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

1. Introduction.

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.)

This paper is the result of my investigations into the problems involved in the mathematical prediction of (tertiary, 3-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 [205]) and invariants in knot theory (Jones polynomials) have been discussed in this context; see, e.g., Sumners [311]. An extensive recent report [218] from the U.S. National Research Council on the mathematical challenges from theoretical and computational chemistry shows the protein 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

* Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria.
email: `neum@cma.univie.ac.at` WWW: `http://solon.cma.univie.ac.at/~neum/`

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 3-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 nuclein acids and genome analysis see, e.g., a recent U.S. National Research Council report by LANDER & WATERMAN [178].)

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 infectuous diseases, designing drugs with specific therapeutical properties (see, e.g., BALBES et al. [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 of being 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 one or the other recent paper 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 and to aid investigations in other molecular modeling problems not directly related to protein folding, I also included (less complete) remarks and pointers to the literature regarding attempts of more detailed or 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 in mathematical modeling of a complex scientific problem.

For further information, we refer to the introductory paper RICHARDS [252] in *Scientific American*, to the books by BROOKS et al. [34] and CREIGHTON [62], which contain thorough treatments of the subject, to the *Reviews in Computational Chem-*

istry edited by LIPKOWITZ & BOYD [191] with many excellent articles on related topics, to the recent survey of CHAN & DILL [48], which contains many additional pointers to the recent literature related to the physics, chemistry and biology of protein folding, and to PARDALOS et al. [229] for algorithmic aspects of the optimization problems associated with the problem. Two books providing a general background in computational chemistry are CLARK [53] (an introductory overview with little theory) and, more oriented towards biological applications, WARSHEL [339].

Further useful books on the subject are [30, 34, 44, 110, 138, 253], and another survey, emphasizing the biological aspects, is JAENICKE [156].

More and more, useful material becomes available electronically on the World Wide Web. At `http://solon.cma.univie.ac.at/ neum/protein.html`, there is a necessarily biased and incomplete list of links, collected while working on this study.

Acknowledgments. Much of the research necessary for writing this survey was done during a year I spent at AT&T Bell Laboratories, Murray Hill.

I'd like to thank David Gay and Margaret Wright for introducing me to the subject, Frank Stillinger for providing me with background literature and for patiently answering my questions, and Tamar Schlick for her many comments on a draft version of the paper and additional pointers to the literature.

Thanks also go to an anonymous referee who made many suggestions that widened the perspective of the survey so that it covers the entire field of protein structure prediction; and to Erich Bornberg-Bauer who interpreted the referee's condensed remarks by providing the literature relevant to meet his requests, and who produced most of the figures.

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 2, 3, several, or many residues, respectively. Each *amino acid* (except proline) has the structure given in Figure 1, where R stands for the *side chain* characteristic for specific amino acids.

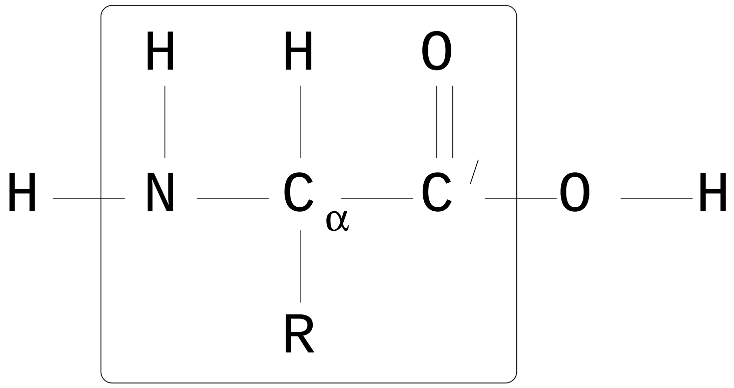

FIG. 1. **An amino acid with side chain** R

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. [210]. For generalities on biochemical nomenclature see [155].

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 Figure 2. Bonds joining two residues (called *peptide bonds*) hydrolize (i.e., break under consumption of a water molecule) in a sufficiently acid 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.

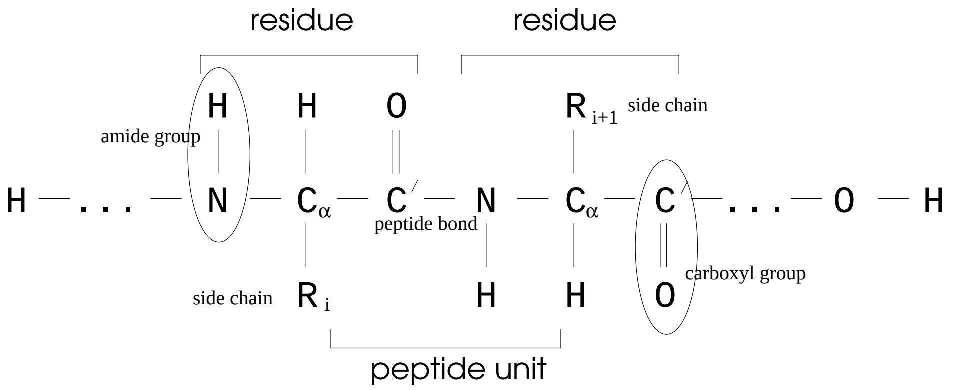

Fig. 2. **The chemical structure of a protein**

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 a way characteristic 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, 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 10000. One of the smallest proteins, BPTI (bovine pancreatic trypsin inhibitor), with 58

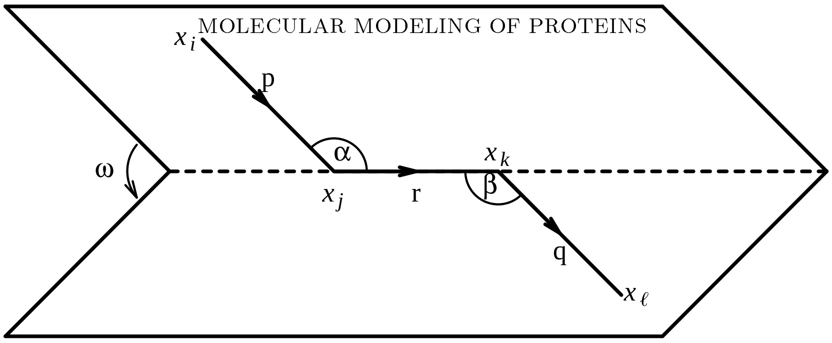

FIG. 3. **Bond vectors, bond angles, and the dihedral angle**

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 3-dimensional *coordinate vector*

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

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* $\alpha = \sphericalangle(i\text{-}j\text{-}k)$ 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 $\alpha \in [0^o, 180^o]$), where

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

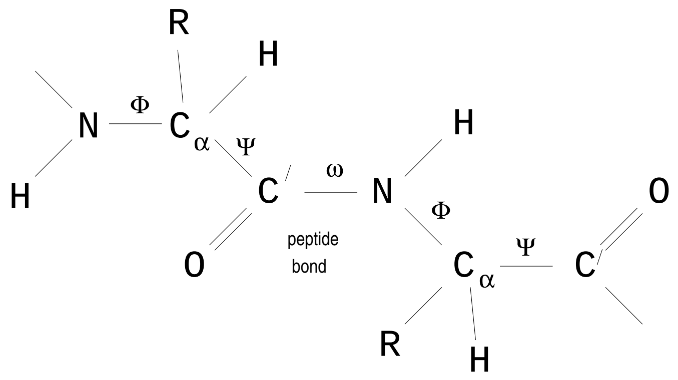

FIG. 4. **Backbone dihedral angles of a protein**

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* $\omega = \sphericalangle(i\text{-}j\text{-}k\text{-}l) \in [-180^o, 180^o]$ (or the complementary *torsion angle* $180^o - \omega$) 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 $(q\times p, r)$.

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 useful to specify in addition a number of so-called *out-of-plane bending* or *improper torsion angles* $\tau = \sphericalangle(i\text{-}j\text{-}k\text{-}l)$, that 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.

In a protein, the bond angles are usually denoted by the letter θ, and the dihedral angles describing the torsion around the backbone N–C_α, C_α–C', and C'–N bonds by the letters φ, ψ and ω, respectively; dihedral angles in the side chain by χ.

Under biological conditions, the bond lengths and bond angles are fairly rigid (with a standard deviation of less than $0.2\AA$ for bond lengths and of about 2^o for bond angles, see HENDRICKSON[137]; recent experimental values are reported, e.g., in ENGH & HUBER [90]). Therefore, the dihedral angles along the backbone (usually labeled as in Figure 4) determine the main features of the final geometric shape of the folded protein.

