria that involve biological macromolecules and the unexpectedly excellent theoretical estimates that the simulations produce. The first successes were obtained in studies of the hydration of small molecules and ions, in which the free energy of transfer of a small molecule to bulk water could be compared with accurate experimental data (see following section on “Solvation”).

An important feature of the ongoing research program on macromolecular equilibria is that investigators are carefully iso lating a small subset of the global problem to avoid overwhelming available computers. For example, in studying enzyme inhibition equilibria, McCammon's group is using molecular dynamics simu lations to estimate the differences in binding free energy of a series of small inhibitors to the enzyme trypsin. A complete calculation consists of two parts, one in which one substrate bound to the protein is replaced with another, and one in which the first substrate solvated in water is replaced by the other. As can be seen from the following thermodynamic cycle (E is enzyme, S1 and S2 are two different substrates),

$$\begin{tikzcd}
\mathbf{E\ (solvated)\ +\ S1\ (solvated)} \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & \mathbf{E-S1} \arrow[d, leftrightarrow] \\
\mathbf{E\ (solvated)\ +\ S2\ (solvated)} \arrow[r, leftrightarrow] & \mathbf{E-S2}
\end{tikzcd}$$

the difference of the two free energy changes obtained from the simulations (indicated by vertical arrows) is equal to the difference of free energy of binding the two substrates to the enzyme (indi cated by horizontal arrows). The agreement between theory and experiment is of the order of a few kJ/mole, which amounts to a factor of two to three in the equilibrium constant. These methods are easily adapted to the estimation of differences in affinity of substrates and inhibitors caused by alteration of the enzyme by site-directed mutagenesis (e.g., Bash et al., 1987b). In one study of the interaction of a protein and a small molecule, the binding of xenon gas to myoglobin, Hermans and Shankar (1987) found that the molecular dynamics simulation was able to give a direct estimate of the binding equilibrium constant, which was within a factor of two of that observed experimentally.

Similar methods are being used to study conformational equi libria of macromolecules. Most thoroughly studied are conformational equilibria of the alanine dipeptide, a well-known low molecular weight model of a polypeptide. The most important result concerns the equilibrium between two conformations, al pha and beta, which correspond to different helical structures of polypeptides. In an environment of water molecules, simulations performed by two groups with different methods gave similar results: a preference by a factor of two to five for the (extended) beta conformation. Experiment indicates, imprecisely, a value of around 10.

Prognosis of Developments

The success of the free energy simulations has suddenly changed the scope and emphasis of molecular dynamics simulations. The early simulations either clarified properties of proteins that were difficult to study experimentally (kinetics and dynamics on the picosecond time scale) or else gave unsatisfactory agreement with experiment (mean atomic positions). However, free energy calculations suffer from neither of these problems. In addition, the results are in a field that is traditionally of great interest to biochemists. Biochemists routinely and accurately achieve exper imental determination of free energies of binding (from binding equilibrium constants) of inhibitors and substrates to enzymes. Furthermore, the agreement between theory and experiment is so good that molecular dynamics simulations are widely believed to be a useful tool to predict the inhibitory power of new compounds. This tool will at least screen out a large fraction of possible in hibitors, and thereby greatly reduce the synthetic work required in the search for the perfect inhibitor. Replace “inhibitor” with “drug,” and one realizes the potential of these tools. Add to this the possibility of predicting the properties of genetically altered proteins produced by the biotechnology industry and the demand for such tools soars.

This new application has created sudden and perhaps unex pected demands for computer time for two reasons. First, investi gators suddenly have a seemingly limitless number of technologically and biochemically interesting questions to answer; one may envision the possibility of rationalizing the inhibition constants of all studied inhibitors of any one enzyme and its mutants (the latter designed and manufactured in the laboratory on order). Second, as the emphasis has shifted from problems of structure and dynamics to problems of equilibrium thermodynamics, there is less reason to analyze the details of most trajectories and conformations. This is because free energy simulations typically pass through a series of artificially constructed intermediates that are physically un realizable. Thus, each researcher will be able to perform more simulations without being overwhelmed by the time requirements of analyzing the results. Consequently, one researcher can more effectively use more computer time.

Accordingly, progress in free energy simulations, although po tentially very rapid, is heavily limited by available computer time. As recently as 5 years ago the demands of these calculations ex ceeded the available computer power. At present, each of several research groups is using hundreds of machine hours of Cray time. In addition, a number of groups have been able to acquire Star array processors, which may have the power of a Cray but a much lower price. Dedicating one or more array processors full time to the single task of molecular dynamics is extremely efficient in terms of total cost of hardware and programming. Similarly, the economics of building a hard-wired special-purpose machine for molecular dynamics may be justified in terms of the economics of building and operating several copies of the final product, and such a machine is almost complete. In spite of these rapid developments, the scope of free-energy simulations is still severely limited.

Within a short time, we will need a radical increase in computer time to realize possibilities that are now clearly defined. An immediate 10-fold increase appears needed, and does not seem an extravagant objective with proper planning (i.e., duplicate existing hardware that is already programmed and otherwise inexpensive). The pharmaceutical and biotechnology industries will make some investment since companies' efforts at rational drug design require simulation capability that works in parallel with NMR and x-ray determination of physiologically crucial enzymes.

Apart from drug and protein design, others within and out side of industry will apply these techniques to the broad problems of protein-protein and protein-nucleic acid recognition, by using a combination of molecular dynamics simulations and the results of site-directed mutagenesis. Once we have dealt with the problem of rationalizing these equilibria in terms of molecular interactions in atomic detail, our attention will shift to the application of the newly acquired skills to problems of the dynamics of interaction of proteins with other molecules, which will presumably require just as much computer time. In contrast to the protein folding prob lem described in the previous section of this report, the problem of computer modeling of the dynamics of protein interactions can be tackled in a series of small, increasingly complex steps, each of which solves a discrete problem of immediate biochemical inter est, yet also provides additional insight and experience needed to advance the technical expertise.

Beyond the need for sufficient computer time, we must improve forcefields for proteins, carbohydrates, and nucleic acids by determining better values for the parameters and extending these to include drug-type molecules; attempts should also be made to adapt molecular dynamic techniques in ways designed to over come some of the intrinsic imperfections of existing forcefields. At present, there is a disturbing trend towards the development and commercialization of proprietary software instead of free exchange of programs and subroutines, and there appears to be a parallel development of a proprietary forcefield. The members of this committee hope that these trends are temporary; the emphasis on commercialization is appearing too early in the scientific process. If the trend persists, a national agency should commission the development of state-of-the-art programs and forcefields that would be available to all workers.

SOLVATION AND ELECTROSTATICS IN COMPUTER SIMULATION OF BIOPOLYMERS

Biomolecular systems function in vivo in an aqueous solution environment. That environment includes solvent as well as a substantial component that consists of a variety of salt ions. Because these components interact significantly with each other and typically also with the macromolecular species, this environment can contribute substantially to the observed state of macromolecules in solution. Manifestations of its influence include the relative stability of various macromolecular conformations and the binding constants that characterize the association of macromolecules with each other or with other biochemically significant molecules.

The simplest effect of water can be thought of as that of a high dielectric medium that screens the interaction of charged and polar groups. Hence, the interaction is in effect much smaller than the corresponding interaction in the absence of solvent. However, solvent influence cannot be described solely in such terms. For example, it has long been appreciated that nonpolar moieties are preferentially excluded from water, and such “hydrophobic” effects have long been believed to be a major component in protein conformational free energies (Kauzmann, 1959). For all po lar interactions, hydrogen bonding in particular, the*net* energy is determined primarily by the mismatch between short-ranged solvent-solute and solute-solute interactions. In the case of the highly charged nucleic acid polymers, the identity and distribution of solution counterions also seems to be particularly significant.

As this brief discussion makes clear, we cannot expect to succeed in the quantitative treatment of biopolymer structure and function without paying due attention to the molecular role of the solution environment. In particular, we must take into account the significant role played by the environment in determining the strength of ligand binding, as well as the relative stability (free energy) of the*varied* structures that must be encountered during protein folding and that may accompany function.

In the following section, we describe briefly alternative frame works for considering environmental effects and discuss the current state of theory in this area. We focus our attention on the limi tations of currently available results and methods, as well as on the potential for significant progress in the near future. Finally, we point out important areas for attention in the short term, and comment on the prospects for successful quantitative treatments in the longer term.

All-Atom Modeling

Detailed molecular models for the solution environment take the same form as those used for the biopolymer as such. That is, the solution components are represented as a collection of sites, typically atomic sites, that carry partial electrostatic charges and are each associated with a spherical short-ranged potential that accounts for short distance interatomic repulsion and for London attractive forces. Molecular entities are usually specified through sets of bond length and angle constraints. For water, those models that most successfully reproduce experimental liquid data (Jor gensen et al., 1983) include an additional charged site that is not aligned with any of the three nuclei of each molecule. The sol vent molecules (and ions) then interact with each other and with macromolecular components by a superposition of electrostatic and short-ranged terms.

This format for the potential is itself an approximation, and the quantitative limitations of this form are not yet completely determined. The most significant limitation is that, in reality, the polarity of a molecule in solution is substantially influenced by its surroundings due to electronic polarization. For hydrogen-bonded liquids such as water, it is known that a successful nonpolarizable model must include electrostatic site charges for each molecule that correspond to an increased dipole moment with respect to the gas phase; the increase represents the average additional polarization induced by neighboring molecules (Stillinger, 1975). Such effects appear, in principle, for all components of the solution, including macromolecular species. Although we have no evidence that the neglect of explicit polarizability is now limiting the predictive power of such models, it should be kept in mind as a potential limitation to quantitative prediction. Whether such effects are included in modeling efforts is limited primarily by computational rather than theoretical capabilities, so even in the worst case, such effects can eventually be added later.

Carrying out any computational study of macromolecular be havior taking full atomic account of the solvent is an extraordi narily demanding task, since the surrounding solvent constitutes most of the whole system. In the presence of finite concentrations of ions, the problem is substantially more difficult, since in that case, the solvent associated with ionic dilution must be included as well (25 ion pairs require 15,000 water molecules to dilute to 0.1M). Such a computer simulation study remains at least an order of magnitude beyond what is now feasible. Nevertheless, consider able progress is being made in simulating macromolecular systems and related model compounds, taking full account of the solvent environment. This progress is due in large part to the vast increase in available computational power.

In particular, in the area of model systems, this power has per mitted a few studies of small molecule conformational equilibrium (Jorgensen, 1982; Rosenberg et al., 1982; Zichi and Rossky, 1986) and one direct investigation of the conformational free energy of a dipeptide (Mezei et al., 1985). Carrying out such studies requires specialized sampling techniques, termed umbrella sampling, that allow the accurate determination of relative populations of conformers that are separated from one another by significant free energy barriers. Such techniques and their efficient use are the products of recent research on simulations of molecular model systems.

The results of these studies are consistent with the limited experimental data available, as documented in the cited reports. A few other studies have been carried out on peptides in water without any attempt to fully explore conformational space (Brady and Karplus, 1985; Hagler et al., 1980a).

Recently, a number of studies have been carried out that aimed to evaluate directly the relative hydration free energies of small molecules, amino acids, and nucleotide bases (Bash et al., 1987a; Jorgensen and Ravimohan, 1985; Lybrand et al., 1986; Singh et al., 1987). Such studies are essential for calibrating the potentials in use. These relative free energy quantities are amenable to calculation using a thermodynamic perturbation approach, another tool added recently to simulators' methods (Postma et al., 1982).

The results of the studies are encouraging in that the investi gators obtained relative free energies within about 1 kcal of exper imental determinations. Although this level of accuracy may not be sufficient to determine quantitatively the stability of systems involving many such groups, it strongly suggests that the potential functions are sufficiently close to being right that relatively small adjustments may adequately finish the job.

Similarly encouraging results have been obtained in small model system binding equilibria. Studies of relative affinities of ions for a cyclic ionophore (Lybrand et al., 1986) and of base pair stacking and hydrogen bonding (Bash et al., 1987) have been carried out. In the latter case, the results do not agree quantita tively with experimental estimates but are, again, close enough to warrant optimism.

In parallel, several groups have carried out true macromolec ular binding studies. These will be discussed in detail in the following section on molecular dynamics of biopolymers.

In the area of globular protein structure*per se*, several at tempts have now been made to compare the structural and dynamic behavior of the hydrated model to experimental hydrated crystal structures and to the results obtained in the absence of solvent (Krüger et al., 1985; Teleman, 1986; van Gunsteren and Berendsen, 1984; van Gunsteren and Karplus, 1982; van Gun steren et al., 1983; Wong and McCammon, in press). The results of these studies are encouraging in that they show that the ad dition of solvent makes simulated structures agree better with experimental crystallographic atomic positions. Significant quan titative differences remain, however. Further, for hydrated single proteins, a simulation initiated in the crystallographic protein structure was found to produce an increasingly deviant structure (as measured by the R factor) from the crystallographic structure as the simulation proceeds (Krüger et al., 1985; Teleman, 1986). In a recent hydrated crystal study, similar behavior appeared to be present (Berendsen et al., 1986). One possible interpretation is that these studies simply sample fluctuations that are not fully av eraged, and the deviation is only an apparent one—the simulations are relatively short, less than 100 picoseconds. Alternatively, for the noncrystalline simulations, this result may reflect real differ ences between crystal and solution structures. However, one also should suspect the underlying theoretical interaction potentials as the source of the deviation, and much more testing is required to narrow the alternatives.

In this context, it is important to emphasize that the state of the art has not yet reached a level where computational complex ity is the only limiting issue, as a simple example can illustrate. Although short-range solute-solvent forces play a very important role, we have already noted that the dielectric screening of solute charges is of substantial importance. In light of this, it is no table that for those popular water models for which the dielectric constant has been determined, the agreement with experiment is not very good; at room temperature the so-called MCY model yields a value near 35 (Neumann, 1985), while for the structurally and thermodynamically excellent TIP4P model of Jorgensen (Jor gensen et al., 1983) one finds about 50 (Neumann, 1986), and the ST2 model yields about 120 (Steinhanser, 1983), all compared to the experimental result of about 80. Clearly, for the interaction of charges at long range, such discrepancies would be quantitatively disastrous. This does not imply that results obtained, for example, for polypeptide conformational equilibria would have comparable relative errors, but it does indicate that caution is warranted, and that further model development is necessary.

Implicit Environmental Modeling

Since the solvent and small ions as such are often not of primary interest, it is in principle simplest to avoid giving an explicit account of the surrounding solution and treat its influence only implicitly. Formally, this can be done by introducing effective, or so-called solvent-averaged, potentials among the solute atoms of explicit interest. The rigorous existence and formulation for such a reduction is well known. However, such effective potentials are not generally represented by only pairwise interactions, but can be resolved into pair, three-body, and other terms. The pair term, for example, is the potential of mean force between an isolated solute pair in an infinite amount of solvent. The lack of pairwise additivity is present even if the full unreduced system is described by pairwise additive potentials. The use of a continuum dielectric model of an ionic solution represents the simplest form of a pairwise additive effective potential, where, in addition, the pair potential is only roughly modeled.

The question is whether pairwise additivity of the effective potentials is a good approximation. The validity of this approx imation remains largely untested, although for ionic solutions pairwise additive semiempirical potentials adequately reproduce experimental solution thermodynamics up to about 1M concen tration (see Friedman et al., 1973 and references therein).

Current macromolecular modeling of the effects of solution environments typically invokes such effective potentials in a relatively crude form. Most often, an effective dielectric constant typical of a nonpolar, polarizable material is used to account for polarization screening of electrostatic charges (Weiner et al., 1984). In some treatments, a modification to the potential to account for short-range, molecular, solvent effects is then added in some treat ments (Gibson and Scheraga, 1967; Némethy et al., 1978; Hodes et al., 1979a, 1979b; Kang et al., 1987). This added “hydration shell” term introduces a free energy bonus or penalty associated with the close approach of solute atoms, typically proportional to the overlap volume of the first solvation shells of the approaching polypeptide atoms. This approach is closely analogous to the method widely used in models of ionic solutions pioneered by Friedman et al. (1973).

A significant problem of the implicit approaches to solvents now in use is that they use an ad hoc form for effective potentials, the reliability of which is not well established. The ability of such potential functions to produce correct quantitative results for protein/nucleic acid systems is obviously difficult to assess since the system is complex, with many theoretical parameters and relatively little experimental data for comparison. However, Gō and Scheraga (1984) have demonstrated that such approaches are useful in analyzing differential hydration effects in specific cases. The current procedure is to establish the parameters for the potentials from the thermodynamics of hydration (Némethy et al., 1978; Hodes et al., 1979a, 1979b; Kang et al., 1987); this procedure is valid at the present stage. However, in the near future, we should emphasize more direct comparison to spectroscopic and NMR results for small molecules such as oligopeptides.

In fact, some skepticism of the current forms of the potentials is warranted, given the results obtained for the most simple so lute systems. It is known, for example, that for ionic solutions, a detailed molecular solvent treatment of the interionic potential of mean force does not closely resemble the hydration shell model, although both are consistent with observed thermodynamics (Pet titt and Rossky, 1986). The true effective potential exhibits large oscillations as a function of distance. The minima are shifted in spatial position compared to the simpler model, but the depth of the alternative potentials appears to compare favorably. Hence, the hydration shell model may be a viable way to estimate solvent effects associated with native versus completely unfolded states, but not for intermediate structures. This last hypothesis is consis tent with the use of aqueous thermodynamic data to parameterize the potential. It is clear that the folded state prediction *per se* is of great import in the *a priori* prediction of protein structure and function.

Recent efforts to generalize the molecular solvent theory avail able for ionic solutions to the atoms that make up peptides appear promising (Pettitt and Karplus, 1985; Pettitt et al., 1986), but no quantitative comparison to experimental data has yet been made.

Even if we accept a purely continuum fluid description of the solvent environment, the issue of dielectric screening itself is a major one. If one is studying a protein crystal, a small dielectric constant may well be appropriate. However, the use of such a value to determine structures does not seem warranted. Particularly for atomic charges near the solvent interface in a folded protein, one expects that the pure solvent value is more relevant. Recent work is aimed directly at rigorously examining the usefulness of different distance-dependent dielectric functions for such interactions within a dielectric continuum picture, but for dielectrically inhomogeneous systems (Gilson et al., 1985; Klapper et al., 1986). This work may well produce an optimal and well-founded treatment within the scope of this simplified model.

The approach for nucleic acid modeling is less refined than for proteins because of the ubiquitous charges and requisite coun terions present in solutions of nucleic acids. For these macromolecules, it is insufficient to deal with solvent alone. Current ap proaches have considered various polyionic charges, solvent, and, in some cases, counterions. In many studies, the polyionic phos phate charges have been artificially reduced to about 25 percent of the physical value to account crudely for counterion association (Hingerty and Broyde, 1982; Tidor et al., 1983). This value arises from the counterion condensation formalism, which describes a required fractional counterion screening for counterions that are far from the polyelectrolye (Manning, 1978). This approach is a valuable qualitative tool, but we do not expect quantitative results from such a strong approximation.

Only recently have initial all-atom studies of polynucleotide ion-solvent systems been carried out (Corongiu and Clementi, 1981; Seibel et al., 1985; van Gunsteren et al., 1986), but it is clear that the exceptionally time-consuming nature of these simulations with ions present does not yet permit such calculations to be very informative in practical ways. To simulate a duplex oligonucleotide without added salt for 2 nanoseconds (a relevant motional time scale for the polymer) would require roughly 1,000 hours of supercomputer time.

In the case of nucleic acids, an intermediate ground state exists that is not relevant for many proteins. That is, the set of explicitly simulated atoms can be extended to include the macromolecule and ambient ions, while retaining the implicit treatment of only the solvent. In terms of the number of atoms to be followed, the simplification is substantial.

In any of the cases discussed above in which the solvent is treated implicitly, one still must implement realistic potentials of mean force, or, at least, invoke a firmly based dielectric continuum treatment. Since the potential payoff of knowing viable implicit solvation routes is very large, it is important to encourage research into implicit modeling in biopolymer related systems.

A potentially useful approach to the ionic atmosphere that avoids even the intermediate-level treatment of ions is the implementation of solvent and ion-averaged potentials within the biopolymer. The use of reduced phosphate charges is an ad hoc form of such a potential function. An oversimplified but well-founded alternative is the use of a Debye-Hückel-like screening between polymer sites (Hesselink et al., 1973; Manning, 1978;

Soumpasis, 1984), employing the bulk solvent dielectric constant. The latter approach cannot, however, account for the unusually high degree of ionic association that is present in the immediate vicinity of a polyion.

Another unusually promising approach involves applying more analytical theories for the influence of the solution environment, while retaining a detailed description of the biopolymer. In essence, one evaluates the effective potentials that govern the intrapolymer interactions for fixed polymer configuration by (numerically) solving the relevant equations of an essentially analytical theory. An example of significant recent progress along these lines is the work of Pack and coworkers (Klein and Pack, 1983). They have used an algorithm for solution of the Poisson-Boltzmann equation for the ionic distributions around a detailed DNA model, and from such distributions the relevant free energies of different conformations are, in principle, obtainable. At least for simplified models of DNA, the Poisson-Boltzmann mean field theory has proved accurate compared to computer simulation for the same mod els (Murthy et al., 1985). Closely related approaches have been considered for enzyme-substrate binding involving charged species (Klapper et al., 1986). However, a Poisson-Boltzmann treatment is tied to a dielectric continuum view of the solvent, although dielectric heterogeneity can be readily accounted for within this context.

A brief comment on biopolymer dynamics is appropriate here. Dynamics are clearly connected to the general question of protein folding, and are likely to be significant for function in many cases. Although molecular dynamics may not be directly related to the issue of predicting function, it is clearly connected to the more general question of protein folding. As for the equilibrium time-independent problem, one can, in principle, consider the full atomic description. However, one can also focus only on an explicit subset of solute atoms, such as biopolymer or biopolymer plus ions. The formal theory to be applied when only some of the atoms are considered explicitly is well established (for recent discussions in a variety of contexts see: Adelman, 1982; Ermak and McCammon, 1978; Tully, 1981). The motion proceeds according to the forces prescribed by the effective potential, but with additional random forces and friction due to the implicit solvent collisions. In general, these solvent forces are not simple, but depend on the history of the solute dynamics (memory) and the current solute conformation (hydrodynamic interaction). In the general case, the relevant equation is the so-called generalized Langevin equation with the friction described by a memory function that embodies the sta tistical properties of the solvent collisional correlations in space and time. The approximation that neglects any frictional correlation involves only constant drag coefficients, the so-called ordinary Langevin equation. Further approximation leads to equations of a diffusion type.

