SIAM Rev.
Vol. 39, No. 3, pp. 407–460, September 1997

002

MOLECULAR MODELING OF PROTEINS AND MATHEMATICAL PREDICTION OF PROTEIN STRUCTURE*

ARNOLD NEUMAIER†

Abstract. This paper discusses the mathematical formulation of and solution attempts for the so-called protein folding problem. The static aspect is concerned with how to predict the folded (native, tertiary) structure of a protein given its sequence of amino acids. The dynamic aspect asks about the possible pathways to folding and unfolding, including the stability of the folded protein.

From a mathematical point of view, there are several main sides to the static problem:
– the selection of an appropriate potential energy function;
– the parameter identification by fitting to experimental data; and
– the global optimization of the potential.

The dynamic problem entails, in addition, the solution of (because of multiple time scales very stiff) ordinary or stochastic differential equations (molecular dynamics simulation) or (in case of constrained molecular dynamics) of differential-algebraic equations. A theme connecting the static and dynamic aspect is the determination and formation of secondary structure motifs.

The present paper gives a self-contained introduction to the necessary background from physics and chemistry and surveys some of the literature. It also discusses the various mathematical problems arising, some deficiencies of the current models and algorithms, and possible (past and future) attacks to arrive at solutions to the protein-folding problem.

Key words. protein folding, molecular mechanics, transition states, stochastic differential equations, dynamic energy minimization, harmonic approximation, multiple time scales, stiffness, differential-algebraic equation, molecular dynamics simulations, potential energy surface, parameter estimation, conformational entropy, secondary structure, tertiary structure, native structure, conformational entropy, global optimization, simulated annealing, genetic algorithm, smoothing method, diffusion equation method, branch and bound, backbone potential models, lattice models, contact potential, threading, solvation energy, combination rule

AMS subject classifications. Primary, 92C40; Secondary, 65L05, 90C26

PII. S0036144594278060

It is God's privilege to conceal things,
but the kings' pride is to research them.
(Proverbs 25:2; ascribed to King Solomon of Israel, ca. 1000 B.C.)

1. Introduction.

This paper is the result of my investigations into the problems involved in the mathematical prediction of (tertiary, three-dimensional) protein structure given the (primary, linear) structure defined by the sequence of amino acids of the protein. This so-called *protein folding problem* is one of the most challenging problems in current biochemistry and is a very rich source of interesting problems in mathematical modeling and numerical analysis, requiring an interplay of techniques in eigenvalue calculations, stiff differential equations, stochastic differential equations, local and global optimization, nonlinear least squares, multidimensional approximation of functions, design of experiment, and statistical classification of data. Even topological concepts like the Morse index (Mezey [204]) and invariants in knot theory (Jones polynomials) have been discussed in this context; see, e.g., Sumners [310]. An extensive recent report [217] from the U.S. National Research Council on the mathematical challenges from theoretical and computational chemistry shows the protein

*Received by the editors December 1, 1994; accepted for publication (in revised form) September 30, 1996.
http://www.siam.org/journals/sirev/39-3/27806.html
†Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria (neum@cma.univie.ac.at).

folding problem embedded into a large variety of other mathematical challenges in chemistry.

The aims of the present paper are to introduce mathematicians to the subject, to provide enough background that the problems in the mathematical modeling of proteins become transparent, to expose the merits and deficiencies of current models, to describe the numerical difficulties in structure prediction when a model is specified, and to point out possible ways of improving model formulation and prediction techniques.

Molecular biology is mankind's attempt to figure out how God engineered His greatest invention—life. As with all great inventions, details are top secret; however, even top secrets may become known. I find it a great privilege to live in a time where God allows us to gain some insight into His construction plans, only a short step away from giving us the power to control life processes genetically. I hope it will be to the benefit of mankind, and not to its destruction.

After the successful deciphering of the genetic code that defines how the amino acid sequences of proteins are coded in the DNA, one of the major missing steps in understanding the chemical basis of life is the protein folding problem—the task of understanding and predicting how the information coded in the amino acid sequence of proteins at the time of their formation translates into the three-dimensional structure of the biologically active protein. (Actually, there are also folding problems in connection with nucleotide sequences in DNA and RNA, but this survey is limited to protein folding only. For the mathematics of nucleic acids and genome analysis see, e.g., a recent U.S. National Research Council report by Lander and Waterman [177].)

Proteins are the machines and building blocks of living cells. If we compare a living body to our world, each cell corresponds to a town, and the proteins are the houses, bridges, cars, cranes, roads, airplanes, etc. There are huge numbers of different proteins, each one performing its specific task.