Structural information for the proteins with known geometry is collected world wide in the quickly growing Brookhaven Protein Data Bank (BERNSTEIN [19]), accessible through the WWW at `http://www.pdb.bnl.gov`; see also WALSH [338], STAMPF et al. [302].

3. Molecular mechanics. In this section we 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, we replace the family of coordinate vectors x_i in 3-space 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 $1500 - 30000$.

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

$$M\ddot{x} + C\dot{x} + \nabla V(x) = D\dot{W}(t). \tag{1}$$

(The dots denote differentiation w.r. to time. Physicists and chemists often use the term *Langevin dynamics* for such equations. For an exposition of stochastic differential equations from a rigorous point of view but without excessive generality, and with some chapters readable by non-specialists, see KLOEDEN & PLATEN [167].)

The first term of (1) describes the change of kinetic energy, and is the product of the *mass matrix* M and the *acceleration* $\ddot{x}$. In Cartesian coordinates (as we have adopted here), the mass matrix is diagonal, with diagonal entries equal to the mass of an atom at the three positions corresponding to this atom. The second term describes the excess energy dissipated to and absorbed by the surrounding, and is the product of a symmetric, positive definite *damping matrix* C and the *velocity* $\dot{x}$. The third term describes the change in potential energy, and is expressed as the gradient of a real-valued *potential function* V characteristic of the molecule and defined for all x except when two coordinate vectors x_i and x_j coincide. The potential is discussed in more detail in the next section. Finally, the right hand side is a random force accounting for fluctuations due to collisions with the surrounding that dissipate the energy; it is the product of normalized white noise $\dot{W}(t)$ with a suitable matrix D. Here $W(t)$ is the Wiener process. For generalities about modeling with stochastic differential equations (from the physicist's point of view) see VAN KAMPEN [335], GARDINER [103].

Actually, the interaction with the environment is much more intricate, the damping and fluctuation terms are only simplified descriptions; e.g., their dependence on positions and velocities is ignored, memory terms are missing, and time correlations of the noise are neglected, although they figure in more careful elimination procedures of the environment. Rather than improving such contracted formulations, more accurate representations used in practice add the positions of a number of water molecules (and other molecules present in the solvent in a significant amount) to the state vector, and the potential is extended to account for their interactions with each other and the molecule.

Because of the orthogonal symmetry of the multivariate Wiener process, the stochastic differential equation (1) only depends on the covariance matrix of the noise term $\varepsilon(t) = D\dot{W}(t)$,

$$DD^T = \langle \varepsilon\varepsilon^T \mathrm{dt} \rangle$$

(where $\langle ... \rangle$ denotes expectation, and D^T denotes the transpose of D) called the *diffusion matrix* of the fluctuation term. Conservation of energy on the microscopic level requires that the diffusion matrix is related to the damping matrix by the *fluctuation-dissipation theorem*

$$DD^T = 2k_B TC, \tag{2}$$

involving the *temperature* T and the *Boltzmann constant* k_B. For discussions of its validity see, e.g., Deutch & Oppenheim [77] and Bossis et al. [27].

The matrices C and D which model the coupling to the environment are typically matrices with small entries. Since little is known about this coupling, C is usually modeled by a small scalar multiple of the (diagonal) mass matrix,

$$C = \gamma M,$$

with the *damping coefficient* γ as a free parameter determined by some heuristic (see, e.g., [358]), and the random part is then determined by (2) as $D = \sqrt{2k_B T\gamma} M^{1/2}$ (the choice of another solution of (2) is equivalent to this one). As long as the coupling to the environment is small, this can be justified to some degree since the qualitative features of the dynamics are independent of the precise form of the damping and random forces, being mainly governed by the form of the potential.

One can argue for the above default choice from the form of the equation. Indeed, if $C \gg M$, the second-order term can be neglected except in an initial phase, and one would naturally expect that the resulting first order equation follows the solution of the initial value problem

$$M\dot{z} + \nabla V(z) = 0, \ z(0) = z^0, \tag{3}$$

although on a different time scale. (After a suitable transformation, this can be interpreted as the *steepest descent path* in the physically relevant, mass-scaled coordinates.) This requires that C is a scalar multiple of the mass matrix M. An even simpler heuristic argument simply demands that the acceleration $\ddot{x}$ is damped in the direction of the negative velocity $\dot{x}$ only, giving the same dependence of C on M.

However, the choice of C affects the boundaries of the catchment regions for the dynamics, hence may be relevant for more detailed investigations. Realistic choices for C are position (and probably also velocity) dependent non-diagonal matrices. These can be derived in terms of time correlation functions from the interaction with the solvent; see Deutch & Oppenheim [77], Ciccotti & Ryckaert [51] and Chapter 9 of Allen & Tildesley [3]. An estimation of the required time correlation functions can be obtained by molecular dynamics simulations involving explicit solvent molecules.

The low temperature limit. In the low temperature limit $T \to 0$, the covariance matrix (2) vanishes; so this limit corresponds to the absence of random forces. In this limit, (1) becomes the ordinary differential equation

$$M\ddot{x} + C\dot{x} + \nabla V(x) = 0. \tag{4}$$

In this case, the sum of kinetic and potential energy,

$$E = \frac{1}{2}\dot{x}^T M \dot{x} + V(x),$$

has time derivative $\dot{E} = \ddot{x}^T M \dot{x} + \nabla V(x)^T \dot{x} = -\dot{x}^T C \dot{x}$. Since C is positive definite, this expression is negative, and vanishes only at zero velocity, when all energy is potential energy. Thus the molecule continually loses energy until (in the infinite time limit) it comes to rest at a stationary point of the potential, $\nabla V(x) = 0$ (by (4)). This stationary point generally is a local minimum of the potential, since otherwise it is unstable under even tiny random fluctuations.

Under realistic conditions, the temperature is positive, and random forces will continue to add kinetic energy to the molecule, so that it will describe random oscillations around the local minimum. And after sufficient time, rare large random forces may allow the molecule to stray very far from the local minimum, possibly into another valley corresponding to another geometric configuration.

For reasonably rigid molecules, characterized by the fact that there is a unique local (and hence global) minimizer in the part of state space accessible to the molecules, this determines the kinetics of the molecules up to temperatures sufficiently high to break some of the chemical bonds, thus altering the nature of the molecules. Therefore, the geometry defined by the minimizer of the potential energy surface, referred to as the *stable state* of the molecule, gives a correct geometric description of the average shape of rigid molecules.

However, proteins, as other polymers, are not rigid but can be easily twisted along the bonds of the backbone. As a consequence, the potential energy surface becomes very complicated and exhibits a large number of local minima. Thus, from time to time, and more frequently at higher temperature where the random forces are larger, large random excitations allow the molecule to escape from the neighborhood of one local minimum and reach the neighborhood of another one. In the language of chemistry, the local minima of the potential (and their neighborhoods) are now only *metastable states*, and the occasional escapes to other local minima are referred to as *state transitions*.

State transitions. The frequency of transitions depends on the temperature and on the energy barrier along the energetically most favorable path between two adjacent local minima. Any such path has its highest point on the energy surface at a saddle point called a *transition state*. (The most natural definition – the literature is somewhat vague here – of this path, the so-called *reaction coordinate*, is a continuously differentiable, non-constant solution of

$$(\det \nabla^2 V(z)) M \dot{z} + \nabla V(z) = 0, \tag{5}$$

passing a number of stationary points, among them the two local minima, and the saddle point as the unique stationary point in-between. Here $\nabla^2 V(z)$ denotes the Hessian matrix of second derivatives of V at z.)

A state transition thus proceeds from the neighborhood of one metastable state, passing near a transition state, and moving towards a neighboring metastable state. The book MEZEY [205] contains a thorough treatment of the topology of potential energy surfaces, their stationary points and their catchment regions. In particular, it turns out that the transition states (saddle points) are stationary points of the potential whose Hessian has just one negative eigenvalue while the Hessian at metastable

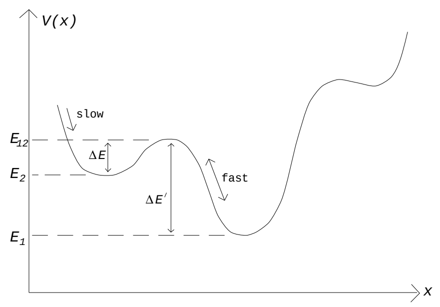

FIG. 5. **Transition energies between adjacent local minima**

states (local minima) has, of course, no negative eigenvalues. (For methods to calculate saddle points and associated reaction paths see McKEE & PAGE [204], CULOT et al. [68], and the references there. These methods appear to be much less developed and are at present much less reliable than methods for minimization.)