As with most areas of theory in physical science, the time-dependent theory lags behind the equilibrium theory in terms of development and application. A few examples of attempts to ap ply these methods to realistic and simplified models exist (Levy et al., 1979; McCammon et al., 1980) but only once does it ap pear that a polypeptide has been examined (Brooks and Karplus, 1986). The problems encountered in any application involve, first, computational limitations, since the dynamics that are of biochemical interest are relatively slow. Perhaps more significant is our extremely limited knowledge of the memory function and hy drodynamic interactions for a complex solute. In principle, we can test our assumptions against all-atom simulations or the relevant functions extracted from these simulations, but this route is itself limited by the current sparsity of the requisite simulations. Never theless, future developments in this area seem very likely, although they are further away than are those in equilibrium theory.

Conclusions

The ability to adequately test predictions of theoretical calcu lations is an element of overriding importance in the future of the modeling of solvation and electrostatics. This testing can occur at two levels: first and foremost by comparing theory and experiment and second, by comparing results obtained through convenient but approximate theory with those derived from accurate theoretical treatment. The first is essential for accuracy and the second for the future development of viable theoretical methods for studying increasingly complex systems. Therefore, we should continue to encourage both experiment and theory for both macromolecular and smaller model compounds.

We cannot expect an immediate and completely satisfactory way to account for the influence of the solution environment on the behavior of macromolecules. The above discussion indicates that some unsolved and many partially solved problems remain. Nevertheless, reasonable approaches already exist that provide adequate grounds for qualitative study. The degree of quantitative accuracy is not yet well established and awaits both further model calculations and further thermodynamic and spectroscopic experimental data, so that we may make unequivocal comparisons between theory and experiment. Based on the steady progress described above over only the past few years, we have every reason to expect rapid incremental progress to continue. A clear view of the capabilities of current models and methods for describing flexible small molecules should be available within only a few years.

At the same time, the large amount of theoretical activity both in the development of well-founded approximate approaches and in the simulation of atomic-level solvated molecules virtually assures our ability to make the appropriate comparisons between the two in the near future. A very significant element in recent developments has come from algorithmic breakthroughs. Biased sampling techniques and thermodynamic perturbation/integration methods are two new methods that contribute essential capabilities to the theoretical effort. Hence, we should encourage theoretical developments as much as computational applications.

The rapidly increasing access to the necessary computer facilities has contributed significantly to progress, and it is essential that this access continue to grow. For all-atom models of the environment, the current limitations on macromolecular simulation are primarily computational. Although limitations of the model theory are also likely, we now have insufficient data to make that judgement. An order-of-magnitude increase in available computing power would be enough to make a dramatic difference in this area; two orders of magnitude would permit simulations into the interesting many-nanosecond regime. Such changes are likely within the next five years through the combined effects of new hardware, improved performance, and lower cost.

To explore adequately events such as protein folding that occur on much longer time scales (or involve vast conformational exploration), these computational improvements would still be inadequate by many orders of magnitude. Thus, a theoretical breakthrough appears necessary if we are to make real progress within the next several years. Such a breakthrough would be, for example, the demonstration of an implicit treatment for the solution environment that yielded accurate biopolymer dynamics on a nanosecond time scale when compared to a full atomic-level simulation. Such a treatment could then be applied for longer times.

Considering the very limited knowledge available about the performance of alternative implicit modeling techniques, we can not now say whether such an approach is workable, even in prin ciple. However, the process of determining the usefulness of these techniques requires a one to two order-of-magnitude gain in computer power.

In summary, the rapid progress in our ability to describe the environmental aspects of bipolymer systems gives solid ground for optimism that this element of biomolecular modeling will not impede development of useful predictive methods. However, for the most challenging aspects, we are at least several years away from demonstrating the ability to mimic accurately solution environmental effects.

HEURISTIC METHODS

There are two major approaches to the prediction of three-dimensional structures of proteins: modeling by extension and hierarchical searching. Both methods can combine heuristic ideas and energy calculations. They differ from the energy calculations described earlier and from each other in the way they arrive at starting structure. Modeling by extension uses the known structure of a protein or proteins with strong sequence homology to the unknown. The hierarchical methods use packing considerations derived from the crystallography of many proteins. The following section describes these approaches in more detail.

Homology

Protein Homology

Proteins occur in families. Evidence for this comes first from protein sequence homology and then from the architectural simi larity of homologous proteins determined by x-ray crystallography and NMR spectroscopy. A family of proteins can be modeled by homology if several conditions are fulfilled. First and most important, the structure of at least one member of the family must be known. Second, the three-dimensional protein to be modeled must be sufficiently homologous to that of the known protein. Many proteins have been modeled over the past five years, and the general consensus is that if two proteins share at least 30 per cent similarity, then computer graphics and energy modeling will be useful. If the global homology is less than 30 percent, then it is difficult but not impossible to say whether the two proteins are in the same family. If there are important conserved residues such as disulfide bridges then even 30 percent homology might be sufficient.

The difficulty of modeling a given protein depends on the range of homology with the known structure. When the homology is between 80 and 100 percent, normally only the surface amino acids are changing. In these cases, there is usually no change in peptide length. With homology of between 50 and 80 percent, again, mainly the surface amino acids are changing, but there may be additions and deletions in the peptide length. The amino acids on the surface of the protein can be easily substituted. Energy minimization and/or molecular dynamics are sufficient to reduce the errors caused by any changes in surface sidechain conformation. Surface-charged amino acids under molecular dynamics ei ther must be neutralized by artificially altering the parameters or by adding a solvent box about the protein.

When interior amino acids are changed from one protein to another, the changes are either to make a long amino acid shorter, thus creating a hole in the interior of the protein, or a long-short pair of amino acids are changed to a short-long pair of amino acids.

When amino acids are added or deleted in a helix, they are generally in multiples of three amino acids. This preserves the hydrophobicity/hydrophilicity relations in the helix. In contrast, beta strands tend to have insertions or deletions of two amino acids, thus preserving the inside/outside relations (Feldmann et al., 1985). The inside amino acids of a protein normally are hy drophobic, while the outside amino acids are normally hydrophilic. Graphic modeling of such changes is accomplished by moving the additions and deletions in helices and beta strands toward the ends of the secondary structure feature. Graphic modeling of loops is easy to do but fraught with large inaccuracies. Insertion or deletion of amino acids on a loop can be accomplished by breaking the peptide chain, making the appropriate change and then bending the loop ends to accommodate the change. Energy modeling un der local conditions of relatively high simulated temperature can be used to explore the local conformational space.

There are two ways to align sequences, automatically and manually. The automatic alignment methods such as Wilbur and Lipman tend to align the sequence for highest local match. Man ual methods (see Feldmann et al., 1985) permit the alignment of secondary structure features which minimize the number of disturbances which must be made to the protein and convert from the crystallographic structure to the model structure.

One of the ways to overcome the uncertainties of the structure of a particular loop is to build a library of representative loops. Alwyn Jones at Uppsala has done this and recently integrated it into FRODO, his modeling program.

After all the changes have been made either by graphic methods or by using a loop library, extensive molecular dynamics sim ulation is generally used to improve the quality of the model. Whether a broad range of scientists can use molecular dynamics calculations depends on the availability of appropriate modeling software and on a variety of displays and workstations. To complete the modeling by molecular dynamics calculations, sufficient computer power must be available either on a personal supercomputer (PSC) or by network access to a national supercomputer.

Modeling by Extension

Twenty years ago, Phillips (1967) made a model of the protein lactalbumin without obtaining a single crystal. This was possible because the amino acid sequence of lactalbumin had been found to be 35 percent identical with that of the enzyme lysozyme, a protein whose crystal structure had recently been determined. The residues for lactalbumin were simply placed in the equivalent po sitions known to be occupied by the residues of lysozyme (Browne et al., 1969).

Subsequently, Warme et al. (1974) underscored the validity of the approach by computational approaches to the structures of the two proteins. The structure of alpha-lactalbumin computed by energy minimization by Warme et al. (1974) has recently been verified by x-ray crystallography (D.C. Phillips, personal commu nication, 1987). Since then, “modeling by homologous extension” has become a common if sometimes casually applied approach, that has benefited from modern computer graphics (Greer, 1985).

A recent example of the utility of modeling by extension is that an important but elusive factor, angiogenin, involved in the biogenesis of blood vessels was isolated after a search of more than 15 years. The sequence of this factor was determined and found to be 45 percent identical to pancreatic ribonuclease. As a result, Palmer et al. (1986) promptly generated a three-dimensional structure.

Naturally, the closer the resemblance of the unknown protein to the one whose structure has been determined, the more accu rate the modeled structure. Recently, however, Moult and James (1986) have shown that it is possible to construct good models even when the sequence resemblance is barely recognizable. In many instances, then, all that is needed is the family relationship of the new protein.

Exon Shuffling

Many recently evolved proteins exhibit evidence of "exon shuf fling." In this phenomenon, mosaic proteins result from the ge nomic rearrangement of segments that encode small portions of different proteins. For such proteins, it is thought that the peptide segments, which often range from 30 to 90 amino acids in length, all fold independently (Doolittle, 1985); as such, they constitute domains in the truest sense. As of mid-1987, about six such domains had been found in a variety of proteins in different three-dimensional settings. Most of them are tightly folded and contain disulfide bonds that hold the structure in place. They include such well-known motifs as: the "EGF domain," the "Kringle," and the "fibronectin fingers." They are readily identified by rou tine computer searches of sequence, and, when such a structure has been identified, can be immediately modeled in place. Exon shuffling, which is due at least in part to the additional recombination that ensues from the presence of introns between exons, is not restricted to the small set of stable structures listed above, and it is anticipated that hybrid and mosaic proteins of many sorts will be identified. In all these situations, prior knowledge of any structural motif will aid in interpreting the overall structure. Doubtless, other motifs will emerge as more sequences are determined, compared and analyzed.

HIERARCHICAL MODELS OF PROTEIN FOLDING

The major goal of hierarchical modeling is to build three-dimensional structures that incorporate directly or indirectly the architectural principles by which nature constructs globular proteins. The direct approach uses empirical rules or models that capture recognized aspects of protein folding. The indirect ap proach uses homology to provide the basic structure and explores the local environment with energy calculations or other modeling efforts. This latter work was described in the previous section. Here, we assess the success of rule-based procedures.

Investigators have used these procedures at various levels of formality. These efforts generally follow the same plan: predicting secondary structure followed by packing secondary features. They also use the Kauzmann hydrophobic model as the basic packing principle. Classifying protein domains by structure has been particularly important because it provides major rules (for a discussion and extensive references see Richardson, 1981; Cohen et al., 1983). Such rules include statements about the geometry of helices packing against helices and beta sheets and beta sheet-beta sheet packing. The efforts have also led to the development of a list of rules concerning the ordering of strands within a beta sheet. We should note clearly that rules such as these only sum marize what is observed in known structures; they are not derived from first principles of physics or chemistry. Nevertheless, many of the idealized structures built to be consistent with such rules do look recognizably like protein domains and some are rather close (average error 4 Å) to the crystallographic result.

PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE

We differentiate the techniques of pattern recognition from those of artificial intelligence, especially its subdiscipline of ex pert systems. Pattern recognition is usually defined as including numerical techniques for clustering observed data into binary or higher order categories (but see below). Artificial intelligence is usually defined to include use of rule-based systems to express and use empirical knowledge of a subject.

The usual techniques of pattern recognition, such as linear discriminate analysis, cluster analysis, and other parametric and nonparametric methods, have not proved useful in the analysis of protein structures and functions. These methods pose difficulties in determining the statistical significance of a derived classification. Given the relative lack of knowledge about structure-function relationships in complex molecules such as proteins, it is very difficult even to pick a reasonable set of structural descriptors upon which to build a clustering scheme.

Other definitions of pattern recognition, however, are less controversial in technique, if not in interpretation of results. These techniques involve, for example, the presentation of three-dimensional protein structures in the form of C-alpha distance maps (Kuntz, 1975; Rao and Rossmann, 1973) to infer the presence of secondary and super secondary structures and location of intron/exon boundaries (Gō, 1981).

The technology of expert systems may be applied when empirical knowledge, which can be expressed in rule-based systems, can be used to solve problems. For example, production rules of the form IF (x) THEN (y) may be an integral part of a hierarchical pattern comparison described previously (Figure 4-1). To achieve high performance, such rule-based approaches are often supplemented with methods and data from other sources. Two systems under development, KARMA (Klein et al., 1986) and PROTEAN (Hayes-Roth et al., 1986) illustrate this point. The KARMA system employs rule-based proposal and evaluation of small molecules and their predicted binding activities in receptor sites of known proteins. A variety of mathematical and graphic procedures are used to evaluate candidate structures and their affinities for binding. PROTEAN uses artificial intelligence techniques with interatomic (nonbonded) distance information from NMR and a variety of mathematical and graphic techniques to explore structural possibilities for proteins of known primary structure.

Cohen et al. (1983, 1986b) have carried out one of the most extensive projects in their studies of alpha/beta domains and four helix bundles. In the former case, they identified secondary features by using pattern matching and then built tertiary structures from exhaustive combinatorial packing of the secondary elements. In the most favorable case, flavodoxin, they could generate a unique prediction for the alpha carbons involved in helix or sheet of the protein. Similar predictions for molecules such as interleukin-2 have been made based on a core structure of four helices (Cohen, et al., 1986b).

The important strengths of such projects are (1) they achieve low resolution structures of the central residues of proteins that contain many protein features, including a prediction of the "ac tive" portion of the molecules; (2) the computational labor is modest; and (3) the rule system can be tested directly against known structures and their homologs. In some sense, they are the best solution currently available to the folding problem. On the negative side, the low resolution is an obvious limitation. Details of loops are often neglected. Only certain classes of proteins can be dealt with successfully.

The next several years should bring improvements in all as pects. More realistic models will reduce errors. Loops and side chains can be treated either from rule-based or energy-based ap proaches. Expansion to more protein structural classes is proceeding rapidly. It is difficult to see the ultimate limitations of these heuristic methods. We are hopeful that they yield good first-order approximations that can be refined by the energy minimization and molecular dynamics calculations.

7.

Functional Aspects of Proteins and Nucleic Acids

We turn our attention from structural considerations to the more complex questions surrounding biological function. This section contains discussions of enzyme catalysis, protein design, and ligand/substrate design.

CATALYSIS

The Theory of Enzyme Catalysis

Enzyme catalysis is one of the most crucial and certainly the most intriguing aspects of the kinetic behaviors of proteins. The central question about catalysis is, which aspects of protein structure and dynamics cause the often enormous enhancements of rates of reaction over the rates in water? This question is, at best, incompletely answered and at worst very much open. One can not expect that molecular dynamics, with its use of a classical mechanical forcefield, will, by itself, provide definitive answers. Empirical potentials in molecular dynamics or molecular mechanics calcu lations are approximations chosen to model thermodynamically stable minima in energy surfaces. Catalytic reactions are by their very nature dependent on barriers or maxima in these surfaces. Not even the form of the potential used in most classical calculations would be correct. However, it is reasonable to think that this problem will eventually be solved by an approach that combines molecular dynamics and quantum mechanical methods. Molecular dynamics techniques can handle the motion of many atoms, and quantum mechanics can be used to represent events along the reaction pathway at the catalytic site and in the reactants (substrates), that is, wherever chemical bonds are broken and/or formed. Recent studies of simple organic reactions in solution by Jorgensen (in press) and studies of enzyme mechanisms by Warshel (1986) exemplify this combined approach.

In work on simple organic reactions in solution, the reaction path has been investigated by a series of ab initio quantum me chanical calculations of the reactants in vacuo in different states of reaction, and molecular mechanical (Monte Carlo) simulation of the solvation of each state. When combined, the results of these two calculations yielded estimates of the free energy profile along the reaction coordinate, from which reaction kinetics can be estimated. Although the quantum mechanical calculations did not take into account the response of the reactants to the solvent environment, Jorgensen (in press) nevertheless obtained very promising results for three different reaction types: SN1, SN2, and addition reactions.

Parallel studies of enzyme mechanisms pose additional prob lems, simply because the systems, including the reacting species, contain many more atoms. Except in a few simple reactions such as catalysis by carbonic anhydrase, the substrates are much larger than the reactants in the models chosen by Jorgensen. Also, in many enzyme-catalyzed reactions, a chemical bond forms between enzyme and substrate in an intermediate product of the reaction. In these cases the ab initio quantum mechanics calculation, which by its nature is currently restricted to systems of a few atoms, will have to be performed on a fragment or fragments of the chemically reacting species. It is not yet clear how this can be done without introducing large errors. Warshel has introduced the use of a more approximate quantum mechanics method, the ab initio empirical valence bond (EVB) method, into this problem to replace the quantum mechanical component. Although it can handle more atoms, the EVB method initially had the drawback of having to be calibrated by simulations designed to predict known molecular properties. In this instance, acidities of ionizing groups were used. However, this was done successfully, and Warshel and Sussman (1986) and Hwang and Warshel (1987) found that the method now can rationalize observations of changes in catalytic efficiency of mutant enzymes. These results emphasize the critical role of stabilization of the reaction's transition state by electrostatic interactions. Because of the empirical character of the EVB method and a lack of general experience with it, these conclusions await confirmation through further study with the EVB method and ab initio methods.

Given the great interest in the theory of enzyme catalysis, investigators have already begun to apply a combination of ab initio quantum mechanics and molecular dynamics (Rao et al., 1987). This will generate new problems to be solved. As noted, a major problem will be encountered for those enzyme reactions in which a chemical bond is formed between enzyme and substrate in an intermediate step. It will also be necessary to establish the magnitude of the systematic error caused by transfer of a quantum mechanical result obtained without solvation, not only to a solvated situation, but also to a protein active site environment, where the rates are much enhanced. Some very careful work will be required before this approach can be applied reliably to enzyme mechanisms. However, caveats notwithstanding, these studies are worth doing.

DESIGNING NEW PROTEIN STRUCTURES

The following section discusses the future of protein design, which is one of the key areas of growth in macromolecular modeling.

In the same way as Levinthal's (1966) pioneering studies initiated computer-assisted modeling, Richardson's (1981) work on the anatomy and taxonomy of proteins signaled the transition from molecular archeology to molecular design, for protein chemists. We call our drug design project computer-assisted molecular design—there are more types of molecules than proteins. Richardson presented a framework for understanding the organization of protein architecture. This framework condenses the observations of protein structure and architecture from individual structure solutions into a form that allows us to think of protein architecture as manipulable. Site-directed mutagenesis of protein structure is a simple way of altering proteins without really altering protein architecture. The technique of producing chimeric proteins by gross manipulation of gene structure is another early approach to the manipulation of protein architecture. At this point, the problems involved in designing new proteins are formidable. Consider the process of designing an ordinary protease. Proteases typically have 250 amino acids. Since there are 20 amino acids, there are 20^{250} possible proteins of length 250. The only sensible way to reduce the number of possible proteins to one design is to use several levels of decomposition that specify how portions of the protein are to be organized architecturally, spatially, and functionally. Our understanding of the rules of thumb triggered by the Richardson paper is now evolving rapidly.

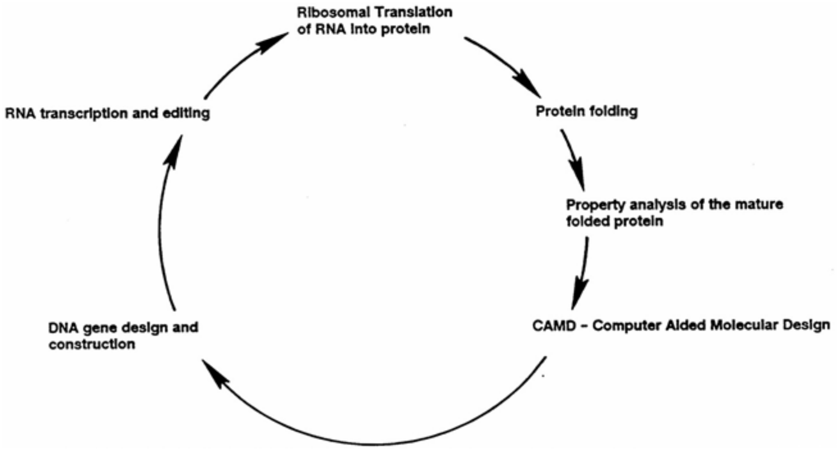

Ribosomal Translation of RNA into protein
Protein folding
Property analysis of the mature folded protein
CAMD - Computer Aided Molecular Design
DNA gene design and construction
RNA transcription and editing

FIGURE 7-1 The cycle of protein design and expression.

Protein design cannot be divorced from the issues of protein expression and folding. The complete cycle for the design, expression, and folding of proteins can be represented as inFigure 7-1.

To be able to design a protein effectively, one must be able to traverse this design cycle rapidly and often. At present, several important conceptual problems prevent us from completing this cycle at all. Suppose that we could design a hypothetical protein using the architectural concepts. The output would be a three-dimensional model of the protein embodying a particular amino acid sequence. Given this hypothetical design, the next task would be to construct a gene. The problem here is that the amino acid sequence of the hypothetical protein, in general, specifies only two of the three bases in each codon of the gene. One way to resolve this issue is to choose a random third base (See Figure 7-2).