Since it is known already how to use genetic engineering to produce proteins with a given amino acid sequence, knowledge of how such a protein would fold would allow one to predict its chemical and biological properties. If we were able to solve the protein folding problem, it would greatly simplify the tasks of interpreting the data collected by the human genome project, understanding the mechanism of hereditary and infectious diseases, designing drugs with specific therapeutical properties (see, e.g., [12]), and of growing biological polymers with specific material properties.

The literature on the various aspects of protein folding is enormous, and I made no attempt to be complete in the coverage of papers; instead I simply quote the papers that I have found useful in the preparation of this study. Given the current amount of activity in this broad field and my own time limitations, it is probably inevitable that I also omitted some recent papers with new developments, and I'd appreciate being informed about any serious omissions.

However, I tried to draw a complete picture of the physical and chemical background needed to understand modeling details and to be able to read more specialized literature. To allow an assessment of the approximations made in the traditional modeling process, I also included (less complete) remarks and pointers to the literature regarding attempts at more detailed or more accurate modeling (e.g., quantum corrections) even if these are (in the near future) unlikely to be relevant to practical calculations with macromolecules. Thus the paper can also be viewed as a case study of mathematical modeling of a complex scientific problem.

For further information, I refer to the introductory paper Richards [251] in *Scientific American*; to the books by Brooks, Karplus, and Pettitt [35] and Creighton [62],

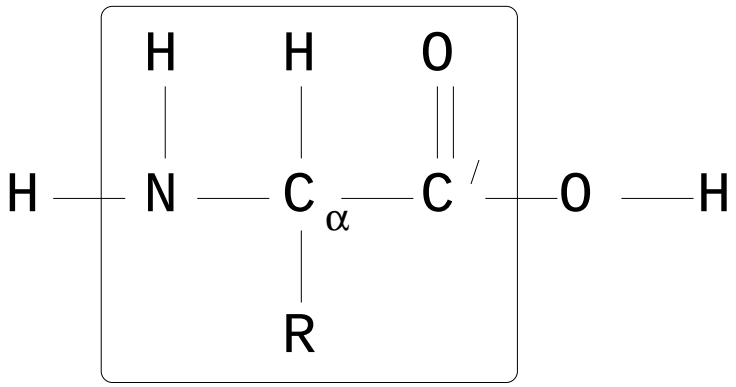

FIG. 1. *An amino acid with side chain R.*

which contain thorough treatments of the subject; to the *Reviews in Computational Chemistry* edited by Lipkowitz and Boyd [190] with many excellent articles on related topics; to the recent survey by Chan and Dill [49], which contains many additional pointers to the recent literature related to the physics, chemistry, and biology of protein folding; and to Pardalos, Shalloway, and Xue [228] for algorithmic aspects of the optimization problems associated with the problem. Two books providing a general background in computational chemistry are Clark [54] (an introductory overview with little theory) and, more oriented toward biological applications, Warshel [338].

Further useful books on the subject are [30, 35, 45, 110, 137, 252], and another survey, emphasizing the biological aspects, is Jaenicke [155].

More and more useful material becomes available electronically on the World Wide Web (WWW). At `http://solon.cma.univie.ac.at/~neum/protein.html`, there is a good, but necessarily biased and incomplete list of links that I have found useful while working on this study.

2. Proteins.

Chemical structure. From a purely chemical point of view, a *protein* is simply a polymer consisting of a long chain of amino acid residues. More precisely, polymers of this type are called *di-, tri-, oligo-,* or *polypeptides* if they consist of two, three, several, or many residues, respectively. Each *amino acid* (except proline) has the structure given in Fig. 1, where R stands for the *side chain* characteristic for specific amino acids.

The proteins in living cells contain 20 different residues, with side chains having 1–18 atoms. The residues are usually abbreviated with three identifying letters of the corresponding amino acid, giving the list

Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile,

Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val.

A set of (ECEPP-)geometries for these amino acids can be found, e.g., in Momany et al. [209]. For generalities on biochemical nomenclature see [154].