The energy difference ΔE between the energies at a local minimum and at a transition state is called the *activation energy* for the transition. Using suitable approximations (see, e.g., VAN KAMPEN [335], Section XI.6-7 or GARDINER [103]), one can derive from the stochastic differential equation (1) the *Arrhenius law*, that predicts the mean transition frequency k as

$$k = \frac{k_B T}{h} \exp\left(-\frac{\Delta E}{k_B T}\right), \tag{6}$$

with h being Planck's constant. Note that in the literature on reaction dynamics, energies are often normalized to correspond to one mole of a substance instead of to single molecules; then the constant appropriate in place of the Boltzmann constant is the *gas constant* R, and the Arrhenius law takes the more familiar form

$$k = \frac{RT}{h} \exp\left(-\frac{\Delta E}{RT}\right).$$

The exponential term shows that transitions become vastly more difficult and more rare as the activation energy increases, while at higher temperatures, transitions become easier. This is in accordance with the intuition that at higher temperatures more random energy is available, which more frequently exceeds the amount required to pass the transition state.

We now consider two adjacent local minima with potential energies E_1 and E_2 and the corresponding transition state with energy E_{12}. (See Figure 5 for a cross section of the energy surface along the reaction coordinate.) Suppose that $E_1 < E_2$. The activation energy $\Delta E_{1\to 2} = E_{12} - E_1$ needed to cross from state 1 to state 2 is larger than the activation energy $\Delta E_{2\to 1} = E_{12} - E_2$ needed to cross from state 2 to state 1. Using the Arrhenius law (6) we find that the corresponding transition frequencies satisfy

$$k_{2\to 1} : k_{1\to 2} = 1 : \exp\left(\frac{E_2 - E_1}{k_B T}\right)$$

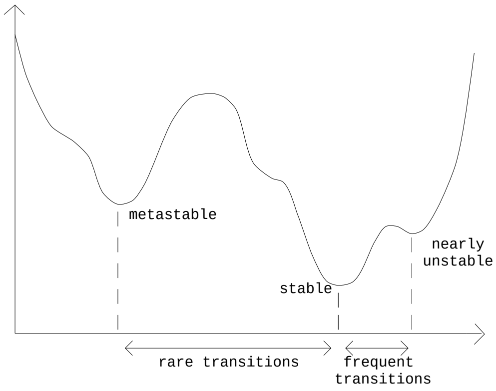

FIG. 6. **A metastable state with high energy barrier**

independent of the transition state energy. In particular, unless both metastable states have nearly the same potential energy, transitions from the state with the higher energy to that with the lower energy are *much* more frequent than transitions to a higher energy level.

This implies that over sufficiently long time scales, a molecule spends most of its time in the deepest valley near the global minimizer of the potential. It also shows that in a collection of many (independent) molecules of the same kind, most molecules are in a conformation close to the state with absolutely smallest potential energy. This is the reason why most experts (exceptions will be discussed later) expect that the geometry defined by the *global* minimum of the potential energy surface is the correct geometry describing the conformation observed in folded proteins. However, if the energy barrier of a metastable state is sufficiently high, transition frequencies may be so low that within a biologically meaningful time the metastable state behaves like the stable state given by the global minimum.

It should be noted that the Arrhenius law, based on classical dynamics and bistable potentials (i.e., with only two local minima), is only approximately valid. For a discussion of quantum corrections see VANGUNSTEREN & BERENDSEN [334] and VOTH & O'GORMAN [337]. A complete review of the subject from the classical and the quantum point of view is given by HÄNGGI et al. [125].

4. The harmonic approximation. In a molecular system (such as a protein) where the atoms are highly mobile, the potential energy surface has a complicated shape; the varied topography of our earth gives an impression of what is possible in two dimensions, and the possibilities in the high-dimensional state space are even greater. Therefore it seems impossible to attain the ideal of studying all features of the dynamics; in practice one must be content with the exploration of sample paths through the state space by means of so-called *molecular dynamics calculations* [26, 138, 142, 202]. These 'solve' the stochastic differential equation by simulating sample paths of (1) using pseudo-random techniques. (Actually, the texts on molecular dynamics calculations are not clear about this relation; they usually treat the stochastic terms in an ad hoc manner that makes it difficult to assess the accuracy of the results obtained.) A detailed description of the intricacies of Monte Carlo and molecular dynamics simulations is given in the book by ALLEN & TILDESLEY [3]; a useful survey is VAN GUNSTEREN [331]. For a discussion of numerical methods for general stochastic differential equations see, e.g., GREINER et al. [121], HONERKAMP

[145], Iniesta & de la Torre [151], Kloeden et al. [168], Sobczyk [300], and van Gunsteren & Berendsen [333]. For an error analysis of numerical methods for stochastic differential equations see the book by Kloeden & Platen [167] and references there, and Bishop & Frinks [21]. The gap to the present knowledge in the numerical analysis of deterministic differential equations (see, e.g., Hairer et al. [128, 129]) seems enormous, leaving much scope for research.

However, the high frequency behavior of the motion of molecules at low temperature can be studied in a simpler way by means of the so-called *harmonic approximation*. The reason is that at high frequencies, only tiny motions are possible, and at low temperature, the motion is confined with high probability to a neighborhood of a local minimizer x_{loc}. Thus it is justified to expand the potential into a Taylor series around x_{loc}, truncating it after the quadratic term. Since the gradient vanishes at a local minimizer, we obtain the approximation

$$V(x) \approx V(x_{loc}) + \frac{1}{2}(x - x_{loc})^T K(x - x_{loc}). \tag{7}$$

Here $K = \nabla^2(x_{loc})$, the *Hessian* of the potential, is (in the absence of degeneracy) a positive definite symmetric matrix, called the *stiffness matrix*. Under our assumptions we may also neglect damping and random forces, and obtain from (1) the linear differential equation for the harmonic approximation,

$$M\ddot{x} + K(x - x_{loc}) = 0. \tag{8}$$

This differential equation has the general solution

$$x = x_{loc} + \sum_l e^{i\omega_l t} u_l, \tag{9}$$

where the *frequencies* ω_l and the *normal modes* u_l (describing the vibration patterns corresponding to these frequencies) are the eigenvalues and corresponding eigenvectors of $M^{-1}K$. (Here we assumed for simplicity that no multiple eigenvalues occur.) The frequencies are observable as spectral lines, and (by linear response theory) the normal modes are observable, too.

The highest frequencies occurring in proteins are of the order of 10^{14}/sec and the corresponding normal modes essentially correspond to stretching a C-H bond (with small compensating changes in the other bonds and the angles). Vibrations corresponding to bond-angle bending have frequencies of the order of 10^{13}/sec. Non-vibrational internal motions are geometrically distinguishable at time scales of around 10^{11}/sec Creighton [63]. These involve non-local changes and roughly correspond to lower frequency normal modes. While low frequencies are irrelevant for the real dynamics since the assumptions used to derive the harmonic approximation are no longer valid, the invariant subspace spanned by all low frequency eigenvectors is the space in which the long time dynamics takes place.

For large molecules, the eigenvalue problem is in itself already a nontrivial numerical task, with work growing like $O(n^3)$ for n degrees of freedom if the full spectrum is wanted. To obtain only the low frequency eigenmodes, iterative methods like the Lanczos algorithm or subspace iteration can be used; see, e.g., Parlett [231]. The work can also be reduced by various approximations. Fixing bond lengths and bond angles removes the high frequency spectrum (Levitt et al. [187], Gibrat et al. [108]). Splitting the molecule into suitable pieces allows an approximate divide-and-conquer approach by solving eigenvalue problems for the pieces and for a condensed matrix (Hao & Harvey [131]).

Time scales. The frequency analysis has consequences for the molecular dynamics simulations. Indeed, current algorithms for tracking sample paths for solutions of stochastic differential equations need to proceed in time steps significantly smaller than the smallest time scale of the oscillations in order to properly trace the effect of the interaction between these oscillations and the random forces. Thus, time steps of the order of $10^{-15}sec$ are called for. Since, as will be explained in the next section, the potential evaluation needed for each time step is rather expensive, only of the order of 10^7 time steps can be performed in a reasonable amount of time on present day computers, corresponding to an interval of up to a few nanoseconds.

These numbers mainly intend to give an idea of typical scales; detailed numbers depend of course on the computer used, on the algorithms used, on the size of the protein, on the potential employed, and on the time one is prepared to wait for the results. In 1985, the limit was at only about 0.3 nanoseconds (LEVY et al. [188]); now there are studies covering several nanoseconds (e.g., TOBIAS et al. [319], GUO ET AL [124], SOMAN et al. [301]) and, according to GODZIK et al. [115], 100ns are feasible in a lattice approximatimation. Progress in algorithmic ingenuity and computer technology will push up the limit further. A recent survey of folding studies is CAFLISCH & KARPLUS [42]; see also DAGGETT & LEVITT [69].