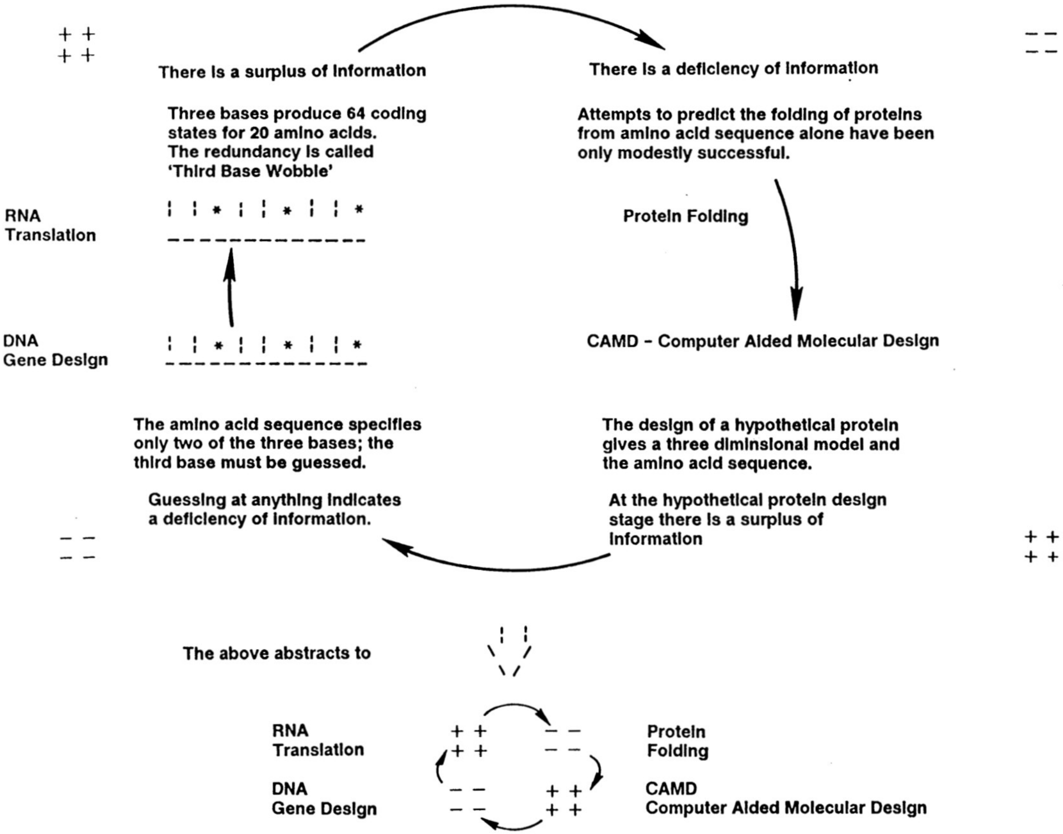

+ +
+ +
There is a surplus of information
Three bases produce 64 coding states for 20 amino acids. The redundancy is called 'Third Base Wobble'
RNA
Translation
DNA
Gene Design
There is a deficiency of information
Attempts to predict the folding of proteins from amino acid sequence alone have been only modestly successful.
Protein Folding
CAMD – Computer Aided Molecular Design
– –
– –
The amino acid sequence specifies only two of the three bases; the third base must be guessed.
Guessing at anything indicates a deficiency of information.
The design of a hypothetical protein gives a three diminsional model and the amino acid sequence.
At the hypothetical protein design stage there is a surplus of information
The above abstracts to
RNA
Translation
DNA
Gene Design
Protein
Folding
CAMD
Computer Aided Molecular Design

FIGURE 7-2 The current pattern of the flow of information in the cycle of protein design and expression.

Once the DNA of the constructed gene has been transcribed to messenger RNA and edited, there is a linear sequence of three-base codes that, as Nirenberg (1965) has described, exist in 64 combinations of the bases. However, the 64 codons code for only 20 amino acids, so there is, at this point in the cycle, a surplus of information. The ribosome translates the messenger RNA codons into nascent polypeptide.

The Anfinsen (1975) experiment involving the denaturation and renaturation of ribonuclease has been used to convince us that proteins fold into their active three-dimensional structure solely on the basis of the information contained in their sequence. Many scientists have been trying for the last two decades to predict the secondary and tertiary structures of proteins from the amino acid sequences alone. Their efforts have met with partial success at best. We lack information about the protein folding portion of the design cycle. The surpluses of information (denoted by pluses in the upper portion ofFigure 7-2) and the deficiencies of information (denoted by the minuses) can be abstracted to form the cycle pattern in the lower portion of the same figure. Clearly, our current perception of the design cycle is flawed. Recent experiments show, for example, that a gene that is moved from its native host to another expression vector does not necessarily produce well-folded proteins. Even when the codon utilization statistics of the new expression vector are mimicked, complete protein folding does not necessarily occur. This suggests that third base redundancy may be partially used to control protein folding, especially for complex proteins.

Experiments should be designed to explore how third base redundancy influences protein folding. With such data, the information from a hypothetical protein design could be used to properly construct the DNA of a gene. A proper gene would then transcribe and translate properly to yield a polypeptide that folds properly. If these conditions were met, the design cycle abstraction would be as shown inFigure 7-3.

If the information flow around the protein design and implementation cycle is preserved, then it should be possible for protein engineers to rapidly traverse this cycle in the design and perfection of novel proteins.

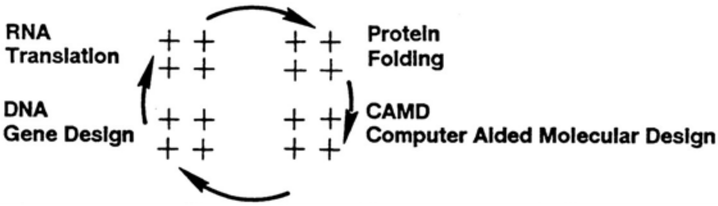


FIGURE 7-3 The ideal pattern of flow of information in the cycle of protein design and expression.

Computer Representation

Computer-assisted modeling of molecules has been evolving since the original work of Levinthal (1966). His work was the first time that a computer, a PDP-1 from the then-infant Digital Equipment Corporation (DEC), was used to draw the three-dimensional structure of a small organic molecule. It used one line segment to represent each chemical bond. With simple software controls, the molecule could be rotated in space and redisplayed. Similarly, the conformation of the molecule could be changed by rotating one portion of the molecule around a bond that formed an isthmus between it and the remainder of the molecule. The display of the molecule was done in pairs of images where the image of one molecule was rotated 5 degrees around the vertical axis. This produced a stereoscopic effect that permitted the three-dimensional structure of the molecule to be perceived without having to rotate it continually. All of molecular graphics has simply been an extension and refinement of these powerful ideas. The number of line segments drawn per second has risen dramatically. Color has been added. Hardware stereo devices have been developed, and very recently powerful array processors have been added to permit the rapid calculation of molecular energetics during modeling.

These techniques for display and modeling were developed and refined in a few academic research laboratories. They began to diffuse to biochemical and genetics laboratories in academic and industrial institutions worldwide. Over the past 20 years, the manufacturers of computer and graphics hardware have begun to recognize that molecular graphics and modeling is a substantial market. We are now at the critical point in this respect. Hardware manufacturers are now willing to design workstations (i.e. integrated computational and graphics machines for individual use) for the molecular modeling and design market. A new class of workstations is expected in the next year. Two members of the class of personal supercomputers (PSCs) have been identified, and collaborations are in place to insure that these machines, when they enter the commercial market, will be fully conditioned "chemistry engines". The four functions, molecular energy computation, molecular configuration control, molecular graphics, and reasoning about molecular structure, will be integrated in one computer system.

The PSCs will provide a nearly ideal package for mass distribution of CAMD capabilities. Market forces can be expected to expand the number of different machines and the features that each machine offers. Standards at various levels, as defined by the International Standards Organization (ISO), will permit existing and new program systems to be transported rapidly onto the PSC class members. The standardization efforts will permit a decoupling of the computational support systems (i.e. hardware, graphics, operating systems, and the molecular modeling and design programs) from the intellectual uses of such systems.

The existence of standards, however, does not guarantee portable program code. Scientists who write new programs must know about these standards and write programs that conform to them. Commercial organizations that take existing scientific programs should shape them towards the standard style because, in the end, the size of the commercial market will depend on the ability of end users to piece together working systems from components made out of various standard programs. If these standardization efforts succeed, then in the future, the molecular modeling community will be able to routinely make smooth transitions to more powerful computer support systems.

Computer graphics representations offer alternative ways of understanding molecular structure and function. They started as the simplest white line drawings on black screens, then progressed to color images, to solid surfaces, to dot surfaces, and to electrostatic surfaces. Intergraph three-dimensional representations and white light hologram representations have been developed and used for molecular structure problems. Intergraph is composed of approximately 20 individual photographs where vertical strips are selected from each photograph and composed into one image. The composite image is viewed through a linear fresnel lens. The next generation of workstation, the PSC, will offer ray-traced images as part of the operating system. In a ray traced image the reflections on a surface are compared by calculating the trajectory of light beams from all possible light sources. This produces in the extreme the reflections of one object on another. A truly three-dimensional representation where the molecule would actually occupy three-dimensional space is needed. A breakthrough in a field such as plasma physics is necessary to make this a reality.

Impediments to Progress

The central bottleneck to progress in protein design is our inability to predict protein tertiary structure from amino acid sequence. The notion put forth by Anfinsen 25 years ago was that the amino acid sequence alone determines tertiary structure. This notion may be too simplistic, and there may indeed be a higher level code than the Nirenberg nucleic acid to amino acid conversion by the ribosome. Since the Anfinsen conjecture and the experimental detail surrounding it are largely prohibitory in nature, they had the effect of discouraging experimentation in expression and folding of proteins. Scientists who are concerned with protein expression are content with the Nirenberg code and explain away anomolous results because they see no need for any other effect. Scientists concerned with protein folding cannot explain how proteins fold, but then are discouraged by the Anfinsen conjecture from asking for more information from the geneticists. A theory and experiment linking codon utilization in gene structure with the folding of protein structure would be a major step toward reconciling these views.

PREDICTING FUNCTION FROM A PREDICTED THREE-DIMENSIONAL STRUCTURE

In principle, the information needed to predict the function of a biological macromolecule is encoded in its three-dimensional structure. We assume that we must know the three-dimensional structure of a macromolecule before we can fully understand its function. The problem is, how do we decode the rules that govern the relationship between structure and function? A subset of this problem will be discussed below: the prediction of the change in the functioning of a protein that results from the binding of a ligand.

Recently, the possibility of computer-assisted drug design based on the three-dimensional structure of target biomolecules has received much attention in the scientific literature (Beddell, 1984; Goodford, 1984; Hol, 1986). Those in the field believe that medicinal chemistry is poised to undergo a revolution as dramatic as the events in the 1950s and 1960s that transformed organic chemistry from a descriptive to a predictive science. Since we are at the beginning of a new age, the many challenges ahead do not diminish the excitement of knowing that the solutions are also on the horizon. We have a sense that, at last, we know what it is that we have to learn and have at least the rudiments of the necessary tools at hand.

In anticipating this revolution, we are presupposing that we can or soon will be able to predict the functions of proteins from their structures. In particular, we would need to be able to predict the ability of a protein to recognize and bind a ligand and to predict the structure of the "optimum" ligand. Beyond that, however, we would need to be able to predict how the protein carries out its function and how it recognizes and interacts with other macromolecules to alter its own functions and theirs. Although we have learned much about these topics, there are unanswered questions that we must be able to answer before we will be able to make accurate predictions.

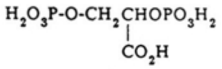
Experience in Ligand Design from Experimental Protein Structures

One illustration of our current state of achievement is given by work on hemoglobin. Ligands affect the function and properties of hemoglobin in complex ways. Investigators began to attempt to design ligands based on the three-dimensional structure of a protein as soon as such structures were available. In the early 1970s, the group headed by Goodford at Wellcome Laboratories in England began to explore the possibilities of ligand design by receptor fit (Beddell, 1984; Goodford, 1984). They used the structure of hemoglobin as determined by protein crystallography and constructed a wire model that was hinged so that they could examine both the oxy- and deoxy-states.

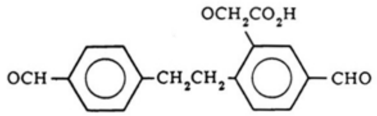
The first studies involved the design of ligands (using mechanical models) to fit the diphosphoglycerate (Figure 7-4, compound 1) binding site and then to mimic its function. The investigators used simple concepts of complementary shapes, electrostatic interactions, and possible covalent bonds. The designed compounds (Figure 7-4, designated compounds 2-4) do indeed mimic the effect of diphosphoglycerate on the dissociation of oxygen from hemoglobin. Subsequent crystallographic work supported the proposed binding mode. In addition, the relative binding energy of various analogues to a number of different hemoglobins was measured for 29 protein-inhibitor combinations. Statistical analysis revealed a highly significant correlation between the strength of binding and the number of covalent and ionic interactions. The use of computer graphics for the design would have accelerated this process since it took three months to construct the physical wire model of the protein.

This work was then expanded in an attempt to design a compound for the treatment of sickle cell anemia. The goal was to develop a compound that would affect the oxygen-dissociation curve in a way opposite to that of diphosphoglycerate. An intensive biochemical, physiological, and structural examination of the problem suggested that a ligand that binds between the alpha subunits of oxyhemoglobin might have the desired effect. Since no natural ligand for this site was known, the ligands were designed from the protein structure alone and designated compounds 5 and 6 (See Figure 7-4). Although the proposed binding mode has not been experimentally verified, the designed compounds did produce the expected change in function of hemoglobin. One of the compounds is now in clinical trials for the treatment of sickle cell disease. Thus, using rather primitive tools, the Wellcome group was able to predict the effect of a small molecule on the function of a protein.

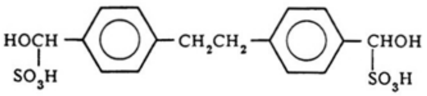
The recent experience of Perutz et al. (1986) emphasizes both the important accomplishment made by these workers and the limits of our molecular understanding. Perutz and coworkers experimentally demonstrated several of the potential binding sites that a molecule might recognize in hemoglobin. Specifically, they solved the crystal structure of eight ligand-hemoglobin complexes and showed that there are at least six different positions on the protein at which a ligand might form a tight complex. Since the ligands were selected on the basis of their perceived structural



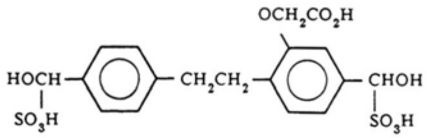
1



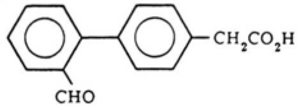
2



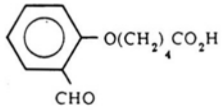
3



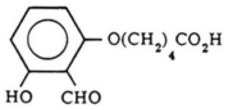
4



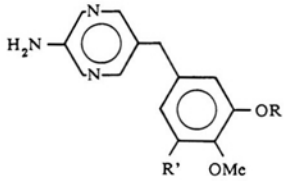
5



6a



6b



7 $R = Me, R' = DMe$

8 $R = (CH_2)_5CO_2H$ $R' = OMe$

9 $R = CH_2CH_2NHSO_2\phi R' = H$

10. $R_1 = CH_3, R_2 = NH_2, R_3 = HC = N$

11. $R_1 = H, R_2 = OH, R_3 = H_2C - NH$

FIGURE 7-4 Some molecules synthesized to aid in elucidating the relation of structure to biological activity of some macromolecules. See text for details.

similarity, this result must be carefully considered by those who attempt to design ligands to fit a particular site on a protein. Three of the molecules bind in sites that overlap; two of these raise the minimum gelling concentration of hemoglobin S, whereas the third lowers it. Each of the bound ligands changes the structure of the proteins so little that the change is barely detectable, yet some of the ligands increase the gelling concentration, some decrease it, and others do not change it at all. This work makes it clear that even after studying structure and function of hemoglobin for 25 years, an investigator may still be puzzled by the functional consequences of the minute structural changes that accompany ligand binding.

The work of Perutz et al. (1986) also illustrates that the design of a drug is more complicated than the mere design of a tightly bound ligand. Clofibrate raises the gelling concentration of hemoglobin S. Also, it has been used clinically for other disorders and so is known to be absorbed, metabolized, and nontoxic. Yet it cannot be used to treat sickle cell disease because it is so tightly bound to serum albumin that it is not available to bind to the hemoglobin. The lesson to be drawn from this is that when we design a new drug from theoretical principles, we must somehow incorporate the possible interaction of the proposed ligand with all other macromolecules of the body.

The work cited and other studies on hemoglobin underscore both the promise and the challenges in predicting the changes in the function of a protein that are brought about by formation of a protein-ligand complex. This work also highliglits the further challenges of predicting all of the interactions of the ligand with the organism.

The design of inhibitors of dihydrofolate reductase was aided by the three-dimensional structures of the proteins. Research efforts of two pharmaceutical companies and their collaborators culminated in the crystallographic determination of the structure of the dihydrofolate reductase from several species, some with bound ligands (Beddell, 1984; Blaney et al., 1984; Goodford, 1984). Each group designed a trimethoprim (Figure 7-4, compound 7) analogue (Figure 7-4, compounds 8 and 9), that was proposed to form a new interaction with a nearby arginine. Goodford (1985) was able to verify this interaction (Figure 7-4, compound 8) by crystallography. Both new analogues show greatly enhanced binding affinity. However, neither shows enhanced antibacterial activity, presumably because they do not easily penetrate into the bacterial cell. Additionally, neither has any antibacterial activity in whole animals, so they are not candidates for development as new therapeutic agents. Thus, although this work did successfully predict the local functional consequences of structural modification of a previously marketed drug, it did not predict all of those properties needed to convert a biologically interesting compound into a therapeutically useful one.

In summary, efforts to design inhibitors of dihydrofolate reductase have shown that, although knowing the structure of the target biomolecule is very useful when designing a ligand, this alone is not enough information to design a drug. Furthermore, these protein structures have been available for approximately five years, yet neither company has capitalized on them, nor have other research groups used the published enzyme structures to design new compounds.

Many other groups have used crystal structures of proteins to design analogues of known ligands. In an early example, a physical model of lysozyme and a proposed mechanism of its hydrolytic action were used to design a transition-state inhibitor (Goodford, 1984). The compound inhibits the enzyme; it binds 32 times more strongly than the corresponding substrate. The proposed binding mode has also been confirmed by x-ray diffraction studies.

In the first use of color computer graphics in the design of ligands to bind to a protein, workers at the University of California at San Francisco used interactive potential energy calculations and molecular graphics to dock thyroxine analogues into the binding site in the crystallographic structure of pre-albumin (Blaney et al., 1982). They designed and synthesized several more strongly bound ligands. The predictions were confirmed experimentally. Since the function of pre-albumin appears to be transport of thyroxine, no further predictions of the consequence of ligand-protein interaction were made.

Experience in Ligand Design from Predicted Protein Structures

Protein structures modeled by homology to proteins whose three-dimensional structure is known have also proved useful in the design of novel ligands. For example, workers at two different pharmaceutical companies have used a structure of the enzyme renin that was modeled from other members of the aspartic proteinase class (Anonymous, 1986; Boger, 1986). Such models suggested the structures of new inhibitors. The compounds were shown to be potent inhibitors in vivo as well as in vitro.

Approximate target macromolecule structures have also been used to design new agents. The classic example is the design of captopril (Figure 7-4, compound 9), an inhibitor of angiotensin-converting enzyme and a clinically successful antihypertensive agent (Petrillo, 1982). Captopril was designed from a proposed structure of the substrate when bound to the enzyme. The structure of the enzyme was assumed to be similar to that of carboxypeptidase A because of mechanistic similarities between the two enzymes.

Inferring Binding Sites

Much of this section has addressed issues related to determining and analyzing the structures of proteins of known sequence but unknown three-dimensional structure. Once these structures are known, detailed studies can be carried out of the relationships between those structures and the corresponding functions of a protein. Proteins express their functions through binding of other molecules, often termed effectors, with or without concomitant transformation of the effector, e.g., degradation or chemical reactions at functional groups. We have discussed structure/activity studies of the binding of effector molecules to putative sites in a protein of known structure. A separate body of research has focused on a complementary problem: relating the structures of several effector molecules to one another in order to determine information about binding sites, often termed active sites, in proteins of unknown structure. When such studies are successful, they can obviously provide structural information about a protein that can be used in conjunction with some of the techniques for structure determination discussed earlier in this report.

The general problem of inferring binding sites can be stated simply. Given a set of molecules that are presumed to bind to the same site in a given protein of unknown structure, we must infer the size, shape, and binding characteristics of the active site. Several problems are subsidiary to this general problem. We mention them here, but a detailed analysis is beyond the scope of this report. For example, one must consider the process of*recognition* of the effector molecules prior to the actual binding in the active site. One must consider the possibility of conformational changes of both an effector and an active site during recognition and binding. One must perform very careful studies to ensure that the measured biological or chemical responses for several effectors are in fact due to binding in the same active site. A useful introduction to this research area, with leading references, has been presented (Olson and Kristoffersen, 1979).

At least three approaches have been used to infer the structure of receptor sites:

- the*receptor mapping* aproach of Humber et al. (1979);
- the*active analog* approach of Marshall et al. (1979);
- the DYLOMMS program of Wise et al. (1983).

These approaches are closely related. All follow the same basic principles, but in different ways. All begin with the assumption that similar effectors possess related*pharmacophoric patterns*, i.e., similar dispositions in three-dimensional space of similar structural features important for binding. Independent studies are used to postulate pharmacophoric patterns of active molecules, generally using the most conformationally rigid molecules to form hypotheses. Once such a pattern is assigned, all possible conformations of each effector are examined to determine if there are low energy conformations that present the pattern. This both tests the hypothetical pattern and begins building a set of molecules that can be superimposed based on the pattern. Once superpositions are established, the volume occupied by the molecules can be used to define the cavity of the active site. Molecules of related structure that can yield the pharmacophoric pattern, but that display no activity, can be used to define the walls of the cavity, further elaborating its shape.

Recently, practicing medicinal chemists have become enthusiastic about these uses of computational and computer graphics techniques to compare the three-dimensional structures of ligands that bind to a receptor. They use the common features of the aligned structures to propose tentative maps of the receptor topography. These maps are then used to design new compounds (Ghose and Crippen 1985; Hopfinger, 1985; Humblet and Marshall, 1981). These techniques have benefited from the knowledge gained through protein crystallography. In particular, current applications of receptor-mapping methods usually compare the location of the projection of ligand atoms to possible binding sites, rather than identifying the location of the ligand atoms themselves as had been done previously.

The final category of computer-assisted prediction of the biological properties of a small molecule is also the oldest. This type of methodology, Quantitative Structure/Activity Relation ships (QSAR) uses statistical or pattern recognition methods to explore the possible relationship between the biological and physical or substructural (presence or absence of certain functional groups) properties of molecules. Given the known utility of QSAR methodology to predict the potency of untested analogues (Hopfinger, 1985; Martin, 1981), it is important that the developers of this methodology are actively pursuing the challenge of evaluating the reliability of linear free-energy equations for cases in which the protein structure is known. In the case of dihydrofolate reductase, several investigators have compared the conclusions from QSAR and molecular graphics modeling of the inhibitors (Blaney et al., 1984). The conclusions derived from the two methods agree closely, confirming the proposal that the QSAR equations contain information about the types of noncovalent interactions between the inhibitors and the enzyme. However, a major advantage of QSAR over other computer-based methodologies is that one can attempt to develop equations for any biological response. For example, equations have been developed for the enzyme inhibition, antibacterial, and whole-animal antitumor activity of dihydrofolate reductase inhibitors (Blaney et al., 1984). Thus, QSAR is a logical complement to the more structure-based computer methodologies. It could be used to model the potential whole-animal activity of new ligands and perhaps to search for unanticipated interactions with other macromolecules.