Under the influence of RNA containing the genetic information coding for the amino acid sequence, amino acids polymerize in a specific sequence to a chain with the structure as given in Fig. 2. Bonds joining two residues (called *peptide bonds*) hydrolyze (i.e., break under consumption of a water molecule) in a sufficiently acid

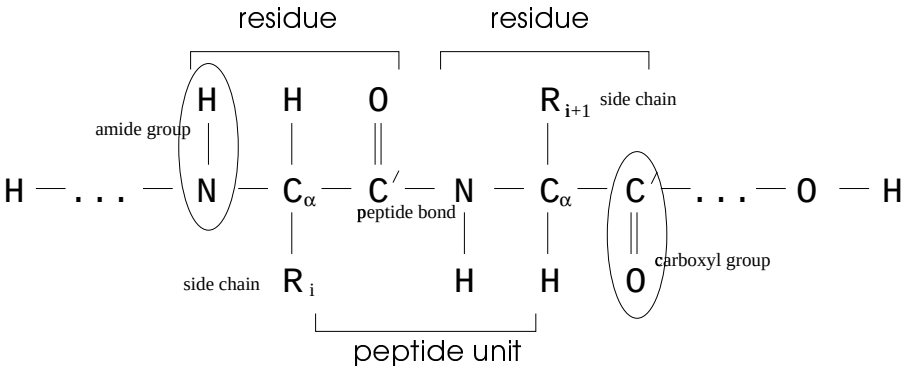

FIG. 2. *The chemical structure of a protein.*

environment, and this can be used to determine the precise sequence of residues in a given protein. Sometimes, the end groups of a protein, the NH_2 *amino group* and the COOH *carboxyl group*, are substituted by other groups; e.g., so-called *blocked* polypeptides have CH_3 *methyl groups* at both ends. Since amino acid residues are asymmetric, two distinct proteins correspond to a chain of residues and the chain in reversed order.

The repeating $-NC_\alpha C'-$ chain of a protein is called its *backbone*. Although looking linear in the diagram displaying the bond structure, interatomic forces bend and twist the chain in characteristic ways for each protein. They cause the protein molecule to curl up into a specific three-dimensional geometric configuration called the *folded state* of the protein. This configuration and the chemically active groups on the surface of the folded protein determine its biological function.

Consequently, biochemists are very keen in wanting to understand how the *primary structure* (the sequence of the residues) gives rise to the *tertiary structure* (the folded state). Intermediate between the two is the *secondary structure*, i.e., local systematic patterns or motifs like helices, recognizable in shorter pieces of many proteins. The *quaternary structure*, i.e., the pattern in which proteins crystallize, is less interesting from a biological point of view. (The naming reflects the fact that the primary structure, coded in the cell genome, is the basic information from which the synthesis of proteins in a cell proceeds. While folding, the secondary structure appears and is modified until the folded tertiary structure is established; the quaternary structure is the latest stage, if it is attained at all.)

The smallest proteins, hormones, have about 25–100 residues, typical globular proteins about 100–500; fibrous proteins may have more than 3000 residues. Thus the number of atoms involved ranges from somewhat less than 500 to more than 10,000. One of the smallest proteins, BPTI (bovine pancreatic trypsin inhibitor), with 58 residues and 580 atoms only, has become a well-studied model protein from both the computational and the experimental point of view; very accurate data for the crystal structure are available. Another small protein that has found considerable attention is Crambin (with 46 residues).

Local geometry. The geometry is captured mathematically by assigning to the ith atom a three-dimensional *coordinate vector*

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix}$$

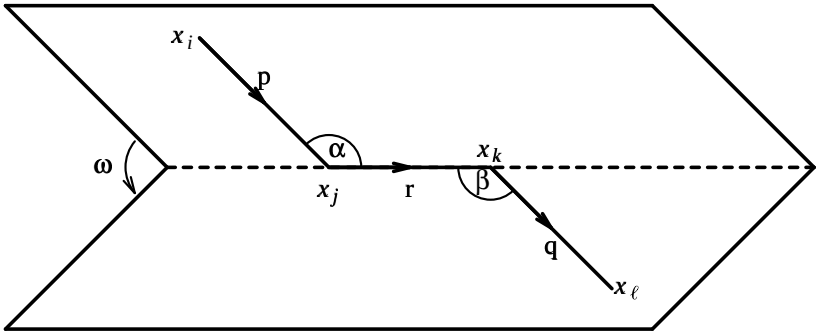

FIG. 3. *Bond vectors, bond angles, and the dihedral angle.*

specifying the position of the atom in space. If two atoms with labels j and k are joined by a chemical bond, we consider the corresponding *bond vector*

$$r = x_k - x_j,$$

with *bond length*

$$\|r\| = \sqrt{(r,r)},$$

where

$$(p,q) := p_1q_1 + p_2q_2 + p_3q_3$$

is the standard inner product in $\mathbb{R}^3$.