On the other hand, the experimentally accessible time resolution (see, e.g., RADFORD & DOBSON [246]) is of the order of milliseconds, and typical times observed experimentally for a protein to fold (in the absence of catalyzing enzymes) are in the order of $\sim 10^{-1} - 10^3$ seconds. This shows that we are still very far away from a computational treatment of the dynamics of protein folding. A critical evaluation of the results currently obtainable with molecular dynamics simulations on some practical problems in drug design is given in KÖPPEN [170]. See also MCCAMMON & HARVEY [202], MCCAMMON & KARPLUS [203]. Time saving techniques using so-called multiple time-steps are based on the fact that different parts of the forces change at different time scales; see TELEMAN & JÖNSSON [317], TUCKERMAN et al. [322, 323], and WATANABE & KARPLUS [340].

The fastest time scales can be suppressed by fixing the fastest changing variables, especially bond lengths. (Fixing also bond angles distorts the dynamics significantly; see VAN GUNSTEREN & BERENDSEN2 [333].) Enforcing a vector of constraint equations $B(x) = 0$ in a dynamical equation requires adding to the differential equation a force term proportional to the gradient of $B(x)$. (This can be justified by variational principles.) The result is the *differential-algebraic equation*

$$\begin{gathered} M\ddot{x} + C\dot{x} + \nabla V(x) - \nabla B(x)\lambda = D\dot{W}(t), \\ B(x) = \text{const.} \end{gathered} \tag{10}$$

for the pair (x, λ) consisting of the state vector and the Lagrange multiplier. In the chemical literature, people speak of *constrained dynamics*; see, e.g., RYCKAERT et al. [258], VANGUNSTEREN & BERENDSEN [332, 333] and MIYAMOTO & KOLLMAN [208]. A mathematical analysis of the widely used SHAKE algorithm [258] is given by BARTH et al. [14]. For the mathematics and the solution of (non-stochastic) differential-algebraic equations in general, see the books by BRENAN, CAMPBELL AND PETZOLD [31] and HAIRER et al. [127], and the survey article MÄRZ [193].

Some speed can also be gained by considering instead of a full molecular representation of the protein a reduced representation in terms of *extended atoms*, where the hydrogen atoms responsible for the highest frequencies are not modeled explicitly but are treated as part of the atoms they are bonded to, thus producing extended atoms like CH_2, OH, NH, etc. and using large time steps. However, the results obtained in this way can only be considered as rough approximations. (The models discussed in MOMANY et al. [212] and TROYER & COHEN [321] make even more drastic simplifications to reduce the size of the problem further. See also CHAN & DILL [48].) MONGE et al. [213] gains speed in a different way by assuming known secondary structure (cf. Section 7), that is frozen to limit the number of degrees of freedom.

An important development concerns the numerical methods for solving stochastic differential equations, that at present are mostly explicit methods. To cope with the oscillatory stiffness introduced by frequencies on multiple time scales, implicit methods used successfully for the numerical solution of stiff systems of ordinary differential equations [129], in particular so-called A-stable methods, can be adapted to take account of random forces. Implicit Euler-steps were studied in PESKIN & SCHLICK [237]. The solution of the resulting implicit equations can be achieved by local optimization of a dynamical energy function (SCHLICK [272]); see Appendix 1 for more details. Since the high frequencies are damped, this approach is suitable for macroscopic models where the high frequencies are absent, or virtually uncoupled, from the slow modes. For applications to DNA supercoiling, see [275, 248].

An interesting possibility explored in ZHANG & SCHLICK [358] is to combine the harmonic approach and the stochastic molecular dynamics approach to get a direct handle on the fast oscillations and simulate essentially the behavior on the slow modes. ZHANG & SCHLICK [359] improve this further by solving the linearized stochastic differential equations exactly. Together, these developments offer a promising way out of the limitations described above, and allow already much larger step sizes of up to 10^{-12}sec, valuable for sampling. Indeed, significant speedup was achieved very recently by a simplified version of this approach in BARTH et al. [15]. The resulting scheme can be considered as a method with two different timesteps: $\delta\tau = 0.5 \times 10^{-15}$sec for solving the harmonic model, and $\delta t = 5 \times 10^{-15}$sec for updating the harmonic model.

Multiple time scales also arise in the dynamics of electrical circuits; multirate strategies developed in that context (see, e.g., the papers by DENK, by GÜNTHER & RENTROP and by WRIEDT in BANK et al., [13]) may also prove useful to the dynamics of proteins.

Robustness questions of numerical methods, related to the symplectic (Hamiltonian) nature of the (undamped) dynamics and discussed, e.g., in SANZ-SERNA [263] and SKEEL et al. [292], may also turn out to be relevant. Numerical methods taylored to Hamiltonian problems are treated in a recent book by SANZ-SERNA & CALVO [264]. Hamiltonian formulations are also worthy of investigation in the case of constrained dynamics. The use of symplectic integrators in molecular dynamics calculations is discussed in GRAY et al. [119].

A good discussion of the numerical problems involved in tracing the dynamics of molecules, together with further references, is given in the section on molecular dynamics algorithms of [218].

5. Modeling the potential. Strictly speaking, the dynamics of atoms in a molecule is governed by the quantum theory of the participating electrons. For chemical applications, the so-called *Born-Oppenheimer approximation* is usually considered to be adequate. In the Born-Oppenheimer approximation, one obtains the energy $V(x_1, \ldots, x_N)$ at fixed nucleus positions x_i as the smallest eigenvalue of an associated partial differential operator for the electron wave function (the Hamiltionian of the electron system). Approximations of such eigenvalues (and of their partial derivatives with respect to the nucleus positions) can be computed by so-called *ab initio* methods. However, for complex molecules, quantum mechanical calculations are far beyond the computational resources likely to be available in the near future.

Hence chemists usually use a classical description of molecules in terms of bonds and effective atomic interactions, the only trace left of the electrons being partial charges on the atoms. Quantum theoretical calculations are restricted to the calculations of properties of small constituent parts of the molecule (such as amino acids), and phenomenological models are constructed from the data obtained in this way and from experiment to allow extrapolation to larger molecules. Some general references on molecular modeling are Burkert & Allinger [37], Gund & Gund [122] and Hirst [138].

The interactions of the atoms in proteins can be classified into bonded and non-bonded interactions. Bonded interactions depend on the nature of the bond: At the energies and time scales of interest, *covalent bonds* (the bonds drawn as lines in chemical formulas) are considered un-breakable, *disulfide bonds* (joining two close sulfur atoms, S - - - S) are slow to form and to break, and *hydrogen bonds* (joining hydrogen atoms with close oxygen atoms, H $\cdots$ O) are fairly easily formed and broken.

Atoms far apart are subject to non-bonded interactions: If both atoms carry partial charges, there is the long range, slowly decaying *electrostatic (Coulomb)* interaction, and for all pairs of atoms there is the short range, fast decaying *van der Waals* interaction.

Hydrogen bonds and non-bonded interactions are particularly relevant for the interaction of the molecule with the atoms of the solvent (water). The Coulomb interaction is modified by polarization effects due to the presence of solvent, and this is modeled in the simplest case by a (often distance-dependent) *dielectric constant* D. However, this is somewhat inadequate, and there have been a number of recent studies that treat the electrostatic effects due to the solvents by adding energy terms defined by so-called continuum solvation models. Two promising methods are in use (see Cramer & Truhlar [60] for a detailed, up-to-date review): Cheaper models are based on solvent-accessible surface area (Richmond [251], Wesson & Eisenberg [343], Schiffer et al. [270]); cf. Appendix 2. More realistic models (e.g., Gilson et al. [113], Nicholls & Honig [223, 147]) are based on an approximate solution of the *Poisson-Boltzmann equation*

$$\nabla \cdot [\varepsilon(x)\nabla\varphi(x)] - \kappa(x)^2 \sinh(\varphi(x)) = -4\pi\rho(x).$$

The solution of this nonlinear partial differential equation for the electrostatic potential $\varphi(x)$, given model assumptions defining the spatial dielectric function $\varepsilon(x)$ (often discontinuous at the molecule boundary), the Debye-Huckel parameter $\kappa(x)$ (often taken constant), and the charge distribution $\rho(x)$ (typically a sum of partial charge delta functions), is a nontrivial numerical problem. (For background information on dielectrics, see, e.g., Fröhlich [101]. In a constant external electric field E, there is an additional electrostatic contribution $-\mu(x) \cdot E$ to the potential involving the *molecular dipole moment* $\mu(x)$, but we do not discuss this further since usually an electrically neutral environment is assumed.) For molecular dynamics studies of solvent effects, obtained by embedding the protein molecule in a large set of explicit water molecules, see, e.g., Schreiber & Steinhauser [278, 279, 280] and Tapia [315].