Computer Tools for Ligand Design from Three-Dimensional Protein Structure

The recent excitement in computer-assisted drug design has arisen because scientists now have available the elements of each of the important tools for such an activity. Two types of computer hardware are necessary: high-speed color graphics and affordable but powerful computers dedicated to modeling. In addition, a growing body of data on the three-dimensional structure of proteins is becoming available, as our understanding increases of some of the relationships between structure and function of proteins.

Finally, software is also available for the graphics display of the molecules and for modeling the energetics and thermodynamics of the binding.

Specialized graphics tools for molecular design have also been developed. Some of these arose from the related activity of docking a known ligand into a protein. The display of the surface of the binding site is more useful for ligand design if it is color-coded to suggest the preferred type of noncovalent interaction at that point in space. For example, through such displays we can distinguish between surfaces near positively charged, negatively charged, hydrogen-bond accepting, hydrogen-bond donating, and hydrophobic regions of the protein.

Another helpful tool used with the graphics display is the immediate read-out of energy values as the ligand is docked into a putative binding site and as bonds in the ligand and/or the protein are rotated to facilitate the docking. Design of ligands at the computer screen is aided by stereoscopic viewing devices and implements that allow one to move an object being displayed (such as a ligand) in three dimensions while keeping the rest of the display as it was. Experience has shown that molecular mechanics energy minimizations are necessary to evaluate the geometry and energy of the proposed complexes (Pincus and Scheraga, 1979).

It was noted previously that one persistent but often hidden problem in ligand design is that a ligand may bind to a protein in a different orientation or at a totally different site than the investigator anticipated. Kuntz et al. (1982) have devised a computerized means of evaluating such possibilities based on shape alone.

The design of a new ligand molecule is aided by the graphics display of the energetically preferred sites on the protein for interaction with various types of possible ligand atoms (Goodford, 1985). Such sites are identified as the energy of interaction of the probe atom at each point on a three-dimensional grid surrounding a protein. The ligand would be designed to interact at as many of these sites as feasible.

Once a proposed ligand is designed, its thermodynamics of binding can be predicted with the free-energy perturbation method if it is a reasonably close analogue of a known compound.

If there are data on the relative energy of binding of other ligands to the protein, a QSAR or receptor mapping analysis discussed above may suggest regions on the target that are conformationally more flexible than the experimental structure may suggest. QSAR (or at least consideration of physical properties) is expected to also be useful in the design of ligands that will have the appropriate whole-animal properties.

Impediments to Ligand Design from Protein Structures

Proteins are conformationally mobile. They are not the static structures that the graphics display of the crystal structure suggests. For example, molecular dynamics calculations on myoglobin have shown that within 300 picoseconds, 2,000 different conformational minima are sampled (Elber and Karplus, 1987). The root-mean-square difference in the location of the atoms in the most different structures is 2 Å; this means that many atoms move substantially more than that.

Proteins also change conformation when ligands are bound to them; hence, ligand design methodologies must be able accurately to predict such movements. For example, when the antiviral compound VIN 52084 is bound to the human rhinovirus, 13 residues of the protein undergo measurable conformational change (Smith et al., 1986b). The main chain moves as much as 3 Å, the channel to the binding pocket opens to the solution, the isoelectric point of the system changes from 6.9 to 7.1, and the occupancy of Ca^{++} at a distant point on the virus increases.

Conformational responses to ligand binding may be part of the function of the protein. For example, in response to Ca^{++}, the channel-forming proteins of the gap junction between cells show small cooperative rearrangements of the relative orientation of the subunits. This rearrangement results in the narrowing of the diameter of the Ca^{++} channel within the cell by 18Å and thus closes the channel to Ca^{++} passage (Unwin and Ennis, 1984). During this rearrangement, the conformation of each subunit does not change appreciably, only the orientation of each subunit changes with respect to the others.

Conformational responses to ligand binding may form the basis of the selectivity of ligands for very similar proteins. Evidence from crystallography, QSAR, and molecular graphics suggests that conformational changes in the enzyme in response to the binding of ligands is responsible for the selectivity of trimethoprim for bacterial dihydrofolate reductases in contrast to vertebrate enzymes

(Blaney et al., 1984). In the chicken liver enzyme, a tyrosine residue moves 5.4 Å in response to the binding of trimethoprim.

Since there is no experimentally established three-dimensional structure of a membrane-bound receptor, for this type of protein we depend on indirect observation and inference for our notions about conformation and conformational changes in response to ligand binding. Current concepts of receptor function usually invoke a conformational change as part of the transduction of the signal of a binding event into the ultimate biochemical and physiological response. Thus, it is possible that the regulatory and second messenger binding sites on receptor proteins might become available only in the presence of the ligand. Furthermore, that certain compounds only partially activate a receptor suggests the possibility that a whole family of receptor conformations is available.

Thus, to use protein structure design a ligand that influences the action of a protein whose function requires more than one conformation or in which the putative binding site is very flexible, we would like to know the relevant three-dimensional structures of that protein and be able to predict the conditions under which each is stable. In other words, we would find it difficult to predict the function of a new ligand unless we had available structures of these protein conformations. We see this as a problem that will require at least as much study as the problem of finding the global minimum energy structure.

The ligand binds to a protein that is part of a system. In solution, a protein is part of a complex with water, ions, and cofactors. Alternatively, it may function while interacting with a membrane. These other species affect the strength of binding of the ligands of interest. For example, trimethoprim binds to dihydrofolate reductase with a 10,000-fold increase of affinity in the presence of its cofactor compared to its absence.

The covalent structures of some proteins are modified during the course of their function. The large family of receptor kinases are responsible for phosphorylatation of receptors as a means of regulating that receptor function. Thus, the addition of a single phosphate group to a protein can dramatically alter its function.

Other proteins are not functional until they are structurally modified after their synthesis. For example, sperm binds to its receptors on the egg only if these receptors are glycosylated. Additionally, posttranslational processing can impart subtle variations in properties to a protein. For example, it is thought that the benzodiazepine receptor is the same protein throughout the brain, but that it is glycosylated to a different extent in different regions of the brain. These differences in glycosylation are reflected in different relative affinities of the receptor for various ligands.

Thus, for accurate and realistic models on which to base theoretical ligand design, we need to be able to include such other species in the calculation. Unfortunately, there are often not molecular mechanics parameters for such cofactors and transition metals. Including these additional ions and molecules increases the complexity and time of the calculation enormously, partly because the number of atoms is increased but more dramatically because the search for the stable arrangement of atoms is much more complicated. This is the multiple-minimum problem, but with even fewer experimental constraints on the solution of the problem. Furthermore, we cannot use traditional molecular mechanics concepts for transition metals because they undergo changes in oxidation and spin states that dramatically affect the optimum geometric arrangement of ligands. To include such ions, we need a combination quantum and molecular mechanics calculation. Although progress has been made in such calculations (Warshel, 1981; Singh and Kollman, 1986), they still need refinement and testing and tend to be calculations that strain available computers. Thus, we see promise that the tools required will be available, but they are not yet in routine use.

There may be more than one binding made for the ligand. The experience with the binding of ligands to hemoglobin and the different binding orientations of methotrexate (Figure 7-4, structure 10) and dihydrofolate (Figure 7-4, structure 11) to dihydrofolate reductase highlight this problem (Blaney et al., 1984). The method of matching ligand shapes to protein cavities is helpful in predicting such alternate binding modes. However, it is currently limited because it considers only the correspondence of the shape of the ligand and the binding site and not their possible flexibility or electrostatic and hydrophobic contributions to binding energy. In principle, this problem could be solved by examining the relative energy of all potential conformations of the protein and the ligand and all potential relative orientations of the two. As noted above, for such calculations water and cofactor molecules and associated ions should also be included. Even if there are only two conformations of the protein each with two binding sites and two conformations or enantiomers of the ligand, the problem increases eight-fold! The challenge escalates when we consider that, in drug design, we would like to consider many possible analogues for synthesis. Thus, much more sophisticated techniques for pruning conformational and orientation hyperspace need to be developed before detailed calculations of this magnitude will be possible.

Even if we could predict the mode and strength of binding of a ligand to a protein, the effect of such binding on the function of the protein in the cell might not be obvious. The simplest case would seem to be the design of an enzyme inhibitor. If an enzyme is inhibited, we would expect that fewer substrate molecules would be transformed in a given unit of time. However, this is not necessarily true. For example, current evidence is that receptor kinases are present in the cell in high concentrations: the rate of phosphorylation of the receptor is apparently governed by the concentration of the cyclic nucleotide and the conformational state of the receptor and not the level of the enzyme. Inhibition of such an enzyme by even 90 percent might have no observable physiological effect. In other cases, the level of a particular enzymatic activity is regulated by feedback control. Inhibition of such an enzyme would be overcome by production of more enzyme. Alternatively, inhibition of an enzyme might simply lead to the presence of higher levels of substrate but the same rate of turn-over of substrate through the biochemical system. The physiological effects of such agents may be impossible to predict.

The situation is even more complex in proteins that have multiple domains that control multiple functions. A compound that prevented sickling of hemoglobin S would be useless as a drug if it also prevented oxygen binding or release or if, when bound, it promoted the crystallization of hemoglobin in a different crystal form.

A further complication in trying to understand function from structure is that a single protein may interact with several small molecules and other proteins in a complex regulatory scheme. Different subunits of domains of a protein may have different but interrelated functions. For example, all four subunits of*Torpedocalifornica* acetylcholine receptor are necessary to elicit a nicotinic response to acetylcholine, whereas only the alpha subunit is refquired for binding the antagonist alpha-bungarotoxin (Mishina et al., 1984). Thus, the structure of the alpha subunit might help in the design of a ligand, but the structure and function of all four subunits might be needed to predict whether the compound would be an agonist or antagonist.

Other factors might make a ligand useless as a therapeutic agent. When a ligand is administered to an animal, it must survive the metabolic and structural defenses of the animal in order to reach its proposed site of action at the required concentration. The ligand may be a substrate for any one of many enzymes, some of which appear to have evolved broad specificity in order to metabolize foreign substances and thereby protect the organism from its unpredictable environment. Ultimately, we expect to be able to predict the biotransformations of small molecules from the structures of the enzymes involved, but we cannot do so today.

The ligand may also fortuitously bind to other macromolecules in the body and, as a result, may not be available to the target protein. The ligand may have the correct physical and chemical properties to be rapidly excreted into the urine or bile before it has a chance to move to its target. Finally, the ligand may be so slightly soluble that it cannot achieve high enough concentrations in the blood or gastrointestinal tract for it to be distributed to its site of action. Again, we have some informal rules that allow us to attack these problems, but lack the basic knowledge we need to make true predictions. A ligand might also be useless in curing disease because it or one of its metabolites produces toxicity in the animal.

To use a ligand as a drug, it must be technically feasible to do so. This means that it must be possible to produce the compound in the required quantities and purity; it must be stable enough to ship to the patient; and an acceptable pharmaceutical form of the compound must be devised. A major advance has been made in the computer-assisted design of pathways for the synthesis of compounds. However, further enhancements would make this tool even more useful.

Economic factors also figure into feasibility; if the compound is to be sold, the patentability of the compound, the cost of its manufacture, the cost and effectiveness of competing therapy, and the expected incidence of the disease for which it is effective will also be issues in the decision to market the compound.

Other complications may also emerge when one is predicting function or designing ligands from predicted three-dimensional protein structures. First, the confidence in the exact coordinates of the protein structures will be lower. This greater uncertainty will complicate the investigation of proposed function or the design of ligands because the exact dimensions of the possible binding sites will be uncertain, as will the conformation of residues on the surface of the protein. In principle, these questions can be answered using extensive molecular dynamics and minimization calculations. The prediction of function might be straightforward if the unknown protein shows a strong sequence homology with a protein of known structure and function.

Another complication with the use of predicted structures is that we may be unaware of posttranslational modifications of the structure. Ultimately, we expect to be able to predict such modifications from the substrate specificities of the enzymes that perform them. However, we cannot do so today.

Consideration of protein structures based on DNA sequence may obscure the fact that the protein may function as part of a multisubunit assembly. Multiple subunit proteins are common. To predict the function of such a protein, we must realize that it binds to the other subunits. It is not enough to consider other proteins coded on the same chromosome; the genes that code for the two different protein chains that form the subunits of hemoglobin are located on different chromosomes. Hemoglobin illustrates a further complication in using DNA sequences: there are at least four different variants of the beta subunit. Only one of these is produced in quantity by the organism. Thus, to predict the function of the alpha chain of hemoglobin, we would need to recognize that it functions in a tetrameric structure with two subunits of a different type, and that, of those with which the alpha subunits could bind, only the beta subunit is produced in appreciable quantity.

The transcribed protein may have one activity and be transformed into a product that has a different activity. Peptide hormones usually arise by the limited hydrolysis of a larger protein that circulates in serum. Sometimes the same carrier protein can be cleaved at different sites to produce different peptide hormones. At present, we cannot predict such events. Only when we know the sequence of every peptide hormone would we be able to recognize the potential for a particular protein to be a carrier of a hormone.

In summary we do not adequately understand the relationship between the details of the three-dimensional structure of a protein and its function. Without such an understanding, we cannot predict the effect that a bound ligand will have on the function of the protein. We lack this understanding partly because three-dimensional structures of proteins have been determined only recently, and molecular graphics hardware and software are also newly available to experimental scientists. But in many cases, we do not know the three-dimensional structure of the protein of interest, nor do we have a good idea of all of its functions. We know even less about the relationship between structure and function of carbohydrates, because we have so little structural information on them. This is a problem that will not be solved in the short-term.

While there are methods to predict the potency of molecules once a structure is suggested, we need better tools for molecular design to help the chemist suggest molecules to examine experimentally or theoretically. The tools described above are primitive. Although some methods are available to match candidate molecules against proposed shape requirements for binding, it is not possible to also specify the chemical properties of the designed compound with existing software. The current methods process a file of three-dimensional coordinates of candidate molecules; this file is generated from experimental or theoretical studies and so is incomplete. Additionally, we cannot automatically compare a compound proposed by a computer program with those already in the world literature as tested for that activity, nor can we automatically detect if the proposed compound is identical or similar to compounds known to have some biological activity deleterious to that desired. It is expected that many of these tools will be developed rather soon.

8.

Structure and Function of Complex Carbohydrates

Complex carbohydrates are very common in animals, plants, and bacteria. They are constituents of cell membranes, as well as subcellular materials of cells. They are also found in physiological fluids such as blood, tears, milk, and urine. It was estimated recently that the covalent structures of between 4,000 and 6,000 natural carbohydrates have been determined (DOE, 1987). Many complex carbohydrates are unsubstituted at their reducing ends and are referred to as polysaccharides; examples include the oligosaccharides of milk, the cellulose of plant cell walls, and storage forms such as starch and glycogen. Many other naturally occurring complex carbohydrates are covalently connected to other molecules, such as proteins or lipids, by glycosidic linkages of the sugar residues at their reducing ends to form glycoconjugates.

BIOLOGICAL FUNCTION

Glycoproteins have many functions in higher organisms. Collagen is an important structural element in the extracellular space and in cartilage, bone and basement membranes. Mucins are significant as lubricants and protective agents in mucous secretions. Important immunological molecules of the glycoprotein class include the immunoglobulins, histocompatibility antigens, blood group antigens of the ABO and Lewis types, complement in the blood clotting mechanism, and interferon. Many human plasma proteins such as fetuin, transferrin, and ceruloplasmin are glycoproteins, as are several of the hormones such as chorionic gonadotropin and thyrotropin. Most of the animal and plant lectins are glycoproteins, as are the lysosomal enzymes. The recognition and binding of lysosomal enzymes to specific receptors in the Golgi apparatus and on the cell surface involves one or more phosphorylated mannose residues on N-linked oligosaccharide chains. Recognition sites on cell surfaces for binding and uptake of hormones and for interactions with other cells, viruses and bacteria are also glycoproteins.

Many of the cell surface functions of glycoproteins have also been proposed for the neutral and acidic glycosphingolipids. In addition, certain glycosphingolipids of the ganglioside class have been found recently to inhibit the mitogenic response of cell growth factors by allosteric modulation of their cell surface receptors (Bremer et al., 1986). Oncogenic transformation by viral infection or chemical mutagens usually leads to alterations in the cell surface pattern of glycosphingolipids such that certain types increase greatly in quantity. In some cases, there are also qualitative differences due to the expression of genes that are silent in the differentiated normal cells. This is particularly important in tumor cells, where tumor-associated antigens may provide a basis for specific monoclonal antibody-based diagnostic assays and eventually, perhaps, treatment.

The binding between glycosaminoglycans and other extracellular macromolecules contributes significantly to the structural organization of connective tissue matrix. All of the glycosaminoglycans, except those that lack sulfate groups or carboxyl groups, bind electrostatically to collagen at neutral pH because of their remarkable anionic character. Dermatan sulfate, which appears to be the major glycosaminoglycan synthesized by arterial smooth muscle cells, binds strongly to plasma lipoproteins, and heparin also interacts with several plasma proteins, including clotting fac tors IX and XI and antithrombin III. Interestingly, the 1:1 stoichiometric binding of heparin to Lys residues of antithrombin III is believed to induce a conformational change in antithrombin III that increases the binding of antithrombin III to thrombin. This binding inactivates the thrombin. Hyaluronic acid is deposited on the surface of Petri dishes by cells growing in tissue culture, giving them a substratum for attachment during growth. The proteoglycans have also been implicated in the regulation of cell growth, possibly through nuclear effects on chromatin structure and activation of DNA polymerase, and may mediate cell-cell communication and the shedding of cell surface receptors.

BIOSYNTHESIS OF N-LINKED GLYCOPROTEINS AND GLYCOSPHINGOLIPIDS

The role of carbohydrates in biological function poses a particularly challenging problem for the future. The synthesis of these glycoconjugates occurs during their intracellular transport from the site of initial assembly of a lipid-linked intermediate (glycoproteins) or ceramide (glycosphingolipids) in the endoplasmic reticulum, through the Golgi apparatus, to the cell surface, intracellular organelles, or extracellular space. Their synthesis requires a family of activated sugar donors called sugar nucleotides that are synthesized in the cytosolic fraction of cells from sugar phosphates and nucleoside triphosphates. An interesting exception is the sugar nucleotide of sialic acid, called cytidine monophosphate sialic acid (CMP-NeuAc), which is synthesized in the nucleus from free sialic acid and CTP. The enzymes involved in glycoconjugate biosynthesis are glycosyltransferases that catalyze the transfer of sugar residues from the sugar nucleotides to the nonreducing end of a growing carbohydrate chain.

The distinction between glycoconjugate biosynthesis and protein synthesis is key; the latter occurs on a template of messenger RNA and is therefore determined by the genetic code for a single structural gene.[1] In sharp contrast, glycoconjugate synthesis is accomplished by the stepwise addition of sugar units using a different enzyme for each step. Therefore, no single DNA sequence is involved in determining the primary structure of the complex carbohydrate, since the order in which sugars are added depends on the substrate specificities and kinetic characteristics of the different glycosyl transferases, each of which is coded by a different structural gene. It is clearly impossible to predict the primary structures of complex carbohydrates from DNA sequences. Therefore, the three-dimensional structures of glycoproteins, glycosphingolipids and other complex carbohydrate-containing molecules can never be completely predicted without experimental structural analysis of the carbohydrates.

Snider (1984) reported that glycoproteins of the N-linked type are synthesized as a cotranslational event in the rough endoplasmic reticulum. While the polypeptide chain is being translated on a messenger RNA and concurrently passed through the endoplasmic reticulum membrane into the cisternal space (lumen), a single oligosaccharide is coordinately synthesized on a phosphorylated polyisoprenoid alcohol (dolichol in higher animals and smaller, similar substances in insects, yeast, and plants). The entire precursor oligosaccharide is then transferred to appropriate asparagine residues on the nascent polypeptide chain (probably before folding into a tertiary structure) according to rules of specificity that are not completely understood. Transfer requires an Asn-X-Ser or Asn-X-Thr sequence but additional factors are involved as well. Accessibility of the Asn residue may be one such factor and assessment of this possibility could be made by the predictive methods described in this report.

The second stage of N-linked glycoprotein synthesis involves extensive posttranslational modification of the protein-linked precursor oligosaccharide by the removal and addition of sugars. In many cases the protein moiety is also modified by partial proteolytic cleavages and/or the addition of function-modifying groups on specific amino acid residues. Posttranslational modification is initiated in the rough endoplasmic reticulum by the removal of the three glucose residues by two specific membrane-bound glucosidases. These glucose residues appear to have the sole function of enabling transfer of the oligosaccharide chain from dolichol pyrophosphate to nascent polypeptide chains. It will be interesting to determine from three-dimensional structures and predicted conformations how these groups interact with the transferase enzyme involved at this step. Mature high mannose oligosaccharide chains are synthesized by the subsequent removal of up to four mannosyl residues from the three branches of the precursor structure. At least three different alpha-mannosidases in the Golgi apparatus are involved in this process. These enzymes and the two glucosidases are hydrolases like lysosomal glycosidases but their activities are

[1] Actually, it is more appropriate to refer to “one cistron-one polypep tide”. This is no longer strictly accurate either, as more than one gene may contribute to the primary structure of a protein, i.e., immunoglobulins.

greatest at neutral pH, in contrast with lysosomal enzymes that have their greatest catalytic activity at an acid pH.

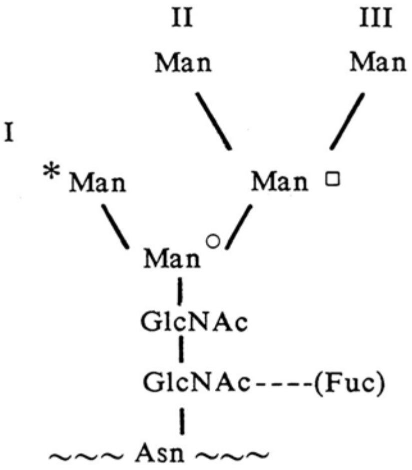


FIGURE 8-1

Intermediate partially processed asparagine—linked carbohy drate chain of a glycoprotein.