Similarly, for two adjacent bonds i-j and k-l, we have the bond vectors

$$p = x_j - x_i, \quad q = x_l - x_k.$$

The *bond angle* $\alpha = \sphericalangle(i\text{-}j\text{-}k)$, of Fig. 3, can then be computed from the formulas

$$\cos\alpha = \frac{(p,r)}{\|p\|\|r\|}, \quad \sin\alpha = \frac{\|p \times r\|}{\|p\|\|r\|}$$

(together with $\alpha \in [0°, 180°]$), where

$$p \times r = \begin{pmatrix} p_2r_3 - p_3r_2 \\ p_3r_1 - p_1r_3 \\ p_1r_2 - p_2r_1 \end{pmatrix}$$

is the cross product in $\mathbb{R}^3$. The bond angle $\beta = \sphericalangle(j\text{-}k\text{-}l)$ is similarly found from

$$\cos\beta = \frac{(q,r)}{\|q\|\|r\|}, \quad \sin\beta = \frac{\|q \times r\|}{\|q\|\|r\|}.$$

Finally, the *dihedral angle* $\omega = \sphericalangle(i\text{-}j\text{-}k\text{-}l) \in [-180°, 180°]$ (or the complementary *torsion angle* $180° - \omega$) measures the relative orientation of two adjacent angles in a chain i-j-k-l of atoms. It is defined as the angle between the normals through the planes determined by the atoms i, j, k and j, k, l, respectively, and can be calculated from

$$\cos\omega = \frac{(p \times r, r \times q)}{\|p \times r\|\|r \times q\|}, \quad \sin\omega = \frac{(q \times p, r)\|r\|}{\|p \times r\|\|r \times q\|}.$$

In particular, the sign of ω is given by that of the triple product $(q \times p, r)$.

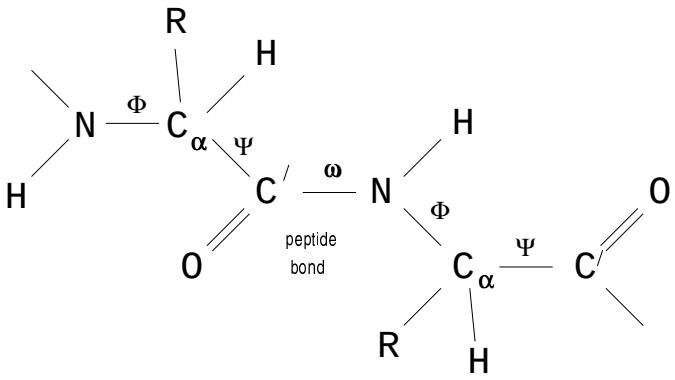

FIG. 4. *Backbone dihedral angles of a protein.*

A full set of bond lengths, bond angles, and dihedral angles already fixes the geometry of a molecule (and often overdetermines it). However, the geometry is quite sensitive to small changes in the angles, and, to reduce the sensitivity, it is also useful to specify a number of so-called *out-of-plane bending* or *improper torsion angles* $\omega = \sphericalangle(i\text{-}j\text{-}k\text{-}l)$, which are defined in a similar way for any tetrahedron formed by an atom k with three adjacent atoms i, j, l. Clearly, bond lengths, bond angles, dihedral angles, and improper torsion angles are invariant under translation, rotation, and path reversal. However, dihedral and improper torsion angles change sign under reflection; their signs therefore model the *chirality* (left- or right-handedness) of subconfigurations.

Under biological conditions, the bond lengths and bond angles are fairly rigid (with a standard deviation of less than $0.2\mathring{A}$ for bond lengths and of about 2° for bond angles [136]; recent experimental values are reported, e.g., in [90]). Therefore, the dihedral angles along the backbone (usually labeled as in Fig. 4) determine the main features of the final geometric shape of the folded protein.

Structural information for the proteins with known geometry is collected worldwide in the quickly growing Brookhaven Protein Data Bank [19], accessible through the WWW at `http://www.pdb.bnl.gov`; see also [337], [301].

3. Molecular mechanics. In this section I look at the physics governing the motion of the atoms in a protein (or any other molecule). To reduce the formal complexity of the discussion, the family of coordinate vectors x_i in 3-space is replaced by a single coordinate vector

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ \vdots \\ x_{N1} \\ x_{N2} \\ x_{N3} \end{pmatrix}$$

in a $3N$-dimensional *state space*, where N is the total number of atoms in the molecule. Since x contains three coordinates for each atom, we see that for a real protein, the dimension of x is in the range of about 1500–30,000.

The force balance within the molecule and the resulting dynamics can be approximated mathematically by means of the stochastic differential equation