The static forces in a molecule are fully determined by the formula defining the potential $V(x)$, so modeling the molecule simply amounts to specifying the contribution of the various interactions to the potential. The models – also called *force fields* –

$$
\begin{array}{llll}
V(x) = & \displaystyle\sum_{\text{bonds}} & c_l(b-b_0)^2 & (b \text{ a bond length}) \\
+ & \displaystyle\sum_{\text{bond angles}} & c_a(\theta-\theta_0)^2 & (\theta \text{ a bond angle}) \\
+ & \displaystyle\sum_{\substack{\text{improper}\\ \text{torsion angles}}} & c_i(\tau-\tau_0)^2 & (\tau \text{ an improper torsion angle}) \\
+ & \displaystyle\sum_{\text{dihedral angles}} & \mathrm{trig}(\omega) & (\omega \text{ a dihedral angle}) \\
+ & \displaystyle\sum_{\text{charged pairs}} & \dfrac{Q_iQ_j}{Dr_{ij}} & (r_{ij} \text{ the Euclidean distance from } i \text{ to } j) \\
+ & \displaystyle\sum_{\text{unbonded pairs}} & c_w\varphi\left(\dfrac{R_i+R_j}{r_{ij}}\right) & (R_i \text{ the radius of atom } i).
\end{array}
$$

Table 1
The CHARMM potential

currently in use (see [7, 33, 52, 72, 212, 219, 341, 342]; a short comparative description of many force fields is given in the discussion part of Cornell et al. [58]) derive their basic structure from the times when molecular mechanics was only used to match the observed structural and spectral data for rigid molecules (or molecules of very limited mobility) to the available theory. In particular, local expansions around equilibrium data could be used without difficulties, and this still shows in the current models.

However, local expansions are much more questionable for global optimization and global dynamical calculations since the potential must now be approximated correctly over much larger regions. We shall therefore describe in some detail one particular modeling approach implemented in the molecular mechanics software package CHARMM (Brooks et al. [33]), mention some of the numerical difficulties reported in the use of this package, and discuss the problems associated with the use of the CHARMM potential for global purposes. The analysis leads naturally to the proposal of a revised model that avoids both the numerical difficulties and the non-physical aspects of this model.

The CHARMM potential. The CHARMM model represents the potential essentially as a sum of six kind of terms given in Table 1.

The Q_i are *partial charges* assigned to the atoms in order to approximate the electrostatic potential of the electron cloud, and D is the dielectric constant. The quantities indexed by 0 are reference bond lengths, bond angles, and improper torsion angles near their equilibrium values; different constants apply depending on the names of the atoms in the various atomic sequences, and sometimes on their location in the functional group, too. The coefficients of the trigonometric terms $\mathrm{trig}(\omega)$, (linear combinations of cosines of multiples of ω), and the *force constants* $c_{\bullet}$ (whose magnitude reflects the strength of the respective forces) are determined, too, by the names of the atoms corresponding to the term in question. That these constants are indeed independent of the molecule is a basic assumption of molecular mechanics called *transferability*, an assumption not always unquestioned (VEENSTRA et al. [336]).

Some further terms, accounting specifically for disulfide bonds and hydrogen bonds, are also present, but will not be discussed here. (For the modeling of hydrogen bonds see, e.g., SCHEINER [267] and LADANYI & SKAF [176].) There are also more complicated alternative versions for the electrostatic interaction; cf. WILLIAMS [345].

The van der Waals interactions (defined by the final sum in the potential) depend on the interatomic pair potential φ that, in the simplest case, is taken as the *Lennard-Jones potential*

$$\varphi(\frac{R_0}{r}) = \left(\frac{R_0}{r}\right)^{12} - 2\left(\frac{R_0}{r}\right)^{6}. \tag{11}$$

The first term drastically decreases for small r forcing atoms to repel each other at short distance. The second term slowly increases for large r, causing an attraction of neutral atoms at large distance. The particular linear combination leads to an equilibrium at the minimum of $\varphi(\frac{R_0}{r})$ at $r = R_0$, and thus accounts for a qualitatively correct behavior. The large distance decay of the potential as r^{-6} can be derived from quantum mechanics (see, e.g., KAPLAN [162] or KIHARA & ICHIMARU [165]), whereas the power 12 in the attractive term, modeling strong physical repulsion, is chosen mainly for easy calculation (by squaring the second term). However, details of the attractive term may change the nature of global opima; e.g., HOARE & MCINNES [140] report that, for the simpler problem of inert gas crystals, softer Morse potentials favor regular crystals (that are believed to be the global optimum of the 'true' potential).

To make the interatomic potential more realistic quantitatively, terms with other powers of $\frac{R}{r}$ are included in CHARMM. (For more in depth studies of pair potentials see STEELE [303], MAITLAND [195].) It is quite possible that more accurate potentials will also need to take three-body forces into account, such as those given by so-called *Axilrod-Teller terms*, predicted by quantum mechanics (AXILROD & TELLER [11], KIHARA & ICHIMARU [165]).

Taking the equilibrium distance R_0 between two atoms as the sum $R_0 = R_i + R_j$ of the atomic radii is a simple intuitive instance of a *combination rule* designed to reduce the number of parameters that need to be supplied. However, the form of the best combination rules is not clear; e.g., MAPLE et al. [196] use instead the rule $R_0^6 = R_i^6 + R_j^6$. Thus $R_0 < R_i + R_j$, i.e., the atoms overlap in their equilibrium position. (Some arguments for a particular combination rule are given in SLATER & KIRKWOOD [297] GILBERT [111]. From a quantum theory point of view, there is of course no atomic radius, and what we here call atoms are just balls with semiempirical nominal radii R_i around the nucleus positions.)

The sum over the charged terms is the Coulomb interaction. There is some disagreement on how to assign the partial charges (see the discussion in WILLIAMS [345] LEE et al. [181]), and there is even some indication that, except for molecules with net charge, better results might be obtainable without such Coulomb terms CLARK et al. [52]).

As stated, the Coulomb interaction is valid only for a homogeneous dielectric medium. However, the protein in solution is really inhomogeneous, with a position-dependent dielectric constant that near the atoms of the protein is only about one eighth of that of the solvent water. It is not clear how this affects the effective Coulomb interaction between the atoms of the proteins, and so far, only heuristic corrections (such as a distance-dependent D) are in use. The free energy of solvation can also be accounted for by terms proportional to the surface area exposed to the solvent; see, e.g., WESSON & EISENBERG [343] PERROT et al. [235], SCHERAGA [269] and SCHIFFER et al. [270]. See also the molecular surface review by CONOLLY [56]. Possibly, the solvation energy can also be accounted for by modifications of the pair potentials and the combination rules; see Appendix 2. For the explicit modeling of water molecules, which is much more expensive but avoids all these problems, see STILLINGER [305] and WARSHEL [339].

To speed up the potential evaluation, the potential is further modified by introducing a cutoff distance beyond which the interatomic potential is neglected. (Recent evaluations of the effect of cut-offs on molecular dynamics simulation include SMITH & PETTIT [299], STEINBACH & BROOKS [304] and SCHREIBER & STEINHAUSER [278, 279, 280].) With such a cutoff, only a small part of the $O(n^2)$ terms in the last sum in the potential of Table 1 need to be calculated. To make full use of the resulting sparsity (which varies from iteration to iteration), efficient data structures must be maintained; see, e.g., SCHREIBER et al. [281]. More recently, fast multipole expansions [25, 120, 284] and variants [70, 91] of the Ewald method (EWALD [92]) were used as an alternative to cut-off methods, with a significant increase in quality YORK et al. [353, 354]. An improvement of a divide-and-conquer method by APPEL [8] for fast potential evaluation is discussed in XUE et al. [352]. While useful for the simulation of fluids, the break-even point seems to be too high to make it useful for protein calculations.

Improving potential features. It is easy to see that the first three sums in the potential derive their form from a truncated Taylor expansion around equilibrium values. Linear terms are missing since they can be incorporated into the quadratic term by changing the value of the equilibrium constants. But cross terms like $(\omega - \omega_0)(\theta - \theta_0)$, considered in HAGLER [126] and to be expected in any multivariate Taylor expansion (BOWEN & ALLINGER [28]), are missing, too, and the main reason (to be discovered by reading between the lines of a number of standard texts on molecular mechanics) seems to be that – in the past – the available data were not sufficient to estimate the corresponding coefficients!