In eukaryotic cells, the high mannose oligosaccharide with five mannose units (See Figure 8-1) is the direct precursor of complex and hybrid structures. The initial step in the Golgi apparatus is the addition of an N-acetylglucosamine residue to the last remaining Man on branch I (*), after which the remaining two Man residues on branches II and III can be removed by alpha-mannosidases that are almost certainly different from those involved in earlier steps. Additional branches may be made at this point to produce tri- and tetra-antennary structures, and the final stages of processing are carried out by the addition of galactose, N-acetylglucosamine, sialic and fucose residues to give mature, complex, N-linked chains. An interesting N-acetylglucosaminyltransferase may add a beta-1,4-linked GlcNAc residue to the branched beta-linked mannose residue of the inner core region (0) to give a "bisected structure." This step has been the subject of intensive study by Carver and coworkers, who have been interested in the structural specificity of the enzyme with different conformations of the precursor oligosaccharides (Carver and Brisson, 1984).

It is likely that predictive methods will be employed in studies of processing pathways and the extent of processing of oligosaccharide chains. If control arises from an enzyme specificity for a particular three-dimensional structure of the substrate, it may be possible to determine these preferences and, from predictions of the distributions of three-dimensional structures of the oligosaccharide attached to the glycoprotein substrate, predict how far the carbohydrate chain will be processed.

Lysosomal enzymes contain one or more phosphate groups on mannose residues of the high mannose type oligosaccharide chains. The mannose-6-phosphate groups are specific recognition markers that are involved in the transport of lysosomal enzymes from the Golgi apparatus or outside the cells into lysosomes. Two membrane-bound mannose-6-phosphate receptors have been discovered in the plasma membrane; at least one of them also resides in the Golgi membranes. Although their binding specificities have been probed in some detail, other aspects have not been determined: the nature of the interaction of the phosphorylated mannose residues with the receptors and the three-dimensional structures of the lysosomal enzyme-receptor complexes.

Another interesting aspect of lyosomal enzyme synthesis involves the determination of structural domains on the folded proteins recognized by the enzyme that initiates phosphorylation of mannose residues, which is an N-acetylglucosamine-phosphotransferase (GlcNAc-P transferase) in the Golgi apparatus. This is the mechanism by which only lysosomal enzyme proteins are selected for phosphorylation. It is especially important because one form of a genetic lysosomal storage disorder, called mucolipidosis II, results from a defect in the binding domain of the GlcNAc-P transferase for lysosomal enzyme proteins. Perhaps this problem can be solved only by computer modeling to predict the three-dimensional structures of both proteins.

Glycosphingolipids are synthesized in an analogous manner, except that ceramide serves the function served by dolichol for glycoproteins and transfer occurs directly from a sugar nucleotide to the acceptor glycolipid. Ceramide is an acceptor for either glucose (from UDP-Glc) or galactose (from UDP-Gal), giving glucosylceramide or galactosylceramide. These simple glycosphingolipids predominate in human plasma and the brain, respectively, and also serve as precursors for more complex glycosphingolipids. In most organs, including the brain, the major pathways involve conversion of glucosylceramide to lactosylceramide, Gal-beta-1,4-Glc-Cer. Lactosylceramide is the substrate for several glycosyltrans ferases, the products of which are the first intermediates in the synthesis of related glycosphingolipids that may be classified according to their general structural characteristics. More than 100 different glycosphingolipids have already been characterized, and new compounds are still being discovered. Although some of the glycosphingolipids may contain between 15 and 35 or more sugar residues, most of the commonly occurring types have between 4 and 10 residues in the oligosaccharide chain.

ANALYSIS OF PRIMARY AND TERTIARY STRUCTURE

A complete understanding of the interactions between carbohydrates and proteins (enzymes, lectins, antibodies, and cell surface receptors) will depend on the determination of accurate three-dimensional structures of both kinds of molecules. As was noted, the primary structures of the oligosaccharide chains of complex carbohydrates cannot be deduced from DNA sequences and so must be determined by chemical and spectroscopic analysis. Modern chromatographic methods of separation, along with mass spectrometry and nuclear magnetic resonance (NMR), allow us to carry out complete analysis of a primary structure on a one micromole sample. Still to be determined are composition; arrangement of sugar residues; ring size; positions of glycosidic linkages and their anomerity; and the location and the chemical nature of non-carbohydrate substituents such as lipids, sulfate, and phosphate groups.

Three-dimensional structures of carbohydrates represent the spatial arrangements of the individual sugar residues. Most commonly occurring mammalian complex carbohydrates consist of sugar residues that exist in the pyranose ring form, the most stable and rigid conformation of which are the chair forms. When two sugar residues are joined together covalently in a glycosidic linkage, they are free to rotate about the glycosidic oxygen atom between the two rings, and the resulting disaccharide can therefore assume a number of different conformations corresponding to the rotations about these two bonds. It is customary to designate the dihedral angles at the glycosidic linkage (See Figure 8-2) by the

Greek symbols phi (ϕ) and psi (ψ), where the initial conformation (ϕ) = 0°,ψ = 0°) is that conformer where the C-l—H-1 bond eclipses O—C`-X` and C-1—O eclipses C`-X`—H-X`.

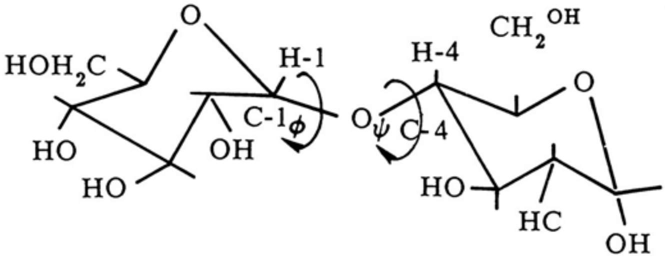


FIGURE 8-2 Dihedral angles determining the spatial relationship of two sugar residues in a disaccharide.

The relative orientations of adjacent sugar residues in an oligosaccharide chain are described by specifying the rotational angles (ϕ), ψ) at each glycosidic oxygen atom. When these angles are the same at each linkage, the chain has a helical conformation with n residues per turn and h unit translation along the helical axis. If n and h are available from x-ray data, then (ϕ) and ψ can be computed and vice versa. If (ϕ) and ψ are different among glycosidic linkages in an oligosaccharide chain, the three-dimensional structure becomes non-periodic and, for extreme variations, assumes a random coil conformation. Information about perturbations can be obtained by light-scattering, viscosity, sedimentation, and diffusion measurements.

X-RAY ANALYSIS OF CRYSTAL STRUCTURES OF CARBOHYDRATES

Of the three major classes of complex biological molecules, we have the least structural information at atomic resolution about carbohydrates. This is because they have not been crystallized, and consequently there is no relevant crystal structure data base other than that of the simple monomers to trimers upon which to model classical or semiempirical quantum mechanical calculations. The blood group-specific oligosaccharides, cord factors, and lipids A and X are typical examples. Exceptions are the cyclodextrins, which crystallize well, but are conformationally a separate class. Structures derived from the fiber-patterns of polysaccharides are model-dependent and do not constitute a source of definitive structural data. Stachyose, an oligosaccharide consisting of four sugar residues, is the largest noncyclic oligosaccharide for which there is a crystal structure analysis, but even in this case, the associated water structure has not been determined.

The crystallinity problem is only partially intrinsic. Carbohydrates do not solvate the same way as proteins, oligonucleotides, or nucleic acids. However, fewer efforts have been made to obtain the significant amounts of*configurationally homogeneous* material needed to conduct crystallization experiments than were made for proteins and nucleic acids. Another aspect of the crystallography of glycoconjugates is that the electron density for the oligosaccharide portion of glycoproteins has rarely been interpreted, even though several crystalline glycoproteins have been studied. This is because the standard refinement programs cannot handle the oligosaccharides, or there is microheterogeneity at the site of glycosylation, and so it is left out of the model. Thus, a potentially valuable source of information is not being exploited for lack of appropriate program development or strategic approaches to deal with microheterogeneity.

Steric considerations about the minimum approach distances between atoms, derived from observed nonbonded distances in various crystal structures, can be used to predict allowed conformations. This “hard sphere” approach, which was originally developed by V.S.R. Rao in the mid-1970s, is a rudimentary method of theoretical calculation that ignores electrostatic effects (hydrogen bonding), but does give a qualitative prediction of structure. This approach was subsequently extended by adapting energy calculations originally used for peptides, where the potential energy is divided into functions that describe discrete contributions such as van der Waals energies, electrostatic interactions, torsional energy, hydrogen bond energy, and bond and angle deformations (Bock, 1983). The data are presented in the form of computer-generated energy contour maps.

In much of the recent literature, conformational energy calculations have been made using a form of Rao's parameters with an added torsional potential about one of the glycosidic bonds (exoanomeric effect). This approach, which goes by the name HSEA (hard-sphere exoanomeric) method (Bock, 1983), has been used with success by Lemieux and Bock (1983), Carver and Brisson (1984), and others, although it contains a number of untested assumptions. The addition of a hydrogen bond potential (HEAH method) yields energy minimization results that differ from those calculated by the HSEA method, from which geometries can be derived that differ from those obtained by the HSEA method.

NMR SOLUTION STRUCTURES OF CARBOHYDRATES

Proton NMR methods provide detailed experimental data from which three-dimensional structures can be determined and compared with conformations arrived at by potential energy calculations. Carver and Cumming (1987) have generated contour maps of computed NOEs of various high mannose oligosaccharides as a function of the torsional angles φ and ψ. They then related them to experimental results as well as to minimum energy conformations estimated by various potential energy calculations (Carver and Cumming, in press). Brisson and Carver (1983) evaluated the utility of this approach using two biantennary complex type glycopeptides (See Figure 8-3). Since the NOE-derived conformations were within a range centered on the minimum energy conformations derived from potential energy calculations, it was concluded “that motional averaging is confined to a narrow range about one stable conformation” (Brisson and Carver, 1983). It now appears, however, that it is meaningless to seek a single NOE-derived conformation that satisfies a single potential energy minimum, because the molecules in fact may occupy such minima for a very small proportion of the time in solution. “Conformational flexibility must be incorporated into the theoretical treatment” (Carver and Cumming, 1987), and the calculation of energy surfaces becomes extremely important. The latest studies by Cumming and Carver indicate that NOE-determined three-dimensional structures may differ significantly from any minimum energy conformation. They have concluded from this that the NOE-derived conformations in such cases might correspond to “virtual” conformations as defined by Jardetzky (1980) to be computed structures that few if any molecules in solution actually adopt.

Scarsdale et al. (in press) have employed a molecular mechanics-based program in an effort to model conformational averaging of NMR data. Conformations were calculated using a combination of molecular potentials and NMR data for the oligosaccharide moiety of an erythrocyte glycolipid composed of three neutral sugars and an amino sugar. The lowest energy conformer closely resembled a structure proposed earlier. However, fits to data could be improved when two equilibrating conformers were considered. Thus, it may be possible to determine solution conformations of the complex carbohydrates, even in nonrigid cases, using a combination of calculations and constraints imposed from experimental NMR data.

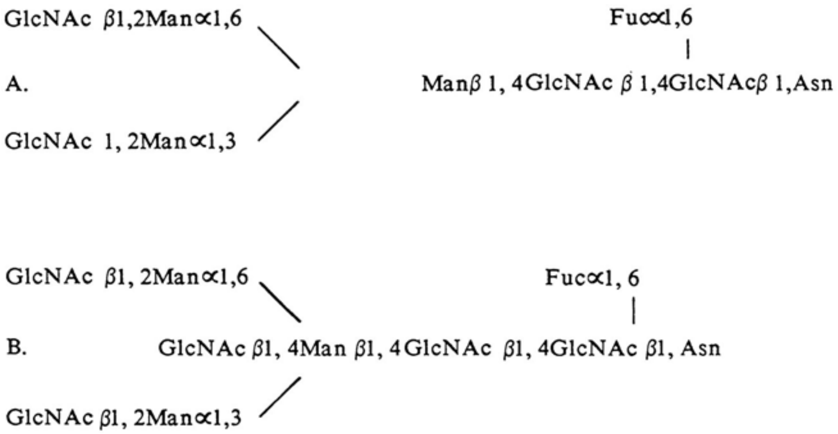


FIGURE 8-3 Structures of two partially processed asparagine—linked carbohydrate chains. The bisecting β1,4GlcNAc of B causes a conformational difference from that of A.

Despite the questions raised about the interpretation of NMR results and the value of potential energy minimizations, some important information has been collected about interactions of carbohydrate antigens with antibodies (Lemieux et al., 1985), oligosaccharides with lectins such as concanavalin A (Sekharudu et al., 1986), and oligosaccharides with glycosyltransferase enzymes (Carver and Cumming, 1987). Further refinements will depend upon the development of an agreed-on set of potential en ergy functions, which can be used with experimentally determined NOE-derived three-dimensional structures to evaluate whether a given molecule is distributed among several low energy conformations or occupies a particular subset of them. Tvaroska and Perez (1986) have recently compared several conformational energy calculations and proposed a general strategy for oligosaccharides.

Computer time and access to appropriate parallel processing array processors are important considerations in determining the level of support of research in this area at the present time. The availability of machines to calculate interatomic distances and van der Waals contributions extremely fast is a question that should be addressed by funding agencies. Interestingly, the several supercomputers currently operating on campuses have not been used to their capacity; perhaps efforts should be directed by appropriate advisory groups at these centers toward developing necessary software in these computers and establishing a policy that would direct a portion of their time for computer modeling of three-dimensional structures.

SUPRAMOLECULAR STRUCTURE

Structures that consist of more than one macromolecule interact as a unit in biological phenomena such as catalysis by many enzymes, binding at a cell surface, signal transduction across cell membranes, and other biological phenomena. Any enzyme that consists of more than one subunit should be thought of as a supramolecular structure. When large numbers of subunits are involved, and perhaps carry out more than one function, special consideration may have to be given to their relative spatial orientations. Examples are the replication of DNA by DNA polymerases, where complexes containing 10 or 12 proteins (called primosomes) are required to initiate replication. Ribosomes are even more complex, requiring at least 75 proteins to translate messenger RNA. Surfaces that consist of more than one macromolecule often behave as a functional unit. For example, the uptake of cholesterol by many cells requires the interaction of a specific cell surface receptor with a polypeptide surface of a complex supramolecular structure called low density lipoprotein (LDL), which consists of protein, cholesterol, phospholipids, and triacylglycerols. Alteration of the LDL protein by acetylation of a Lys residue blocks the binding of LDL to its receptor and uptake of cholesterol by the cell. Several hormones, including norepinephrine and epidermal growth factor (EGF), and other signals such as light (with rhodopsin) induce protein phosphorylation. EGF stimulates the growth of normal fibroblasts by binding to a specific transmembrane protein receptor on the cell surface. The hormone signal in this case is transduced by self-phosphorylation of the receptor on the intracellular side after the hormone binds, followed by other kinase-catalyzed phos phorylations of proteins, internalization of the EGF-EGF receptor complex, and a complex set of consequences in the nucleus and elsewhere in preparation for cell division. Bremer et al. (1986)

recently found that GM3 ganglioside inhibits this process in an allosteric fashion by preventing the self-phosphorylation of EGF receptor after EGF binding. To accomplish this, GM3 in the outer half of the cell membrane must interact with a domain of the polypeptide chain of EGF receptor, probably causing a conformational change that prevents phosphorylation. A similar situation involving a lipid membrane is found with a mitochondrial enzyme, beta-hydroxybutyric dehydrogenase, which is catalytically active only when incorporated into a lipid bilayer composed of certain phospholipids. Computer-assisted mathematical modeling of such supramolecular structures will be necessary to gain a deeper understanding of the organization of biological materials for complex functions.

9.

Hardware

Four functions are essential to computer modeling of molecules:

- molecular energy computation
- configurational control
- graphics
- reasoning

Until recently, the standard hardware configuration of a VAX and an Evans and Sutherland display terminal could only achieve the second and third items. Molecular energy calculation on a VAX is very slow, although these computers were used to develop the programs. The advent of Cray-type supercomputers connected by national communications networks has given scientists access to more computer power for molecular energy calculations. More recently, the development of special purpose array processors made it possible to have in the laboratory computational power roughly comparable to the supercomputers. Reasoning about molecular structure until recently could be done only with special purpose machines which run the programming language LISP.

As the power of computers available to individual scientists increases we expect that these four functions will be brought together. The early VAX computers (for example the 11/780) typically provide 0.5 megaflop (million floating point instructions per second) and 1.0 MIPS (million instructions per second). Typical array processors provide 100 megaflop while typical LISP machines provide 2.0 MIPS. In the last years it was necessary to have one each of these types of machines in order to have reasonable amounts of computational power for the four molecular modeling functions. The next generation of computer described as a personal supercomputer (PSC) will have between 40 and 60 megaflops of number crunching power and between 15 to 20 MIPS of general (i.e. logical) computational power. With this level of numeric and logical computational power available in the next year at a scientific workstation there will be little need for separate machines to perform special functions.

The national supercomputers, however, already in place and operational, constitutes a very real scientific resource. As scientists learn that the supercomputers can effectively carry out molecular energy calculations, these machines will be used to their fullest capacity. However, the technology of the supercomputers is advancing rapidly, and the manufacturers promise that systems with three orders of magnitude more computational power will be available in the next few years.

While the supercomputers grow more powerful, the power of workstations and the PSCs is also increasing. Current workstations have the power of VAXs, but lack the capacity to run all four functions simultaneously. As the PSCs emerge, they will offer a combination of capabilities that will make it possible to run all four functions at once. The PSCs should create the possibility of a new computational and graphic plateau:

1988 - 1995: personal supercomputer
1977 - 1987: E & S display coupled to a MicroVAX II
1970 - 1976: Tektronix display coupled to a DEC system-10.

The Tektronix display and a scientific mainframe gave us the first plateau seventeen years ago. On this plateau it was possible for many scientists to view and manipulate molecules. The VAX computers, and more recently the even less expensive MicroVAX II computers coupled to an Evans and Sutherland display, have established over the last ten years a plateau of graphic capability which has enabled scientists to go over from the physical modeling of macromolecules to completely electronic modeling. The PSCs expected to emerge in the next years will permit scientists to compute and to visualize molecules in much more powerful ways.

Using the PSCs, it should be possible to shape molecular models easily using joystick controls, creating stereo color graphics in multiple modes of representation, while doing energy calculations and molecular reasoning. The only foreseeable problem with the supercomputers is that scientists' appetites for energy calculations may exceed the computational capacities of the PSCs. Configurational control should make it possible to sketch protein models. Using collections of rules, we should be able to use molecular reasoning to generate and evaluate large numbers of possible model states.

Because of the rapidly changing technology of computers, displays, workstations, and PSCs, national effort should be directed to guaranteeing that these devices conform to the various levels of standards of the International Standards Organization (ISO).

Standardization in the United States is achieved by interested parties working together in committees under the auspices of agencies and organizations such as the National Bureau of Standards, American Society for Testing and Materials (ASTM), Institute of Electrical and Electronics Engineers (IEEE) or ISO. Considerable standardization at the level of the computer operating system must be done to make the ISO model work. Hardware vendors must choose between product uniqueness for sales and market development, and intervendor product compatibility. Compatibility has many benefits. Adherence to the standards will make it possible to move programs quickly and easily from one device to another, as well as making it possible to construct a complete system from components supplied by many vendors. The ISO model has several levels, represented below:

1. Ethernet
2. TCP/IP communications protocol
3. NFS - Network File System
4. UNIX operating system
5. VAX/VMS and Cray FORTRAN compatibility
6. X-windows
7. DIALOG-like application program window and functionality specification

The Ethernet originated at the XEROX Palo Alto Research Center. The TCP/IP protocol was developed for the DARPAnet, operated for the Department of Defense, and so is in the public domain. The NFS was developed by SUN Microsystems and placed in the public domain. Bell Laboratories developed UNIX. VAX/VMS FORTRAN was originated by Digital Equipment Corporation (DEC). X-windows originated at the Massachusetts Institute of Technology where they were developed to specify a machine-independent windowing system. DIALOG is an Apollo product that is a first attempt to answer the question of how to write high level mouse-driven applications programs in a high level specification language.

Standards are really the key to future progress in molecular modeling. If all investigators adhere to the ISO standards, then it will be possible to mix various workstations and special purpose computers on a laboratory network. Adherence to standards should lower the price of equipment to end users by enlarging the market. Similarly, with adherence to the standards, it will be possible to send and receive molecular structure data sets all over the world using global communications networks such as BITNET, CSnet, DARPAnet, Japan Universities net (JUnet), and Commonwealth Scientific and Industrial Research Organization net in Australia (CSIROnet).

Special purpose computers offer many possibilities for molecular modeling. Over the years, the National Institutes of Health (NIH) has funded facilities that developed molecular graphics, computation, and control devices. The control systems laboratory at Washington University Medical School developed the MMSX molecular display. The molecular graphics laboratory at the University of North Carolina at Chapel Hill has been instrumental in exploring the development of a variety of stereo, configurational control, and display devices. The molecular graphics laboratory at Columbia University is in the process of developing FASTRUN, a special purpose computer attached to a ST-100 array processor that boosts its molecular dynamics power by a factor of 10. The molecular graphics laboratory at the University of California at San Francisco Medical School has developed stereo and color representation techniques.

Special and general purpose graphics devices are increasingly easy to produce. General Electric in Research Triangle, North Carolina has produced a very fast surface graphics processor that can be used to display different types of objects, including molecules. At least one of the PSCs will have a sphere graphics primitive embedded in a silicon chip. Every effort should be made to encourage the development of special purpose processors. However, these processors should be required to adhere to the emerging computer standards, so that they can be easily integrated into existing laboratory networks.

The last few years have seen the emergence of array processors for laboratory use. The ST-100 array processor from Star Technologies, Inc. has been programmed by microcoding to produce molecular dynamics calculations at a rate comparable to a Cray XMP. The ST-100 is rated at peak 100 megaflops, while the sustained calculation rate is about 30 megaflops. The ST-100 costs about one-thirtieth of the Cray XMP-48. The FASTRUN device currently under development in the laboratory of Cyrus Levinthal at Columbia University will increase the power of the ST-100 by a factor of 10 from 30 average megaflops to 300 average megaflops. Floating Point Systems Inc. is discussing the delivery of a 10 processor FPS-264 system with a peak of 1 gigaflops. Multiple process machines could be added to this list, including the hypercube machines from Intel and NCUBE. All are laboratory machines. The power of supercomputers will obviously be increasing at the same approximate rates.

A very strong relationship exists between the architecture of a special purpose computer and the structure of the scientific problem to be solved. The question is, how much computational power does molecular modeling really need? The protein folding problem seems to be the gauge of this question, since molecular dynamics programs calculate atom position charge in 10^{-15} second time steps. If proteins really take minutes to fold, then computation will have to go from 10^{-15} to 102 seconds. The most powerful array processors available today make it possible to calculate and examine molecular trajectories three orders of magnitude longer than hitherto possible. Extending these trajectories an additional three orders of magnitude might bring us to the range where appropriate protein-folding actions can take place. There is some indication that if amino acids were synthesized at the rate of one per microsecond, then folding would be possible. Then, computing would only have to range from 10^{-15} to 10 $^{-5}$ seconds. This would be seven orders of magnitude less computing. If this estimate is close to correct and computing power increases at a rate of 50 percent per year, then current computer processor development will give us the necessary amount of power in 5 to 10 years.

CENTRAL VERSUS DISTRIBUTED COMPUTING

The National Science Foundation (NSF) supercomputer initiative again brings to the forefront the relationship between central computational services and distributed or personal services. Proponents of centralization argue that certain types of very large calculations are available only on centralized machines. The personal computer revolution showed how profoundly scientists respond to decentralized computation. The capabilities of personal machines increase at the same pace as the supercomputers, but the baseline machines are a market of 10^5 to 10^6 machines, whereas the supercomputers are a market of 10^2 to 10^3. Special purpose boards added to the baseline machine can raise its capabilities for specific functions (i.e., energy calculation, sequence comparison, or graphics) to levels approaching those of supercomputers.

The distribution of personal computation is driven totally by market forces and is not subject to centralized planning. Scientists buy laboratory computers with funds previously allocated for glassware. Postdoctoral students returning to their country of origin bring their personal computers. Floppy disks containing data files and even whole books form a new type of currency in countries operating centrally planned economies.

These modes of behavior form a valuable dichotomy. We need a balance between centralizing and decentralizing efforts. Individual scientists can participate in the planning and use of national supercomputers, while simultaneously helping to specify and buy smaller machines for their personal and laboratory use.

COMPUTER UTILIZATION IN THE NEXT 5 TO 10 YEARS

In the next 10 years, workstations will become ordinary scientific tools, like pocket calculators and balances. The workstations will become more popular with scientists as they acquire larger, faster, and more complex working programs; better graphics; more storage and access to other computers; and new data sources. A few years ago, only specialists searched DNA sequence data bases; now, because many workers have PCs in their laboratories, almost all molecular biologists search these data bases.

Workstation use is likely to follow the same pattern. Now, molecular graphics techniques are used only by departmental or laboratory specialists. In years to come, as all workstations begin to acquire adequate graphics capabilities, all scientists will routinely do molecular graphics, modeling, and energy calculations.

One of the strongest effects in the computer marketplace is the trade-off between constant dollar and constant performance. Because computer power is doubling every two to three years, the manufacturers tend to supply their customers with new models that cost the same but have increasing computational power. A customer, then, can expect to purchase a given level of computational power for a decreasing amount of money.

Twenty years ago, one needed a DEC PDP-10 to search protein or DNA data bases, while 10 years ago one used DEC PDP-11s or DEC VAXs. Now, one can use an IBM PC or one of its many clones to do the same job. In several years, one should be able to do DNA sequence searches on a pocket machine.

The brevity of the computer design and manufacture cycles has begun to overtake our ability to use these machines adequately. Twenty years ago, both manufacturers and consumers could reasonably expect a computer to sell and be worth buying for about 10 years; today, a given level of computational power has a life cycle of 3 years. The cycle length appears to be shortening even further in the sense that special purpose boards can be added to a small general purpose machine to make it functionally equivalent to a machine that costs up to 100 times as much. Why buy a Cray when a PC with a special purpose board will do the same thing? The cure for this problem will probably be a balance of market forces favoring the small mass distribution computers. PCs will rise in power to be general purpose workstations.

THE NATIONAL SUPERCOMPUTER NETWORK

The national supercomputer initiative sponsored by NSF allocates available computer time by a peer-review process. Individual scientist's requests for time must meet granting requirements of quality of the proposed work and size of allocation. From the scientist's viewpoint, the supercomputer network must perform tasks that cannot be done either in the laboratory or at local institutions. Since the network communication rates are 9,600 BAUD, only a limited amount of data can be passed between the scientist and the supercomputer. Essentially, this means that only batch computing can be run on the supercomputers. Large jobs run in the batch mode of computing are only one form of computing.

The highly interactive forms of computing and graphics available on workstations will be even more competitive with the supercomputer network when the next generation of high performance workstation, the PSC, becomes available.

The use of national supercomputers can be left to the discretion of individual scientists as it is in this country or the use of these resources can bemandated. The ability to mandate use depends on the type of the economy or pattern of interaction between scientists and the government. The Australian scientists are also in the midst of this type of central planning (personal communication, 1987, trip to Australia). The government wants scientists throughout Australia to use the centralized supercomputer by paying for the use with funds from the scientists' grants; the scientists see this as a form of taxation. The market forces in Australia will probably dominate when the scientists realize that superior computing and graphics performance can be obtained by purchasing a machine. Once a machine is in a department or laboratory, the problem of centralized national supercomputer access and allocation is essentially ended.

LOCAL AREA NETWORKS

Molecular modeling in the future will probably be done on local networks of computers and displays. For the past 5 to 10 years, advanced scientific laboratories have had one or more minicomputers. Five years ago, laboratory officials, for the most part, took the first hesitant steps to link these computers in a network. In the last two or three years, networking of laboratory computers has become much more common. Laboratory networks contain computers acting as hosts for terminal and computational servers for other workstations. The workstations range in power from the smallest PC to powerful PSCs. As computers age and are replaced because they no longer work or are too expensive to maintain, they will be replaced by networks of a variety of computers and displays.

DATA BASE USE

Access to molecular structure and sequence data bases through global communications networks is an opportunity that will be available in the near future. Currently, most data bases are updated by magnetic tape every three to six months, including the

DNA sequence data bases at the Los Alamos National Laboratory and at European Molecular Biology Laboratory (EMBL) in Heidelberg, the protein sequence data base at the National Biomedical Research Foundation (NBRF) in Washington, D.C., the protein structure data base at the Brookhaven National Laboratory, and the small organic molecule crystal structure data base at Cambridge University. Generating tapes for institutional and random scientific users is becoming an increasing burden for the data base operators. The global scientific networks are organized in such a way that it is possible for the data base operators to send out one copy of the update and have that copy spread throughout the entire scientific community.

For those scientific users who need a particular molecular structure data set for display or further modeling, the global scientific networks are ideal sources of information. Only recently, the Brookhaven protein structure file was tested at the National Research Council in Ottawa. A simple mail request to a BITNET server at the National Research Council produced one or more of the protein structure data sets in a few minutes.

The small molecule organic crystal structure file from Cambridge University in England is being used by scientists for molecular modeling and calculation. The Cambridge crystal file provides an ideal data source for ligand conformations. The data file and a search program have been available on the international commercial computer network for the past 15 years. Technology moves so fast that even while this report is being prepared the panorama with respect to data bases distribution has changed. For several years 5¼ inch laser disks have been on the market for audio. Now this highly developed consumer technology has been applied to the storage and retrieval of molecular structure data. Each laser disk, which costs about $2,000 to master and $10 to reproduce, can hold a complete update for the DNA sequence, protein sequence, protein structure and small molecule data files. The laser disk and associated software will be produced by a small starting company associated with the University of Wisconsin (Fred Blattner, DNAstar, Inc. at the University of Wisconsin, 1987, personal communication).

COMPETITIVENESS

America has a world recognized ability to transfer ideas from their development in an academic setting to practice by the formation of a small commercial enterprise. Then by the infusion of capital in several stages these small companies can be transformed into stable industrial corporations. These corporations are then able to consume the supply of trained scientific personnel produced by the universities. The position of the United States in the world economy is changing very dramatically at present, and certainly will continue to change in the next 5 to 10 years. Our overall competitiveness will be determined by our ability to form links between previously separate activities. It is already clear that biotechnology as an offshoot of our national expertise in molecular biology will be increasingly determined by the way we use computers in computational chemistry, macromolecular modeling, and the design of proteins. We are in the midst of two revolutionary tendencies: genetics and silicon. Computational chemistry is the glue that will bring these tendencies together in a stable form.

10.

Conclusions and Recommendations

CONCLUSIONS

Predicting macromolecular structure from fundamental chemical principles and information on primary structure is a challenging task. Understanding macromolecular function is even more demanding. Identifying important steps toward these goals is possible, however, and we have made considerable progress in various subtasks and specialized areas. There is every reason to believe that major breakthroughs can be expected over the next 10 years.

1. The tools of molecular mechanics and molecular dynamics have proved useful for exploring the conformational space of polypeptides, oligonucleotides, and oligosaccharides. In favorable cases, they identify the most stable conformers and quantitatively probe intermolecular interactions. Although these methods have not yet successfully predicted, a priori, the structures of molecules the size of small proteins, they play a major role in the refinement of experimentally-derived tertiary structures of macromolecules. Some promising results have been obtained in predicting the structural and thermodynamic consequences of local changes in amino acid sequences. Exciting new techniques make it possible to calculate free energies directly by perturbation methods. The techniques can be applied to intermolecular interactions or the changes

in free energy that accompany substitution of one amino acid for another and are readily applied to nucleic acid and polysaccharide problems as well.

2. The major limitations of current methods include:

   - the quality of the potential functions and of their parameters, especially the electrostatic terms;
   - methods for incorporating the solvent;
   - global search algorithms for solving the multiple-minima problem.

   Each of these areas has seen notable developments. While recently introduced procedures may produce solutions, we expect effective solutions to the multiple-minima problem to await new conceptual breakthroughs.

3. Heuristic modeling has been successful in the past, particularly in predicting the double helical structure of DNA, the alpha helix, and the beta-pleated sheet. When applied to globular proteins, this approach has yielded results which, although of relatively low resolution, have proved useful in guiding experiments in pursuit of more definitive data from crystallographic or nuclear magnetic resonance (NMR) techniques.
4. Experimental and theoretical methods can be usefully combined when the goal is to elucidate a new molecular structure based on a known one. When they are appropriate, modeling efforts based on the structural homology of one protein to another are currently the strongest line of attack.
5. Direct experimental approaches to macromolecular structure have been very successful; they cannot always be applied. They are limited by the need for significant quantities of highly purified material. Acquiring sufficient amounts of many interesting proteins, glycosylated proteins, and most nucleic acids is a challenging task. The powerful diffraction techniques all have an absolute requirement for crystals. NMR has molecular weight restrictions and some constraints on ultimate resolution. It takes at best months, and frequently a year or more to deduce a structure through crystallography or NMR.
6. Recent progress in instrumentation for crystallography has included the development of area detectors, which are only now

being fully utilized. Synchrotron sources and new neutron sources offer improved data. Isotopic labeling techniques and improved magnet technology signal new directions for NMR. We expect equally important breakthroughs in crystallization techniques.

7. Even with these advances, the most likely situation in the next decade is a substantial but essentially linear growth in the number of three-dimensional molecular structures elucidated by empirical methods. We estimate from current rates that several thousand protein and nucleic acid structures will be known in 10 years.
8. The explosive growth in the number of known nucleic acid sequences and hence protein primary sequences will continue to accelerate with or without implementation of the human genome project in the United States. Even at current rates, it is reasonable to expect 100,000 protein sequences to be described in the next decade. The overwhelming majority of new protein sequences are likely to be identifiable as members of known families of proteins.
9. Currently, the inventory of three-dimensional protein structural descriptions underrepresents the general distribution of protein families. The opportunities for computer-assisted modeling are enormous and will grow proportionately as more new structures and sequences are determined. Estimates of the number of sequences to be reported in the next decade suggest that existing facilities and resources for structural analysis will be overwhelmed by the avalanche of new sequence data.
10. Effects of covalent modification on structure and function of proteins, nucleic acids, and carbohydrates are diverse and poorly understood. No theoretical basis for predicting these effects exists in many cases. Describing structural relationships and cooperative functional roles in supramolecular systems are embryonic research areas to which modeling methods will contribute. Substantial attention will be directed toward these areas in the coming decade.
11. Computer speed, availability, and storage capacity are important limitations on the types of modeling calculations that can be attempted. Existing equipment is frequently incapable of performing all the necessary control experiments and refining major approximations. A 10-fold increase in computer performance

capability is required for conducting many current projects of biological importance systematically and rigorously. A minimum of a 100-fold improvement is needed for exploring new time scales or studying molecules of greater structural complexity than small proteins. We expect supercomputers, specialized hardware, and personal supercomputers (PSCs) to be significantly more available in the next few years. Most promising is the development during the next decade of high-capacity parallel processors.

12. A national computer network, operating at high speed and linking major government, academic, and industrial research facilities, will be crucial to molecular computation in the coming years. The uses of the network include transmission of sequence and structural data as well as access to computational facilities.
13. Of immense applied potential is the design of ligands to interact preferentially with macromolecular receptors, and the design of receptors to cause alterations of structure and/or function. These programs are in the earliest stages of development, and many hurdles must be overcome on the way from the laboratory to full clinical or commercial utility.
14. The intellectual, practical, and economic benefits of improved understanding of protein folding, macromolecular interactions, and macromolecular function are substantial.

RECOMMENDATIONS

1. The burgeoning volume of new sequence data requires a radical new policy on data banking of protein and nucleic acid sequences. A permanent national facility should be put in place as soon as possible, and considerable attention should be given to developing a data storage format that facilitates data retrieval. There should be no direct charges to the user. The initiation of this new national resource should be undertaken only after a round of detailed proposals has been sought and reviewed. A standing advisory committee of users should be appointed by a consortium drawn from the National Institutes of Health (NIH), National Science Foundation (NSF), and Department of Energy (DOE).
2. Whether the new facility should be allied with a national laboratory, such as Los Alamos, or with the National Library of

Medicine, or should be a completely new academic or commercial enterprise remains to be determined. Until the new unit is functioning, current facilities should be maintained to ensure an orderly transition.

3. Support for the archiving of coordinate and model-derived structures should continue. The Protein Data Bank at Brookhaven and the Cambridge Crystallographic File in England currently serve this need for the national and international community. Inclusion of data from new methods of structural analysis should be encouraged.
4. We recommend in the strongest terms expanding the supercomputer initiative, funding of computer networks, improving access by the scientific community to the existing supercomputer centers at the national laboratories, upgrading those centers, and providing individual research grants for purchasing PSCs. DOE should work closely with the supercomputer project managers at NSF to provide the broadest and most versatile computer network system on a national level. NIH should become more involved in direct support of scientific supercomputer centers.
5. Although the report does not specifically address this issue, the committee felt strongly that educational opportunities in structural biology and molecular modeling should be improved. Several mechanisms are available, such as expanding graduate programs through new training grants. We recommend that NSF and DOE increase graduate fellowship and postdoctoral fellow programs in this area. Workshops have been particularly effective for transferring information and skills. These include formal hands-on training programs in molecular dynamics and molecular graphics, and working meetings of independent investigators to address critical limiting aspects of a particular problem. Such workshops, which also promote crucial interdisciplinary approaches, could be funded by NIH, NSF, or DOE, acting together or independently.
6. Innovative and interdisciplinary research proposals in both theoretical and experimental aspects of structural biology should be directly encouraged through the use of existing funding mechanisms.
7. We see a special role for the national laboratories, which should interact at every level of these recommendations. The

> national laboratories should compete for the National Sequence Data Bank. The national laboratories and DOE have leadership status in the national computer network. They should increase efforts to make supercomputers available to the scientific community. Research efforts are going forward in molecular calculations and structural biology, with major programs at a few locations. Strengthening these efforts will assist the department's Office of Health and Environmental Research to assess the potential health and environmental effects of chemicals involved in energy processes.

Each of our recommendations involves developing some centralized activity. The issues in each area are quite different, however, and should not be taken as a general call for more biotechnology centers.

References

Abarbanel, R. 1984 . Protein Structural Knowledge Engineering . Ph.D. Thesis , University of California , San Francisco . 422 pp.

Abarbanel, R. 1986 . Pattern matching applications in protein secondary structure . Pp. 18-27 in Artificial Intelligence and its Impacts in Biology and Medicine . Proceedings of the I. A. Biomed. September 1986 .

Adelman, S. A. 1982 . Generalized Langevin models and condensed-phase chemical reaction dynamics . J. Phys. Chem. 86 : 1511-1524 .

Alber, T. , S. Dao-Pin , J. A. Nye , D. C. Muchmore , and B. W. Matthews . 1987 . Temperature-sensitive mutations of Bacteriophage T4 Lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein . Biochemistry 26:3754-3758 .

Altman, R. B. , and O. Jardetzky . 1986 . New strategies for the determination of macromolecular structure in solution . J. Biochem. 100:1403-1423 .

Anfinsen, C. B. , and H. A. Scheraga . 1975 . Experimental and theoretical aspects of protein folding . Adv. Protein Chem. 29:205-300 .

Anfinsen, C. B. , E. Haber , M. Sela , and F. H. White, Jr. 1961 . The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain . Proc. Natl. Acad. Sci. USA 47:1309-1314 .

Anonymous . 1986 . Renin inhibitors show early promise as antihypertensive agents . Chem. Eng. News . 64:23-24 .

Arseniev, A. S. , V. I. Kondakov , V. N. Maiorov , and V. F. Bystrov . 1984 . NMR solution spatial structure of 'short' scorpion insectoxin I_5A . FEBS Lett. 165:57-62 .

Barker, W. C. , L. T. Hunt , D. G. George , L. S. Yeh , H. R. Chen , M. C.Blomquist , I. Seibel-Ross , A. Elzanowski , M. K. Hong , D. A. Ferrick , J. K. Bair , S. L. Chen , and R. S. Ledley . 1986 . Protein sequence database, protein identification resource . Natl. Biom. Res. Found . Release 11.0 , Washington, D.C.

Bash, P. , U. C. Singh , R. Langridge , and P. A. Kollman . 1987 . Free energy calculations by computer simulation . Science 236:564-568 .

Bash, P. A . , U. C. Singh , S. K. Brown , R. Langridge , and P. A. Kollman . 1987 . Calculation of the relative change in binding free-energy of a protein-inhibitor complex . Science 235:574-576 .

Beddell, C. R. 1984 . Designing drugs to fit a macromolecular receptor . Chem. Soc. Rev. 13:279-319 .

Bennett, W. S. , and R. Huber . 1984 . Structural and functional aspects of domain motions in proteins . CRC Crit. Rev. Biochem. 15:291-386 .

Berendsen, H. J. C. , W. F. van Gunsteren , H. R. J. Zwinderman , and R. G. Geurtsen . 1986 . Simulations of proteins in water . Ann. N.Y. Acad. Sci. 482:269-286 .

Bernal, J. D. , and D. C. Crowfoot . 1934 . X-ray photographs of crystalline pepsin . Nature (London) 133:794 .

Beveridge, D. L. , and W. L. Jorgensen , eds. 1986 . Computer Simulation of Chemical and Biological Systems . Ann. N.Y. Acad. Sci. , Vol. 482 .

Blaney, J. M. , C. Hansch , C. Silipo , A. Vittoria . 1984 . Structure-activity relationships of dihydrofolate reductase inhibitors . Chem. Rev. 84:333-407 .

Blaney, J. M. , P. K. Weiner , A. Dearing , P. A. Kollman , E. C. Jorgensen , S. J. Oatley , J. M. Burridge , C. C. F. Blake . 1982 . Molecular mechanics simulation of protein-ligand interactions: Binding of thyroid hormone analogues to prealbumin . J. Am. Chem. Soc. 104:6424-6434 .

Bock, K. 1983 . The preferred conformation of oligosaccharides in solution inferred from high resolution NMR data and hard sphere exo-anomeric calculations . Pure Appl. Chem. 55:605-622 .

Boger, J. 1986 . Renin inhibitors: Drug design and molecular modelling . Third SCR-RSC Medicinal Chemistry Symposium . 271-292 .

Brady, J. , and M. Karplus . 1985 . Configuration entropy of the alanine dipeptide in vacuum and solution: A molecular dynamics study . J. Am. Chem. Soc. 107:6103-6105 .

Braun, W. , C. Bosch , L. R. Brown , N. Gō and K. Wüthrich . 1981 . Combined use of proton-proton overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations: Application to micelle-bound glycagon . Biochim. Biophys. Acta 667(2):377-396 .

Braun, W. , G. Wagner , E. Worgotter , M. Vasak , J. H. R. Kagi , and K. Wüthrich . 1986 . Polypeptide fold in the two metal clusters of metallothionein-2 by nuclear magnetic resonance in solution . J. Mol. Biol. 187:125-129 .

Bremer, E. G. , J. Schlessinger , and S. Hakomori . 1986 . Ganglioside-mediated modulation of cell growth . J. Biol. Chem. 261:2434-2440 .

Brisson, J. R. , and J. P. Carver . 1983 . Solution conformation of α-D(1-3)-linked and α-D(1-6)-linked oligomannosides using proton nuclear magnetic-resonance . Biochemistry 22:1362-1368 .

Brooks, C. L. III , and M. Karplus . 1986 . Theoretical approaches to solvation of biopolymers . Methods Enzymol . 127:369-400 .

Browne, W. J. , A. C. T. North , D. C. Phillips , K. Brew , T. C. Vanaman , and R. L. Hill . 1969 . A possible three-dimensional structure of bovine α-Lactalbumin based on that of hen's egg-white lysosyme . J. Mol. Biol. 42:65-86 .

Brunger, A. , R. L. Campbell , G. M. Clore , A. M. Gronenborn , M. Karplus , G. A. Petsko , and M. M. Teeter . 1986a . Solution of a protein crystal structure with a model obtained from NMR interproton distance restraints . Science 235:1049-1053 .

Brunger, A. T. , G. M. Clore , A. M. Gronenborn , and M. Karplus . 1986b . Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraint: Application to crambin . Proc. Natl. Acad. Sci. USA 83:3801-3805 .