However, cross terms lead to much better agreement with vibrational spectroscopy measurements (DERREUMAUX & VERGOTEN [75], MAPLE et al. [196]. Moreover, cross terms can drastically modify the global behavior, especially because of the non-linearities in the non-bonded contributions, and if we ever want to obtain reliable quantitative predictions from global molecular mechanics calculations, these issues must be addressed much more carefully.

Since much more data are available now (and even more can be generated by ab initio calculations) than at the time when the form of the potential was fixed by tradition, the old reasons for such drastic simplifications are no longer appropriate. The literature discusses some other possibilities: Adding so-called Urey-Bradley interaction terms (UREY & BRADLEY [326] and references in Chapter III of BROOKS et al. [34]), of the form

$$c\|r_{ik} - r_0\|^2$$

for atoms bonded as i-j-k has, locally, an effect similar to adding cross terms between bond lengths and bond angles. MAPLE et al. [196] add even some cubic interaction terms, and show that these significantly improve the fit to ab initio data. FOGARASI & PULAY [98] mention (at the end of section III) that using inverse bond lengths instead of bond lengths may be advantageous, and by the same reasoning one can also argue for bond length denominators in cross terms.

For the forces accounting for the relatively easy twisting along the bonds it was well-known that Taylor expansions were not realistic, and this accounts for the more sophisticated trigonometric terms involving the dihedral angles. However, since bond lengths, bond angles, and improper torsion angles involving the peptide bonds are much more rigid, the need for analogous corrections on the corresponding contributions was not apparent as long as the potentials were only used for local (spectral) analysis. However, it is easy to see that globally, the terms for bond angles and improper torsion angles are non-physical: For example, the physically equivalent angles $\theta = 160^o$ and 200^o give rise to different potentials. The minimal change needed to restore a global physically meaningful interpretation is to replace the bond angle contributions by $c(\cos\theta - \cos\theta_0)^2$ and the improper torsion angle contributions by $c(\sin(\tau - \tau_0))^2$, for suitable constants c. This is (approximately) realized in the force field used in [226, 271, 277].

Another defect of the dihedral angles (and similar remarks apply to improper torsion angles) is the fact that these angles are geometrically undetermined when a bond angle is 180^o; the formulas then give the expression $0/0$ for $\cos\omega$ and $\sin\omega$. This invites numerical disaster: for angles close to 180^o, these quotients are numerically very unstable. Thus rounding errors lead to low accuracy or even essentially random values for the dihedral angles, resulting in random energy contributions. Although equilibrium angles are typically far away from 180^o, this is an important defect in global applications; for example it ruins (or produces unpredictable results in) any local optimization routine if one of the angles in an intermediate calculation (such as a line search) happens to come close to 180^o. BROOKS et al. [33] observe (on p. 191/2) “singularities when angles become planar (which is rather common)”; they correct for it in an ad hoc way by using Taylor expansions.

However, the natural resolution of this difficulty is to use in the potential only expressions that are geometrically well-defined for all values of the bond angles. In the notation of Section 2, the natural quantities involving dihedral angles are the products $\sin\alpha\sin\beta\sin\omega$ and $\sin\alpha\sin\beta\cos\omega$ that can be calculated in a numerically stable fashion from the formulas

$$\sin\alpha\sin\beta\sin\omega = \frac{(q\times p, r)}{\|p\|\|q\|\|r\|},$$

$$\sin\alpha\sin\beta\cos\omega = \frac{(p\times r, r\times q)}{\|p\|\|q\|\|r\|^2}.$$

The force field used by SCHLICK [271, 277] avoids the instabilities in a different way by replacing the denominators in the definition of all angles by constant reference values. As a byproduct of these proposed modifications, the evaluation of $V(x)$ even becomes cheaper since no inverse trigonometric functions have to be computed.

The conclusion of our analysis is that, in order to construct globally meaningful potentials that would allow one to hope for a correct quantitative predictions of protein structure, we need to use more carefully designed terms implementing the bonded interactions. In particular, the coefficients of such a revised model must be newly adapted to the data available at present. For additional, experimental support of this conclusion see ROTERMAN et al. [256].

6. Parameter estimation. Currently, the determination of the coefficients in a potential energy model is based on data obtained by one of the following methods:

– X-ray crystallography gives the equilibrium positions of the atoms in crystallized proteins (or rather their average over the high frequency vibrations, which introduces errors due to anharmonic effects). A basic text is GIACOVAZZO [106]; specifically for proteins see, e.g., ZANOTTI [357], FORTIER et al. [99].

– Nuclear magnetic resonance (NMR) spectroscopy, (see, e.g., JARDETZKY & LANE [157], TORDA & VAN GUNSTEREN [320]), gives position data of proteins in solution.

– *Ab initio* quantum mechanical calculations [80, 98, 196, 211, 236] give energies, energy gradients, and even energy Hessians (i.e., second derivative matrices) at arbitrarily selected positions of the atoms, for molecules in the gas phase;

– Measurements of energy spectra give rather precise eigenvalues of the Hessian;

– Thermodynamical analysis gives specific heats, heats of formation, conformational stability information, related to the potential in a more indirect way, via statistical mechanics.

The model parameters are adapted to data from one or several of these sources by using a mixture of least squares fitting and more heuristic or interactive procedures. Starting values for the coefficients come from general knowledge about atomic radii, average bond lengths and angles, and (for the force constants) from the frequencies of the oscillations of these quantities. The resulting rough parameters are then fitted to the data using a least squares approach.

The state of the art of numerical methods for least squares calculations is surveyed in BJÖRCK [22, 23] for the linear case. In addition, recent work by MATSTOMS [201] on multifrontal orthogonal factorizations for large and sparse least squares problems is relevant. The nonlinear case is reduced to the linear case, most commonly by means of damped Gauss-Newton steps, see, e.g., DENNIS & SCHNABEL [74], FLETCHER [97].

For a survey of parameter fitting procedures used in molecular mechanics see HOPFINGER & PEARLSTEIN [148]. More heuristic techniques are used to correct the parameters to match spectral data available, e.g., for the amino acids; see, e.g., [33, 175]. A detailed description of the development of a potential model is given by LIFSON [190]. From a mathematical point of view, the ab initio approach (with second derivative information) poses interesting questions about multidimensional Hermite interpolation (or approximation) and the optimal choice of trial points for the quantum mechanical calculations; since the least squares problems are here linear, traditional methods for the optimal design of experiments (see, e.g., ATKINSON [9], ATWOOD [10], PUKELSHEIM [241] and the books by FEDOROV [94] or SILVEY [286]) should lead to improved fits for the same amount of work.

The assessment of the sensitivity of the parameters with respect to the input data generally received very little attention. MAPLE et al. [196] mention the calculation of parameter uncertainties, but give no details on their magnitude. One reason for the importance of parameter uncertainties is that parameters with large uncertainties are unlikely to be transferable. THACHER et al. [318] discuss other applications of these sensitivities.

The sensitivity of equilibrium internal coordinates with respect to changes in the values of the parameters is discussed in SUSNOV [314].

Knowledge of such sensitivity information gives information on the quality of the model potential, and helps to assess the relative importance of the various terms in the potential; cf. RABITZ [243]. Together with the parameter uncertainties it also gives an idea of the accuracy obtainable in potential minimizations, and hence of the accuracy to which these minimizations are worth computing.

In principle, sensitivity information can be obtained together with the least squares calculation (see, e.g., DRAPER & SMITH [83]); but the improvements gained through the spectral information can be similarly assessed only when this information is combined with the position data available in a larger (and more nonlinear) least squares problem.

Ab initio potentials. Problems of a different kind appear in attempts to derive the potential from basic principles. Strictly speaking, the quantum mechanical ab initio potential is not the right potential to use in molecular mechanics applications since it gives as potential energy the quantity referred to by chemists as the *enthalpy* H. But the potential relevant for calculations at fixed finite temperature $T > 0$, the Gibbs free energy $G = H - TS$, has a correction term involving the *conformational entropy* S of the system. The conformational entropy is proportional to the logarithm of the number of microscopically distinguishable configurations belonging to an observed macrostate and is thus roughly proportional to the logarithm of the volume of the catchment region of a metastable state. Thus large flat minima (having a large catchment region) or large regions covered by many shallow local minima (corresponding to a glassy regime) are energetically more favorable than a narrow global minimum if this has only a slightly smaller potential.

Some early calculations are reported by FARNELL et al. [93]. To estimate the conformational entropy, CREIGHTON [62], p. 161, apparently only counts local minimizers, taking his intuition from simple random polymers where these minima can be assumed to be equidistributed with similar-sized catchment regions. For the much more complex energy surfaces of proteins, this assumption seems questionable. GO & SCHERAGA [118] (see also SCHERAGA [268], who surveys alternative approaches to entropy, too) use second derivative information on all nearly global minima (at $x^{(1)}, x^{(2)}, \ldots$) to approximate the conformational entropy of a minimum at $x^{(m)}$ by

$$S^m = -k_B \log p_m + k_B \log \sum_k p_k, \tag{12}$$

where

$$p_m = \frac{\exp(\frac{-V(x^{(m)})}{k_B T})}{\sqrt{\det \nabla^2 V(x^{(m)})}}. \tag{13}$$

(The $\log \sum$ term can be ignored since it shifts the entropy and hence the free energy by the same term for all minima.) This formula derives from statistical mechanics by replacing the potential near each nearly global minimum by its quadratic Taylor expansion. Anharmonicities (i.e., effects due to non-quadratic higher order terms) are thus not taken into account. Molecular dynamics simulations, started at a nearly global minimizer and exploring the neighborhood over a sufficiently long time interval, would in principle allow the determination also of anharmonic contributions to the conformational entropy. (Formula (12) makes only sense at the local minima themselves. For non-equilibrium configurations, one has to work with partition functions, that are more difficult to handle; cf. again [118].)

Currently, entropy considerations are often addressed computationally in a qualitative fashion only, using very simple (e.g., lattice) models together with techniques from statistical mechanics. A survey of typical results obtainable in that way, and many further references on the statistical mechanics approach, are given in WOLYNES [347].

Entropy of mixing with the solvent is another term that may play a role; see CHEN et al. [49]. See also ABAGYAN [1] for a discussion of the various terms that need to be added to the ab initio potential to get a more realistic potential.

For semiempirical potentials fitted to experimental data, these fine points appear somewhat less important in view of the other approximations made: the resulting potentials are always effective potentials adapted to the form of potential used and the experimental conditions from which the data are derived. However, this implies that caution is needed when combining data from different sources.

7. The native state. We concluded Section 3 with the remark that most experts expect that the geometry defined by the *global* minimum of the potential energy surface is the correct geometry describing the conformation observed in folded proteins. In the present section we look at the structure of the potential landscape. We also take a critical look at the statement that the folded state is given by the global minimum, and discuss some alternatives considered in the literature.

The most challenging feature of the protein folding problem is the fact that the objective function has a huge number of local minima, so that a local optimization is likely to get stuck in an arbitrary one of them, possibly far away from the desired global minimum. People working in the field expect an exponential number of local minima. Estimates I have seen range from 1.4^n to 10^n for a protein with n residues; the highest estimate is from CREIGHTON [62], p. 161. For very general energy minimization problems, combinatorial difficulty (NP-hardness) can be proved by showing that the traveling salesman problem can be phrased as a minimization of the sum of two-body interaction energies (WILLE & VENNIK [344]), but the potential is very contrived, and the result implies nothing for more realistic situations.

For less structured problems with more symmetries, like the problem of the optimal configuration for a cluster of n identical atoms with a Lennard-Jones interaction (11), the number of local minima appears to grow even more violently; an estimate by HOARE [139] is $O(1.03^{n^2})$. However, most of these local minima would have a large potential energy and thus be irrelevant for global optimization; the number of low-lying minima is more likely to grow simply exponential in n, too. A remarkable observation of HOARE & MCINNES [140] reveals that using in place of the Lennard-Jones pair interactions more realistic Morse functons gives cluster potentials with a much smaller number of local minima. Thus it may be that, also in the protein folding case, more realistic potentials will be easier to handle than current simpler models.

A bead model. To get a feeling for the origin of the exponential number of local minima, consider the following intuitive *bead model*, that ignores all non-local interaction and simulates the local interaction by a rubber band. Imagine a chain of n irregularly shaped beads (20 different kinds, corresponding to the amino acids) threaded along a rubber band knotted at one end and held at the other end. The rubber band tries to contract, and the beads arrange themselves in a way as to minimize the potential energy (tension). Now consider fixing the top i beads of the chain while rotating the $(i + 1)$st bead along the chain together with the rest of the chain below. Because of the irregular shape of the beads, the rotated bead and the bead above it move somewhat apart to allow the rotation, but after some local irregularity is overcome the beads can come closer together again. Thus the energy increases at first, passes a saddle point (a local maximum along the reaction coordinate), and moves towards a new local minimum. Depending on the shape of the two adjacent beads, there may be a different number m_i of locally optimal arrangements of the two beads. Noticing that rotations at different beads can be performed independently, we see that the total number of different local optima is $m_1 \cdot \ldots \cdot m_{n-1}$, and if each m_i is greater than one, this gives an exponential number of possibilities.

Of course, this model is very simplistic; interaction in a realistic molecule is much more complex and also non-local. But the model allows one to understand the qualitative origin of the large diversity of local minima possible and is likely to be realistic in this respect. So-called *genetic algorithms* for global optimization (see below) attempt to make use of the insight from such simple analogies by allowing mutations and crossing over between candidates for good local optima in the hope to derive even better ones.

The folded state and molten globule states. Observed in experiments are unique conformations in the folded state (to within a certain accuracy), independent of the history. This strongly suggests the existence of a unique global minimizer with a significantly lower energy than all other local minimizers. This is also supported by experiments that suggest that the approach to the global minimum proceeds in two phases, a rapid phase to reach a nearly folded state, followed by a lag period to complete the folding to the final state (Creighton [63]).

The most natural explanation is the existence of a large barrier with many (but not too many or too deep) saddles around the valley containing the global minimum. For example, Šali et al. [260] discuss a (lattice) scenario where the global minimum is well separated in energy from the other local minima, and the number of transition states leading to the global minimum is large, thus favoring the formation of the folded state from many less compact nearly folded states. They show that about 15% of randomly generated structures in their (simplistic) model had this property and indeed folded correctly within a reasonable number of Monte Carlo simulation steps. An apparently more reliable indicator of good folding is a low mean square distance between different configurations obtained in molecular simulations; see Irbäck et al. [153].

However, there are other possibilities: Quantum mechanical corrections accounting for vibrational zero-point energy might disfavor the global minimum state Mezey [205], Slanina [296]. The main effect is that a slightly non-global minimum in a broader valley may be more highly populated than a global minimum in a valley with steep walls; cf. Richards [250]. (Possibly taking account of these corrections is equivalent to the incorporation of the conformational energy; the effects are quite similar.) In some cases – as, e.g., for the molecule IHI – the ab initio potential surface does not even have a finite local minimum, and the observed metastable state is close to a saddle point of the potential (cf. [205], p.302). However, a remark in [296] suggests that so far there is no evidence that this difficulty occurs in organic molecules.

The folded state might be a metastable state with high energy barriers, or it might just be the lowest local minimizer that is kinetically accessible from most of the state space. The folded state may also correspond to more extended regions in state space where there are many close local minima of approximately the same energy as at the global minimum.

This last situation corresponds to what physicists refer to as *glassy* behavior (see the statistical dynamics treatment surveyed in WOLYNES [347] and references there), and there are some indications from molecular dynamics simulations (ELBER & KARPLUS [89], HONEYCUTT & THIRUMALAI [146]) that this might be the situation in typical energy surfaces of proteins. (Temperature effects also play a role, see ABKEVICH et al. [2].) On the other hand, calculations of IORI et al. [152], using a different model, arrive at the opposite conclusion.

CAMACHO & THIRUMALAI [43] find that, in lattice models, there appear to be exponentially many local minimum structures and exponentially many compact structures, but the number of compact local minimum structures seens to be nearly independent of the number of residues.

More recent studies (KARPLUS et al. [163], LEOPOLD et al. [185] ONUCHIC et al. [227], ŠALI et al. [260, 261, 262], WOLYNES et al. [348]; see also HAO & SCHERAGA [132] and a very informative survey by DILL et al. [78]) seem to reach an agreement in that the native state is a pronounced global minimizer that is reached dynamically through a large number of transition states by an essentially random search through a huge set of secondary, low energy minima (representing a glassy *molten globule* state), separated from the global minimum by a large energy gap. In simulations with crude lattice models (see, e.g., DILL et al. [78], pp. 594-595 for the merits and faults of this simplifying assumption), this scenario appears to be the necessary and sufficient conditions for folding in a reasonable time. (Crude off-lattice models also appear to confirm this statement; see IRBÄCK et al. [153, 154].)