Brunger, A. T. , J. Kuriyan , and M. Karplus . 1987 . Crystallographic R Factor Refinement by Molecular Dynamics . Science 235:458-460 .

Bush, C. A. , Z.-Y. Yan , and B. N. N. Rao . 1986 . Conformational energy calculations and proton Nuclear Overhauser Enhancements reveal a unique conformation for blood group A oligosaccharides . J. Am. Chem. Soc. 108:6168-6173 .

Carver, J. P. , and J. R. Brisson . 1984 . The Three-Dimensional Structure of N-Linked Oligosaccharides . Pp.209-331 in V. Ginsburg and P. W. Robbins , eds. Biology of Carbohydrates . Vol. 2 . John Wiley & Sons , New York.

Carver, J. P. , and D. A. Cumming . 1987 . Site-directed processing of N-linked oligosaccharides. The role of three-dimensional structure. Proc. Intl. Symp. Carbohydr. Chem. , Ithaca , New York .

Case, D. A. , and A. J. McCammon . 1986 . Pp. 222-233 in D. L. Beveridge and W. L. Jorgensen , eds. Computer Simulation of Chemical and Biological Systems . Ann. N.Y. Acad. Sci. , Vol. 482 .

Cech, T. R. 1987 . The chemistry of self-splicing RNA and RNA enzymes . Science 236:1532-1537 .

Cech, T. R. , N. K. Tanner , I. Tinoco , Jr. , B. R. Weir , M. Zuker , and P. S. Perlman . 1983 . Secondary structure of the*Tetrahymena* ribosomal RNA intervening sequence: Structural homology with fungal mitochondrial intervening sequences . Proc. Natl. Acad. Sci. USA 80:3903-3907 .

Chou, P. Y. , and G. D. Fasman . 1974 . Prediction of protein conformation . Biochemistry 13:222-245 .

Clore, G. M. , and A. M. Gronenborn . 1983 . Theory of the time dependent transferred nuclear overhauser effect: Applications to structural analysis of ligand-protein complexes in solution . J. Magn. Reson . 53:423-442 .

Cohen, F. E. , R. M. Abarbanel , I. D. Kuntz , and R. J. Fletterick . 1983 . Secondary structure assignment for*α/β* proteins by a combinatorial approach . Biochemistry 22:4894-4904 .

Cohen, F. E. , R. M. Abarbanel , I. D. Kuntz , and R. J. Fletterick . 1986a . Turn prediction in proteins using a pattern-matching approach . Biochemistry 25: 266-275 .

Cohen, F. E. , P. A. Kosen , I. D. Kuntz , L. B. Epstein , T. L. Ciardelli , and K. A. Smith . 1986b . Structure-activity studies of Interleukin-2 . Science 234:349-352 .

Corongiu, G. , and E. Clementi . 1981 . Simulations of the solvent structure for macromolecules II. Structure of water solvating Na^+-B-DNA at 300K and a model for conformational transitions induced by solvent variations . Biopolymers 20:2427-2483 .

Crippen, G. M. 1984 . Conformational analysis by scaled energy embedding . J. Comput. Chem. 5:548-554 .

Dabrowski, J. , U. Dabrowski , P. Hanfland , M. Kordowicz and W. E. Hull . 1986 . Structure determination of peracetylated glycosphingolipids by one-and two-dimensional^{1}H NMR at 360 and 500 MHz . Magn. Reson. Chem. 23:59-69 .

Dayhoff, M. O. 1978 . Atlas of Protein Sequence and Structure . Vol. 5,suppl.3 . National Biomedical Research Foundation , Washington, D.C.

Diamond, R. 1966 . A mathematical model-building procedure for proteins . Acta Crystallograph . A21:253-266 .

Dixon, R. A. F. , B. K. Kobilka , D. J. Strader , J. L. Benovic , H. G. Dohlman , T. Frielle , M. A. Bolanowski , C. D. Bennett , E. Rands , R.E. Diehl , R. A. Mumford , E. E. Slater , I. S. Sigal , M. G. Caron , R. J. Lefkowitz , and C. D. Strader . 1986 . Cloning of the gene and cDNA for mammalian B-adrenergic receptor and homology with rhodopsin . Nature (London) 321:75-79 .

DOE (U.S. Department of Energy) . 1987 . Summary report of a workshop on a carbohydrate structure data Base . DOE/ER-0310 . Washington, D.C.

Doolittle, R. F. 1985 . The genealogy of some recently evolved vertebrate proteins . Trends in Biochemical Sciences 10:233-237 .

Drew, H. R. , and R. E. Dickerson . 1981 . Structure of a B-DNA dodecamer III. Geometry of hydration . J. Mol. Biol. 151:535-556 .

Eisenberg, D. , R. M. Weiss , and T. C. Terwilliger . 1984 . The hydrophobic moment detects periodic periodicity in protein hydrophobicity . Proc. Natl. Acad. Sci. USA 81:140-144 .

Elber, R. , and M. Karplus . 1987 . Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin . Science 235:318-321 .

Ermak, D. L. , and J. A. McCammon . 1978 . Brownian dynamics with hydrodynamic interactions . J. Chem. Phys. 69:1352-1360 .

Ernst, R. , G. Bodenhausen , and A. Wokaun . 1987 . Principles of Nuclear Magnetic Resonance in One and Two Dimensions . Clarendon Press , Oxford . 610 pp .

Feldmann, R. J. , D. H. Bing , M. Potter , C. Mainhart , B. Furie , B. C. Furie , L. H. Caporale . 1985 . On the construction of computer models of proteins by the extension of crystallographic structures . Ann. N.Y. Acad. Sci. 439:12-43 .

Finer-Moore, J. , and R. M. Stroud . 1984 . Amphipathic analysis and possible formation of the ion channel in an acetylcholine receptor . Proc. Natl. Acad. Sci. USA 81:155-159 .

Fitzwater, S. , and H. A. Scheraga . 1982 . Combined-information protein structure refinement: Potential energy-constrained real-space method for refinement with limited diffraction data . Proc. Natl. Acad. Sci. USA 79:2133-2137 .

Freier, S. M. , R. Kiersek , J. A. Jaeger , N. Sugimoto , M. H Caruthers , T. Neilson , and D. H. Turner . 1986 . Improved free energy parameters for prediction of RNA duplex stability . Proc. Natl. Acad. Sci. USA 83:9373-9377 .

Friedman, H. L. , C. V. Krishnan , and C. Jolicoeur . 1973 . Ionic interactions in water . Ann. N.Y. Acad. Sci. 204:79-99 .

Fujinaga, M. , A. R. Sielecki , R. J. Read , W. Ardelt , M. Laskowski, Jr. , and M. N. G. James . 1987 . Crystal and molecular structures of the complex of α Chymotrypsin with its inhibitor Turkey Ovomucoid third domain at 1.8 Å resolution . J. Mol. Biol. 195:397-418

Ghose, A. K. , and G. M. Crippen . 1985 . Geometrically feasible binding modes of a flexible ligand molecule at the receptor site . J. Comput. Chem. 6:350-359 .

Ghosh, I. , and J. A. McCammon . 1987 . Sidechain rotational isomerization of proteins. Dynamics simulation with solvent surroundings . Biophys. J. 51:637-641 .

Gibson, K. D. , and H. A. Scheraga . 1967 . Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease S-peptide . Proc. Natl. Acad. Sci. USA . 58:420-427 .

Gibson, K. D. , and H. A. Scheraga . 1988 . The multiple-minima problem in protein folding . In M. H. Sarma and R. H. Sarma , eds. Structure and Expression: Vol. 1. From Proteins to Ribosomes . Adenine Press , Guilderland, N.Y.

Gilson, M. , A. Rashin , R. Fine , and B. Honig . 1985 . On the calculation of electrostatic interactions in proteins . J. Mol. Biol. 184:503-516 .

Gō, M. 1981 . Correlation of DNA exonic regions with protein structural units in haemoglobin . Nature (London) 291:90-92 .

Gō, M. , and H. A. Scheraga . 1984 . Molecular theory of the helix-coil transition in polyamino acids. V. Explanation of the different conformational behavior of valine, isoleucine, and leucine in aqueous solution . Biopolymers 23:1961-1977 .

Goodford, P. J. 1984 . Drug design by the method of receptor fit . J. Med. Chem. 27:557-564 .

Goodford, P. J. 1985 . A computational procedure for determining energetically favored binding sites on biologically important macromolecules . J. Med. Chem. 28:849-857 .

Green, D. W. , V. M. Ingram , and M. F. Perutz . 1954 . The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method . Proc. Roy. Soc. A 225:287 .

Greer, J. 1985 . Protein structure and function by comparative model building . Ann. N.Y. Acad. Sci. 439:44-63 .

Hagler, A. T. , J. Moult , and D. J. Osguthorpe . 1980a . Monte Carlo simulation of the solvent structure in crystals of a hydrated cyclic peptide . Biopolymers 19:395-418 .

Hagler, A. T. , D. J. Osguthorpe , and B. Robson . 1980b . Monte Carlo simulation of water behavior around the dipeptide N-Acetylalanyl-Methylamide . Science 208:599-601 .

Hakomori, S. 1986 . Glycosphingolipids . Sci. Amer. 254(5):44-53 .

Hare, D. R. , and B. R. Reid . 1986 . Three dimensional structure of a DNA hairpin in solution: Two-dimensional NMR studies and distance geometry calculations on d(CGCGTTTTCGCG) . Biochemistry 25:5341-5350 .

Harrison, S. C. , A. J. Olson , C. E. Schutt , F. K. Winkler , and G. Brigogne . 1978 . Tomato bushy stunt virus at 2.9 Å resolution . Nature (London) 276:368-373 .

Havel, T. F. , G. M. Crippen , and I. D. Kuntz . 1979 . Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions . Biopolymers 18:73-81 .

Havel, T. F. , and K. Wüthrich . 1985 . An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformation in solution . J. Mol. Biol. 182:281-294 .

Hayes-Roth, B. , B. Buchanan , O. Lichtarge , M. Hewett , R. Altman , J. Brinkley , C. Cornelius , B. Duncan , O. Jardetzky . 1986 . Protein: Deriving protein structure from constraints . Pp. 904-909 in Proceedings of AAAI-86, Fifth National Congress on Artificial Intelligence Vol. 2 . Morgan Kaufman , Los Altos, CA.

Henderson, R. 1979 . The structure of Bacterorhodopsin and its relevance to other membrane proteins . Soc. Gen. Physiol. Ser. 33:3-15 .

Hendrickson, W. A. , and J. H. Konnert . Incorporation of stereochemical information into crystallographic refinement . 1980 . Pp. 13.01-13.26 in Diamond, R. , S. Rameseshan , and K. Venkatesan , eds. Computing in Crystallography . Indian Academy of Sciences , Bangalore, India .

Hendrickson, W. A. 1985 . Stereochemically restrained refinement of macromolecular structures . Pp. 252-270 in H. W. Wyckoff , C. W. Hirs , and S. N. Timasheff , eds. Methods in Enzymology 115 . Academic Press , New York .

Henry, E. R. , M. Levitt , and W. A. Eaton . 1985 . Molecular dynamics simulation of photodissociation of carbon monoxide from hemoglobin . Proc. Natl. Acad. Sci. USA 82:2034-2038 .

Hermans, J. , ed. 1985 . Molecular Dynamics and Protein Structure . Proceedings of a workshop held 13-18 May 1984 at the University of North Carolina . Polycrystal Book Service , Western Springs, Ill. 194 pp.

Hermans, J. and S. Shankar . 1987 . The free energy of xenon binding to myoglobin from molecular dynamics simulation . Isr. J. Chem 27:225-227 .

Hesselink, F. T. , T. Ooi , and H. A. Scheraga . 1973 . Conformation energy calculations. Thermodynamic parameters of the helix-coil transition for poly(L-lysine) in aqueous salt solution . Macromolecules 6:541-552 .

Hingerty, B. , and S. Broyde . 1982 . Conformation of the deosydinucleoside monophosphate dCpdG modified at carbon 8 of guanine with 2-(acetylamino)fluorene . Biochemistry 21:3243-3252 .

Hodes, Z. I. , G. Némethy , and H. A. Scheraga . 1979a . Model for the conformational analysis of hydrated peptides. Effect of hydration on the conformational stability of the terminally blocked residues of the 20 naturally occurring amino acids . Biopolymers 18:1565-1610 .

Hodes, Z. I. , G. Némethy , and H. A. Scheraga . 1979b . Influence of hydration on the conformational stability and formation of bends in terminally blocked dipeptides . Biopolymers 18:1611-1634 .

Hogle, J. , M. Chow , and D. Filman . 1985 . Three-dimensional structure of poliovirus at 2.0 Å resolution . Science 229:1359 .

Hol, W. G. 1986 . Protein crystallography and computer graphics — Toward rational drug design . Angew. Chem., Int. Ed. Engl. 25:767-778 .

Homans, S. W. , R. A. Dwek , and T. W. Rademacher . 1987 . Solution conformations of N-linked oligosaccharides . Biochemistry 26:6571-6578 .

Hopfinger, A. J. 1985 . Computer-assisted drug design . J. Med. Chem. 28:1133-1139 .

Humber, L. G. , A. H. Philips , F. T. Bruderlein , M. Gotz , K. Voith . 1979. Mapping the dopamine receptor: Some primary and accessory binding sites . Pp. 227-242 in E. C. Olson and R. E. Christofferson , eds. 1979 . Computer-assisted Drug Design . ACS Symposium Series 112 , American Chemical Society , Washington, D.C.

Humblet, C. , and G. R. Marshall . 1981 . Three-dimensional computer modeling as an aid to drug design . Drug Dev. Res. 1:409-434 .

Hwang, J. K. , and A. Warshel . 1987 . Semiquantitative calculations of catalytic free energies in genetically modified enzymes . Biochemistry 26:2669-2670 .

Ingolia, T. D. , and E. A. Craig . 1982 . Four small drosophila heat shock proteins are related to each other and to mammalian alpha crystallin . Proc. Natl. Acad. Sci. USA 79:2360-2364 .

Jack, A. , and M. Levitt . 1978 . Refinement of large structures by simultaneous minimization of energy and R factor . Acta Crystallogr . A34:931 .

James, M. N. G. , L. T. J. Delbaere , and G. D. Brayer . 1978 . Amino acid sequence alignment of bacterial and mammalian pancreatic serine proteases based on topological equivalences . Can. J. Biochem. 56:396-402 .

Jardetzky, O. 1980 . Nature of molecular-conformations inferred from high-resolution NMR , Biochim. Biophys. Acta 621:227-232 .

Jones, T. A. 1985 . Interactive computer graphics: FRODO . Pp. 157-170 in Wyckoff, H. W. , C. W. Hirs , and S. N. Timashoff , eds. Methods in Enzymology 115 . Academic Press , New York .

Jorgensen, W. L. 1982 . Monte Carlo simulation of n-butane in water. Conformational evidence for the hydrophobic effect . J. Chem. Phys. 77:5757-5765 .

Jorgensen, W. L. In press . Energy profiles for organic reactions in solution . R. D. Levine , J. Jortner , and S. A. Rice , eds. Adv. Chem. Phys. , special volume. Evolution of Size Effects in Chemical Dynamics . John Wiley & Sons , New York .

Jorgensen, W. L. , J. Chandrasekhar , J. D. Madura , R. W. Impey , and M. L. Klein . 1983 . Comparison of simple potential functions for simulating liquid water . J. Chem. Phys. 79:926-935 .

Jorgensen, W. L. , and C. Ravimohan . 1985 . Monte Carlo simulation of differences in free energies of hydration . J. Chem. Phys. 83:3050-3054 .

Kabsch, W. , and C. Sander . 1983 . How good are predictions of protein secondary structure? FEBS Lett. 155:179-182 .

Kainosho, M. , and T. Tsuji . 1982 . Assignment of the three methionyl carbonyl carbon resonances in*Streptomyces* subtillsin inhibitor by a Carbon-13 and Nitrogen-15 double-labeling technique. A new strategy for structural studies of proteins in solution . Biochemistry 21:6273-6279 .

Kainosho, M. , H. Nagao , and T. Tsuji . 1987 . Local structural features around the c-terminal segment of*Streptomyces* subtilisin inhibitor studied by carbonyl carbon nuclear magnetic resonances of three phenylalanyl residues . Biochemistry 26:1068-1075 .

Kaiser, E. T. , and F. J. Kézdy . 1984 . Amphiphilic secondary structure: Design of peptide hormones . Science 223:249-255 .

Kang, Y. K. , G. Némethy , and H. A. Scheraga . 1987 . Free energies of hydration of solute molecules . J. Phys. Chem. 91:4105-4120 .

Kaptein, R. , E. R. P. Zuiderweg , R. M. Scheek , R. Boelens , and W. F. van Gunsteren . 1985 . A protein structure from nuclear magnetic resonance data . Lac repressor headpiece . J. Mol. Biol. 182:179-182 .

Karplus, M. , and J. A. McCammon . 1983 . Dynamics of proteins: Elements and function . Ann. Rev. Biochem. 52:263-300 .

Kauzmann, W. 1959 . Some factors in the interpretation of protein denaturation . Advan. Prot. Chem . 14:1-64 .

Kendrew, J. C. , R. E. Dickerson , B. E. Strandberg , R. G. Hart , D. R. Davies , D. C. Phillips , and V. C. Shore . 1960 . Structure of Myoglobin . Nature (London) 185:422-427 .

Klapper, I. , R. Hagstrom , R. Fine , K. Sharp , and B. Honig . 1986 . Focussing of electric fields in the active site of Cu-Zn superoxide dismutose: Effects of ionic strength and amino-acid modification . Proteins 1:47-59 .

Klein, B. J. , and G. R. Pack . 1983 . Calculations of the spatial distribution of charge density in the environment of DNA . Biopolymers 22:2331-2352 .

Klein, T. E. , C. Huang , T. E. Ferrin , R. Langridge , C. Hansch . 1986 . Computer-assisted drug receptor mapping analysis . Pp. 147-158 in T. H. Pierce , and B. A. Mohne , eds. Artificial Intelligence Applications in Chemistry . ACS Symposium Series 306 . American Chemical Society , Washington, D.C.

Kopka, M. L. , P. Pjura , C. Yoon , D. Goodsell , R. E. Dickerson . 1985a . The bigning of netropsin to double-helical B-DNA of sequence C-G-C-G-A-A-T-T-^{Br}C-G-C-G: Single crystal x-ray structure analysis . Pp. 461-483 in E. Clementi , G. Corongiu , M. H. Sarma , and R. Sarma , eds. Structure and Motion: Membranes, Nucleic Acids and Proteins . Adenine Press , New York .

Kopka, M. L. , C. Yoon , D. Goodsell , P. Pjura , R. E. Dickerson . 1985b . The molecular origin of DNA-drug specificity in netropsin and distamycin . Proc. Nat. Acad. Sci. USA 82:1376-1380 .

Kopka, M. L. , C. Yoon , D. Goodsell , P. Pjura , R. E. Dickerson . 1985c . The binding of an antitumor drug to DNA; Netropsin and C-G-C-G-A-A-T-T-^{Br}C-G-C-G . J. Mol. Biol. 183:553-563 .

Kosen, P. A. , R. M. Scheek , H. Naderi , V. J. Basus , S. Manogaran , P. G. Schmidt , N. J. Oppenheimer , and I. D. Kuntz . 1986 . Two-dimensional^{1}H NMR of three spin-labeled derivatives of BPTI . Biochemistry 25:2356-2364 .

Krüger, P. , W. Strassburger , A. Wollmer , and W. F. van Gunsteren . 1985 . A comparison of the structure and dynamics of avian pancreatic polypeptide hormone in solution and in the crystal . Eur. Biophys. J. 13:73-88 .

Kubo, T. , K. Fukuda , A. Mikami , A. Maeda , H. Takahashi , M. Mishina , T. Haga , K. Haga , A. Ichiyama , K. Kangawa , M. Kojima , H. Matsuo , T. Hirose , and S. Numa . 1986 . Cloning, sequencing and expression of complementary DNA encoding the muscarinic acetylcholine receptor . Nature (London) 323:411-416 .

Kuntz, I. D. 1972 . Protein Folding . J. Am. Chem. Soc. 94:4009-4012 .

Kuntz, I. D. 1975 . Approach to the tertiary structure of globular proteins . J. Am. Chem. Soc. 97:4362-4366 .

Kuntz, I. D. , J. M. Blaney , S. J. Oatley , R. Langridge , and T. Ferrin . 1982 . A geometric approach to macromolecule-ligand interactions . J. Mol. Biol. 161:269-288

Kyte, J. , and R. F. Doolittle . 1982 . A simple method for displaying the hydropathic character of a protein . J. Mol. Biol. 157:105-132 .

Laue, E. D. , J. Skilling , J. Staunton . 1985 . Maximum entropy reconstruction of spectra containing antiphase peaks . J. Magn. Reson. 63:418-424 .

LeMaster, D. M. , and F. M. Richards . 1985 . ^{1}H–^{15}N Heteronuclear NMR studies of E. coli thioredoxin in samples isotopically labeled by residue type . Biochemistry 24:7263-7268 .

Lemieux, R. U. , and K. Bock . 1983 . The conformational analysis of oligosaccharides by H-NMR and HSEA calculation . Arch. Biochem. Biophys. 221:125-134 .

Lemieux, R. U. , A. P. Venot , U. Spohr , P. Bird , G. Mandal , N. Morishima , O. Hindsgaul , and D. R. Bundle . 1985 . Molecular recognition. V. The binding of the human B blood group determinant by hybridoma monoclonal antibodies . Can. J. Chem. 63:2664-2668 .

Levinthal, C. 1966 . Molecular model-building by computer . Sci. Am. 214(6):42-52

Levitt, M. 1969 . Detailed molecular model for transfer ribonucleic acid . Nature (London) 224:759-763 .

Levitt, M. 1982 . Protein conformation, dynamics and folding by computer simulation . Annu. Rev. Biophys. Bioeng . 11:251-271 .

Levitt, M. , and C. Chothia . 1976 . Structural patterns in globular proteins . Nature (London) 261:552-558 .

Levy, R. M. , M. Karplus , and J. A. McCammon . 1979 . Diffusive Langevin dynamics of model alkanes . Chem. Phys. Lett. 65:4-11 .

Levy, R. M. , M. Karplus , and P. G. Wolynes . 1981 . NMR relaxation parameters in molecules with internal motion: Exact Langevin trajectory results compared with simplified relaxation models . J. Am. Chem. Soc. 103:5998-6011 .

Lewis, P. N. , F. A. Momany , and H. A. Scheraga . 1971 . Folding of polypeptide chains in proteins: A proposed mechanism for folding . Proc. Natl. Acad. Sci. USA 68:2293-2297 .

Lipman, D. , and W. Pearson . 1985 . Rapid and sensitive protein similarity searches . Science 227:1435-1441 .

Lybrand, T. , J. A. McCammon , and G. Wipff . 1986 . Theoretical calculation of relative binding affinity in host-guest systems . Proc. Natl. Acad. Sci. USA 83:833-835 .

Manning, G. S. 1978 . The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides . Q. Rev. Biophys. 11:179-246 .

Markley, J. L. 1987 . One-and two-dimensional NMR investigations of the consequences of amino acid replacements in proteins . Pp. 15-33 in D. L. Oxender , and C. F. Fox , eds. Protein Engineering . A. R. Liss , Inc. , New York .

Marshall, G. R. , C. D. Barry , H. E. Bosshard , R. A. Dammkoehler , D. A. Dunn . 1979 . The conformational parameter in drug design: The active analog approach . Pp. 205-226 in E. C. Olson , and R. E. Christofferson , eds. Computer-assisted Drug Design . ACS Symposium Series 112 , American Chemical Society , Washington, D.C.

Martin, Y. C. 1981 . A practitioner's perspective of the role of quantitative structure activity analysis in medicinal chemistry . J. Med. Chem. 24:229-237 .

McCammon, J. A. , S. H. Northrup , M. Karplus , and R. M. Levy . 1980 . Helix-coil transitions in a sample polypeptide model . Biopolymers 19:2033-2045 .

McIntosh, L. P. , R. H. Griffey , D. G. Muchmore , C. P. Nielson , A. G. Redfield , and F. W. Dahlquist . 1987 . Protein NMR measurements of bacteriophage T4 lysozyme aided by^{15}N isotopic labeling: Structural and dynamic studies of large proteins . Proc. Natl. Acad. Sci. USA 84:1244-1248 .

Mezei, M. , P. K. Mehrotra , and D. L. Beveridge . 1985 . Monte Carlo determiniation of the free energy and internal energy of hydration for the ala dipeptide at 25°C . J. Am. Chem. Soc. 107:2239-2245 .

Michel, H. 1982 . Three-dimensional crystals of a membrane protein complex. The photosynthetic reaction centre from*Rhodopseudomonas viridis* . J. Mol. Biol. 158:567-572 .

Miller, M. H. , and H. A. Scheraga . 1976 . Calculation of the structure of collagen models. Role of inter-chain interactions in determining the triple-helical coiled-coil conformation. I. Poly (Glycyl-Prolyl-Prolyl) . J. Polym. Sci. Polym. Syrup . 54:171-200 .

Mishina, M. , T. Kurosaki , T. Tobimatsu , Y. Morimoro , M. Noda , T. Yamamoio , M. Terao , J. Lindstrom , T. Takahashi , M. Kuno , and S. Numa . 1984 . Expression of functional acetylcholine receptor from cloned cDNAs . Nature (London) 307:604-613 .

Moult, J. , and M. N. G. James . 1986 . An algorithm for determining the conformation of polypeptide segments in proteins by systematic search . Proteins 1:146-163 .

Murthy, C. S. , R. Bacquet , and P. J. Rossky . 1985 . Ionic distribution near polyelectrolytes. A comparison of theoretical approaches . J. Phys. Chem. 89:701-710 .

Némethy, G. , and H. A. Scheraga . 1977 . Protein folding . Q. Rev. Biophys. 10:239-352 .

Némethy, G. Z. I. Hodes , and H. A. Scheraga . 1978 . A model for hydration of peptides and its application to the conformational analysis of terminally blocked amino acids and dipeptides . Proc. Natl. Acad. Sci. USA 75:5760-5764 .

Neumann, M. 1985 . The dielectric constant of water. Computer simulation with MCY potential . 82:5663-5672 .

Neumann, M. 1986 . Dielectric relaxation in water. Computer simulations with the TIP4P potential . J. Chem. Phys. 85:1567-1580 .

Nilsson, L. , G. M. Clore , A. M. Gronenborn , A. T. Brunger , and M. Karplus . 1986 . Structure refinement of oligonucleotides by molecular dynamics with NOE interproton distance restraints: Application to 5' d(CGTACG)$_2$" . J. Mol. Biol. 188:455-475 .

Nirenberg, M. , P. Leder , M. Bernfield , R. Brimacombe , J. Trupin , F. Rottman , C. O'Neal . 1965 . RNA codewords and protein synthesis, VII. On the general nature of the RNA code . Proc. Natl. Acad. Sci. USA 53:1161-1168 .

Nishikawa, K. 1983 . Assessment of secondary-structure prediction of proteins. Comparison of computerized Chou-Fasman method with others . Biochim. Biophys. Acta 748:285-299 .

Noguti, T. , and N. Gō . 1985 . Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins . Biopolymers 24:527-546 .

Okuyama, K. , N. Tanaka , T. Ashida , and M. Kakudo . 1976 . Structure analysis of a collagen model polypeptide, (Pro-Pro-Gly),$_{10}$. Bull . Chem. Soc. Jpn. 49:1805-1810 .

Olson, E. C. , and R. E. Christoffersen , eds. 1979 . Computer-assisted Drug Design . ACS Symposium Series 112 , American Chemical Society , Washington, D.C. 619 pp.

Paine, G. H. , and H. A. Scheraga . 1987 . Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. III. Probable and average conformations of enkephalin . Biopolymers 26:1125-1162 .

Palmer, K. A. , H. A. Scheraga , J. F. Riordan , and B. L. Vallee . 1986 . A preliminary three-dimensional structure of angiogenin . Proc. Natl. Acad. Sci. USA 83:1965-1969 .

Pauling, L. , R. B. Corey , and H. R. Branson . 1951 . The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain . Proc. Natl. Acad. Sci. USA 37:205-211 .

Perutz, M. F. 1965 Structure and function of hemoglobin. I. A tentative atomic model of horse oxyhemoglobin . J. Mol. Biol. 13:646-668 .

Perutz, M. F. , G. Fermi , D. J. Abraham , C. Poyart , and E. Bursaux . 1986 . Hemoglobin as a receptor of drugs and peptides: X-ray studies of the tereochemistry of binding . J. Am. Chem. Soc. 108:1064-1078 .

Perutz, M. F. , J. C. Kendrew , and H. C. Watson . 1965 . Structure and function of hemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence . J. Mol. Biol. 13 : 669-678 .

Perutz, M. F. , M. G. Rossman , A. F. Cullis , H. Muirhead , G. Will , and A. C. T. North . 1960 . Structure of hemoglobin . Nature (London) 185 : 416-421 .

Petrillo, E. W. , and M. A. Ondetti . 1982 . Angiotensin-converting enzyme inhibitors: Medicinal chemistry and biological activity . Med. Res. Revs. 2: 1-41 .

Petsko, G. A. , and D. Ringe . 1984 . Fluctuations in protein structure from x-ray diffraction . Annu. Rev. Biophys. Bioeng. 13: 331-371 .

Pettitt, B. M. , and M. Karplus . 1985 . The potential of mean force surface for the alanine dipeptide in aqueous solution: A theoretical approach . Chem. Phys. Lett. 121: 194-201 .

Pettitt, B. M. , and P. J. Rossky . 1986 . Alkali halides in water: Ion-solvent correlations and ion-ion potentials of mean force at infinite dilution . J. Chem. Phys. 84:5836-5844 .

Pettitt, B. M. , M. Karplus , and P. J. Rossky . 1986 . Integral equation model for aqueous solvation of polyatomic solutes: Application to the determination of the free energy surface for the internal motion of biomolecules . J. Phys. Chem. 90:6335-6345 .

Pfandler, P. , and G. Bodenhausen . 1986 . Automated analysis of two-dimensional NMR spectra of mixtures by pattern recognition . J. Magn. Reson . 70:71-78 .

Phillips, D. C. 1967 . Lysozyme and the development of protein crystal chemistry . Proc. Int. Cong. Biochem. (Tokyo) 7:63-82 .

Pincus, M. R. , and H. A. Scheraga . 1979 . Conformational energy calculations of enzyme-substrate and enzyme-inhibitor complexes of lysozyme. 2. Calculation of the structures of complexes with a flexible enzyme . Macromolecules 12:633-644 .

Plattner, J. J. , J. Greer , A. K. L. Fung , H. Stein , H. D. Kleinert , H. L. Sham , J. R. Smital , and T. J. Perun . 1986 . Peptide analogs of angeniotensinogen. Effect of peptide chain length on renin inhibition . Biochem. Biophys. Res. Commun. 139: 982-990 .

Postma , J. P. M. , H. J. C. Berendsen , and J. R. Haak . 1982 . Thermodynamics of cavity formation in water: A molecular dynamics study . Faraday Symp. Chem. Soc. 17:55-67 .

Rao, N. R. , U. C. Singh , P. A. Bash , and P. A. Kollman . 1987 . Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin . Nature (London) 328:551-553 .

Rao, S. T. , and M. G. Rossmann . 1973 . Comparison of super-secondary structures in proteins . J. Mol. Biol. 76:241-256 .

Richards, F. M. , 1968 . The matching of physical models to three dimensional electron density maps: A simple optical device . J. Mol. Biol. 37:225-230 .

Richards, F. M. 1986 . Protein Design: Are We Ready? Pp. 171-196 in D. L. Oxender , ed. Protein Structure, Folding and Design . Alan R. Liss , New York .

Richardson, J. S. 1981 . The anatomy and taxonomy of protein structure . Pp. 168-340 in C. B. Anfinsen , J. T. Edsall , and F. M. Richards , eds. Advances in Protein Chemistry . Vol. 34. Academic Press , New York .

Robson, B. and D. J. Osguthorpe . 1979 . Refined models for computer simulation of protein folding . J. Mol. Biol. 132:19-51 .

Robson, B. 1986 . The prediction of peptide and protein structure . Pp. 567-607 in A. Darbre , ed. Practical Protein Chemistry - A Handbook . John Wiley & Sons , New York .

Rose, G. D. 1985 . Automated recognition of domains in globular proteins . Methods Enzymol . 115:430-440 .

Rose, G. D. , L. M. Gierasch , and J. A. Smith . 1985 . Turns in peptides and proteins . Pp. 1-109 in Advances in Protein Chemistry . Vol. 37 . Academic Press , New York .

Rosenberg, R. O. , M. Rao , and B. J. Berne . 1982 . Hydrophobic effect on chain folding. The trans to gauche isomerization of n-Butane in water . J. Am. Chem. Soc. 104:7647-7649 .

Rossmann, M. G. , ed. 1972 . The Molecular Replacement Method . Gordon & Breach Publishers , New York .

Rossmann, M. G. , and P. Argos . 1977 . The taxonomy of protein structure . J. Mol. Biol. 109:99-129 .

Rossmann, M. G. , E. Arnold , J. W. Erickson , J. E. Johnson , G. Kamers , M. Luo , A. G. Mosser , R. R. Rueckert , B. Sherry , and G. Friend . 1985 . Structure of a human common cold virus and functional relationship to other picornaviruses . Nature (London) 317:145-153 .

Scarsdale, J. N. , P. Ram , and J. H. Prestegard . In press . A molecular mechanics NMR pseudoenergy approach to the solution conformation of glycolipids . J. Comp. Chem.

Scheraga, H. A. 1984 . Protein structure and function from a collodial to a molecular view . Carlsberg Rev. Commun . 49:1-55 .

Schiffer, J. , and A. B. Edmundson . 1967 . Use of helical wheels to represent the structures of proteins and to identify segments with helical potential . Biophys. J. 7:121-135 .

Schussheim, A. E. , and D. Cowburn . 1987 . Deconvolution of high resolution 2D-NMR signals by digital signal processing with linear predictive singular value decomposition . J. Magn. Reson . 71:371-378 .

Seibel, G. L. , U. C. Singh , and P. A. Kollman . 1985 . A molecular dynamics simulation of double helical B-DNA including counterions and water . Proc. Natl. Acad. Sci. USA .

Sekharudu, Y. C. , M. Biswas , and V. S. R. Rao . 1986 . Complex carbohydrates. 2 . The modes of binding of complex carbohydrates to concanavalin A - A computer modeling approach . Int. J. Biol. Macromol . 8:9-19 .

Sela, M. , F. H. White , Jr. , and C. B. Anfinsen . 1957 . Reductive cleavage of disulfide bridges in ribonuclease . Science 125:691 .

Sheriff, S. , E. W. Silverton , E. A. Padlan , G. H. Cohen , S. J. Smith-Gill , B. C. Finzel , and D. R. Davies . In press . Three-dimensional structure of an antibody antigen complex . Proc. Natl. Acad. Sci. USA .

Singh, U. C. , F. Brown , P. A. Bush , and P. A. Kollman . 1987 . An approach to the application of free energy perturbation methods using molecular dynamics: Applications to the transformation of $CH_3OH \rightarrow CH_3CH_3$, $H_3O^+ \rightarrow NH_4^+$, Gly→Ala, Ala→Phe in Aqueous Solution and to $H_3O^+ (H_2O)_3 \rightarrow NH_4^+(H_2O)_3$ in the gas phase . J. Am. Chem. Soc. 109:1607-1614 .

Smith, J. L. , W. A. Hendrickson , R. B. Honzatko , and S. Sheriff . 1986a . Structural heterogeneity in protein crystals . Biochemistry 25:5018-5027 .

Smith, T. J. , M. J. Kremer , M. Luo , G. Vriend , E. Arnold , G. Damer , M. G. Rossmann , M. A. McKinlay , G. D. Diana , and M. J. Otto . 1986b . The site of attachment in human rhinovirus 14 for antiviral agents that inhibit uncoating . Science 233:1286-1293 .

Smith-Gill , S. J. , J. A. Rupley , J. R. Princes , R. P. Ceerty , and H. A. Scheraga . 1984 . Experimental identification of a theoretically-predicted “left-sided” binding mode for $(Gl_cNA_c)_6$ in the active site of liponyml . Biochemistry 23:993-997 .

Snider , M. D. 1984 . Biosynthesis of glycoproteins . Formation of N-linkedoligosaccharides . Pp. 163-198 in V. Ginsburg and P. W. Robbins , eds. Biology of Carbohydrates . Vol. 2 . John Wiley & Sons , New York .

Soumpasis, D. 1984 . Statistical mechanics of the B→Z transition of DNA: Contribution of diffuse ionic interactions . Proc. Natl. Acad. Sci. USA 81:5116-5120 .

Steinhauser, O. 1983 . On the dielectric theory and computer simulation of water . Chem. Phys. 79:465-482 .

Stillinger, F. H. 1975 . Theory and molecular models for water . Adv. Chem. Phys. 31:1-101 .

Stroud, R. M. , and J. Finer-Moore . 1985 . Acetylcholine receptor structure, function, and evolution . Annu. Rev. Cell. Biol. 1:317-351 .

Sussman, J. L. 1985 . Constrained-restrained least squares refinement of proteins and nucleic acids . Pp. 271-302 in Wyckoff, H. W. , C. W. Hirs ,and S. N. Timasheff , eds. Methods in Enzymology , 115 . Academic Press , New York .

Suzuki, E. , N. Pattabiraman , G. Zon , and T. L. James . 1986 . Solution structure of [d(A-T)5]2 via complete relaxation matrix analysis of two dimensional NOE spectra and molecular mechanics calculations: Evidence for a hydration tunnel . Biochemistry 25:6854-6865 .

Swenson, M. K. , A. W. Burgess and H. A. Scheraga . 1978 . Conformational analysis of polypeptides: applications to homologous proteins . Pp. 115-142 in B. Paullaman , ed. Frontiers in Physical Chemical Biology . Academic Press , New York .

Taylor, W. R. , and J. M. Thornton . 1983 . Prediction of super-secondary structure in proteins . Nature (London) 301:540-542 .

Teleman, O. 1986 . Molecular Dynamics Simulation of Polyatomic Molecules in Aqueous Solution . Thesis . University of Lund , Sweden . 180 pp .

Tembe, B. L. , and A. McCammon . 1984 . Ligand-receptor interactions . Comp. Chem. 8:281-283 .

Tidor, B. , K. K. Irikura , B. R. Brooks , and M. Karplus . 1983 . Dynamics of DNA oligomers . J. Biomol. Struc. Dyn. 1:231-252 .

Tilton, R. F. , U. C. Singh , I. D. Kuntz , and P. A. Kollman . In press . Protein ligand dynamics: A 96 picosecond simulation of a myoglobinxenon complex . J. Mol. Biol.

Tinoco, I., Jr. , O. C. Uhlenbeck , and M. D. Levine . 1971 . Estimation of secondary structure in ribonucleic acids . Nature (London) 230:362-367 .

Tully, J. C. 1981 . Computer simulation of the dynamics of chemical processes . Comp. Chem. 5:159-165 .

Tvaroska, I. , and S. Perez . 1986 . Conformation energy calculations for oligosaccharides: A comparison of methods and a strategy of calculation . Carbohydr. Res. 149:389-410 .

Ultsch, M. H. , A. A. Kossiakoff , J. Burnier , J. A. Wells , D. P. Powers , B. A. Katz , R. R. Bolt , B. C. Cunningham , and S. S. Power . 1985 . Industrial applications of enzyme engineering . World Biotech . Rep. 1:611-616 .

Unwin, P. , and P. D. Ennis . 1984 . Two configurations of a channel-forming membrane protein . Nature (London) 307:609-612 .

van Gunsteren , W. F. , and H. J. C. Berendsen . 1984 . Computer simulations as a tool for tracing the conformational differences between proteins in solution and in the crystalline state . J. Mol. Biol. 176:559-564 .

van Gunsteren , W. F. , and M. Karplus . 1982 . Protein dynamics in solution and in a crystalline environment: A molecular dynamics study . Biochemistry 21:2259-2274 .

van Gunsteren , W. F. , H. J. C. Berendensen , J. Hermans , W. G. J. Hol , and J. P. M. Postma . 1983 . Computer simulation of the dynamics of hydrated protein crystals and its comparison with X-ray data . Proc. Natl. Acad. Sci. USA 80:4315-4319 .

van Gunsteren , W. F. , H. J. C. Berendsen , R. G. Geurtsen , and H. R. J. Zwinderman . 1986 . A molecular dynamics computer simulation of an eight base pair DNA fragment in aqueous solution: Comparison with experimental two-dimensional NMR data . Ann. N.Y. Acad. Sci. 482:287-303 .

Wagner, G. , and D. Bruhwiler . 1986 . Toward the complete assignment of the carbon NMR spectrum of the bovine pancreatic trypsin inhibitor . Biochemistry 25:5839-5843 .

Wallis, M. , S. L. Howell , and K. W. Taylor . 1985 . The Biochemisty of the Polypeptide Hormones . J. Wiley & Sons , New York .

Wand, A. J. , H. Roder , and S. W. Englander . 1986 . Two-dimensional^{1}H NMR studies of cytochrome c: Hydrogen exchange in the N-terminal helix . Biochemistry 25:1107-1114 .

Warme, P. K. , F. A. Momany , S. V. Rumball , R. W. Tuttle , and H. A. Scheraga . 1974 . Computation of Structures of Homologous Proteins. α-Lactalbumin from Lysozyme . Biochemistry 13:768-782 .

Warshel, A. 1981 . Electrostatic basis of structure-function correlation in proteins . Acc. Chem. Res. 9:284-290 .

Warshel, A. , and F. Sussman . 1986 . Toward computer-aided site-directed mutagenesis of enzymes . Proc. Natl. Acad. Sci. USA 83:3806-3810 .

Watson, J. D. , and F. H. C. Crick . 1953 . Molecular structure of nucleic acids . A structure for deosyribosenucleic acid . Nature (London) 171:737-738 .

Weiner, S. J. , P. A. Kollman , D. A. Case , U. C. Singh , C. Ghio , G. Alagona , S. Profeta, Jr. , and P. Weiner . 1984 . A new force field for molecular mechanical simulation of nucleic acids and proteins . J. Am. Chem. Soc. 106:765-784 .

Wetlaufer, D. B. 1973 . Nucleation , rapid folding , and globular intrachainregions in proteins . Proc. Natl. Acad. Sci. USA 70:697-701 .

Wilbur, W. J. , and D. J. Lipman . 1984 . The context of dependent compari son of biological sequences . SIAM J. Appl. Math. 44:557-567 .

Williams, A. L., Jr. , and I. Tinoco, Jr. 1986 . A dynamic programming algorithm for finding alternative RNA secondary structures . Nucleic Acids Res. 14:299-315 .

Williamson, M. P. , T. F. Havels and K. Wüthrich . 1985 . Solution conformation of proteinase inhibitor IIA from bull seminal plasma by^{1}H nuclear magnetic resonance and distance geomety . J. Mol. Biol. 182:295-315 .

Wise, M. , R. D. Cramer , D. Smiths and I. Exman. 1983 . Progress in three-dimensional drug design: The use of real-time colour graphics and computer postulation of bioactive molecules in DYLOMMS . Pp. 145-146 in J. C. Deardon , ed. Quantitative Approaches to Drug Design . Elsevier , Amsterdam .

Wong, C. F. , and J. A. McCammon . In press . Computer simulation and the design of new biological molecules . Isr. J. Chem.

Wüthrich, K. 1986 . NMR of Proteins and Nucleic Acids . John Wiley & Sons , New York . 292 pp .

Xuong, N. G. , S. T. Freer , R. Hamlin , C. Nielsen , and W. Vernon . 1978 . The electronic stationary picture method for high speed measurement of reflection intensities from crystals with large unit cell dimensions . Acta Crystallogr . A34:289-296 .

Yu, R. K. , T. A. W. Koerner , J. N. Scarsdale , and J. H. Prestegard . 1986 . Elucidation of glycolipid structure by proton nuclear magnetic resonance spectroscopy . Chem. Phys. Lipids 42:27-48 .

Zaug, A. J. , and T. R. Cech . 1986 . The intervening sequence of*Tetrahymena* is an enzyme . Science 231:470-475 .

Zichi, D. A. , and P. J. Rossky . 1986 . Molecular conformational equilibria in liquids . J. Chem. Phys. 84:1712-1723 .

Zuker, M. , and P. Steigler . 1981 . Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information . Nucleic Acids Res. 9:133-148 .