It explains the so-called *Levinthal paradox* (LEVINTHAL [186]) that the time a protein needs to fold is by far not large enough to explore even a tiny fraction of all local minima only. However, fast collapse to a compact molten globule state, followed by a random search through the much smaller number of low energy minima could account for the observed time scales, and the energy gap provides the stability of the native state over the molten globule state. The time-limiting step is provided by the task to drive the molten globule into one of the transition states to the native geometry. Experimental characterizations of molten globule states are discussed in DOBSON [81] and MIRANKER & DOBSON [207].

On the other hand, the studies mentioned disagree on many of the details, and the simplified drawings of the qualitative form of the potential energy surface are mutually incompatible between different schools.

This shows that the interpretation of the literature on this point is difficult. At the present stage of development, unless the models are extremely simple, numerical calculations do not allow one to check reliably whether a global minimizer has been found; all that can be said is that the minimizers found were the best ones in an incomplete search. Discrepancies with experiments or with theoretical expectations might as well be interpreted as artifacts created by deficiencies of the potential used; the model accuracy is probably not higher than the distance between several near-global minima. And the papers with the more impressive simulations do not even claim that their (lattice) models represent more than only rough qualitative aspects of real proteins.

A serious limitation of many simulation studies is the fact that the presence of a solvent is poorly accounted for in the models used for folding simulations. Methods based on the Poisson-Boltzmann equation or the solvent-accessible surface are too expensive for long simulations. However, because of the tendency to form hydrogen bonds with water, hydrophilic residues involving charged or polar groups tend to be located at the surface, while the other, hydrophobic groups tend to be buried in the core of the molecule. Without solvent, the tendency is almost reversed since oppositely charged groups now tend to form complementary pairs neutralizing the Coulomb forces. Typical qualitative studies (e.g., Miyazawa & Jernigan [206]) therefore use a (drastic) simplification by employing just two (or three) kinds of residues, polar, hydrophobic (and indifferent).

Systematic studies of how the solvent changes the potential energy surface appear to be missing. However, a particularly striking illustration is given in Novotný et al. [225]. They show that a particular natural protein can be 'folded' by energy minimization (with the CHARMM potential, describing an isolated molecule) into nearly the shape of another protein; the resulting minimal structure had similar energy as the native structure. However, simple modifications of the potential energy function to take account of solvent screening and non-polar surface effects allow the correct discrimination between native and misfolded structure. Similarly, a computational study of LeGrand & Merz [183] reports severe distortions of the minimal energy configuration when the solvent is ignored.

From an evolutionary point of view, the hypothesis of a single, well separated global minimum well is also very likely. Indeed, in order that organisms can function successfully, the proteins performing specific tasks must fold into identical forms. Polypeptides that do not satisfy this requirement lack biological reliability and are not competitive. Thus one expects that at least the polypeptides realized as natural proteins have a single global minimum, separated from nearby local minima by a significant energy gap.

Recently, however, a number of proteins called *prions* were discovered that exist in two different folded states in nature (Prusiner [240]). The normal form appears to be a metastable minimum only, separated by a huge barrier from the sick 'scrapie' form in the global minimum. Under ordinary circumstances, only the metastable form is kinetically accessible from random states; but the presence of molecules in scrapie form acts as a catalyst that reduces the barrier enough to turn the normal form quickly into scapie form, too. Substitution of a few crucial amino acids (caused by mutations of the prion-coding genes) also reduces the barrier.

Another indication that the global minimum picture may be too simple is the existence of an organic compound that crystallizes into a 4:1 mixture of two different geometric conformations of the molecule (Dunitz et al. [86]), indicating two nearly global minima separated by a low energy barrier.

A study of insulin by Hua et al. [149, 150] also suggests that a family of similar but distinct low energy conformations form the biologically functional state.

8. Global optimization. In this section we review techniques and problems related to finding the global minimum and discuss some of the studies made concerning the local and global optimization of macromolecular potentials.

One of the obvious difficulties is that because of the high dimension and the expensive evaluation of the potential, even *local* optimization is slow. (For large molecules, this is the case even when the potential calculations are speeded up using fast multipole expansions [25, 120, 284] or potential cutoffs.) The fastest optimization methods, the *adopted basis Newton-Raphson* (ABNR) method and *truncated Newton* (TN) methods, employed, e.g., in CHARMM [33, 76], combine elements of Newton's method with reduced subspace techniques to reduce storage requirements and to limit the amount of work done at each iteration. A comparison of these methods developed by chemists with some of the recent methods for large scale optimization developed by the optimization community, in particular truncated Newton methods (NASH [217], SCHLICK & OVERTON [276], SCHLICK & FOGELSON [274]), is given in a recent survey article by SCHLICK [273]. For an adaptation of the truncated Newton optimization package TNPACK [274] to the molecular mechanics package CHARMM [33] see DERREUMAUX et al. [76].

However, since the objective function has a huge number of local minima, a local optimization is likely to get stuck before the global minimum is reached. Thus some kind of global search is needed to find the global minimum with some reliability. LEACH [180], SCHERAGA [268], and PARDALOS et al. [229] survey, from different perspectives, the different methods for global optimization that have been used so far on molecular conformation problems. Together with the proceedings of the workshop on global minimization of nonconvex energy functions by PARDALOS et al. [230], these give an up to date bibliography on this part of the literature. Online information on global optimization in general (including public domain software packages) can be found on the World Wide Web, e.g., at the address `http://solon.cma.univie.ac.at/~neum/glopt.html`.

Instead of describing technical details of the various methods (these vary from author to author, and can be found in the citations above), I want to present a vivid informal view of the most useful basic techniques, their strengths and weaknesses. The methods that are considered by the folding community most useful at present are simulated annealing, genetic algorithms, and smoothing methods. All three are based on analogies to natural processes where more or less global optima are reached.

Simulated annealing. Introduced by KIRKPATRICK [166], simulated annealing takes its intuition from the fact that the heating (annealing) and slowly cooling a metal brings it into a more uniformly crystalline state, that is believed to be the state where the free energy of bulk matter takes its global minimum. (Incidentally, even for the simplest potentials, it is still an unsolved problem whether this is indeed true with mathematical rigor. For some results in this direction, see RADIN & SCHULMANN [247]). The role of temperature is to allow the configurations to reach higher energy states with a probability given by Boltzmann's exponential law, so that they can overcome energy barriers that would otherwise force them into local minima. This is quite unlike line search methods and trust region methods on which good local optimization programs are based (see, e.g., GILL et al. [112]).

In its original form, the simulated annealing method is provably convergent (in a probabilistic sense) but exceedingly slow; various ad hoc enhancements make it much faster. In particular, except for simple problems, success depends very much on the implementation used. For results of simulated annealing techniques for protein structure prediction see, e.g., KAWAI [164], SHIN & JHON [285].

Genetic algorithms. Introduced by HOLLAND [143], genetic algorithms make use of analogies to biological evolution by allowing mutations and crossing over between candidates for good local optima in the hope to derive even better ones. At each stage, a whole population of configurations are stored. Mutations have a similar effect as random steps in simulated annealing, and the equivalent of lowering of the temperature is a rule for more stringent selection of surviving or mating individuals.

The ability to leave regions of attraction to local minimizers is, however, drastically enhanced by crossing over. This is an advantage if, with high probability, the crossing rules produce offspring of similar or even better fitness (objective function value); if not, it is a severe disatvantage. Therefore the efficiency of a genetic algorithm (compared with simulated annealing type methods) depends in a crucial way on the proper selection of crossing rules. The effect of interchanging coordinates is beneficial mainly when these coordinates have a nearly independent influence on the fitness, whereas if their influence is highly correlated (such as for functions with deep and narrow valleys not parallel to the coordinate axes), genetic algorithms have much more difficulties. Thus, unlike simulated annealing, successful tuning of genetic algorithms requires a considerable amount of insight into the nature of the problem at hand. For a more detailed discussion of genetic algorithms in general see DAVIS [73]; for applications to protein folding see, e.g., BRODMEIER & PRETSCH [32], KAWAI [164], LE GRAND & MERZ [183], SHINJHON [285].

Both simulated annealing methods and genetic algorithms are, in their simpler forms, easy to understand and easy to implement, features that invite potential users of optimization methods to experiment with their own versions. The methods often work, if only slowly, and lacking better alternatives, they are very useful tools for biochemists, where the primary interest is to find (near-)solutions *now*, even when the reliability is uncertain.

To make simulated annealing methods and genetic algorithms efficient, clever enhancements are essential. However, theoretical work on explaining the effectiveness of useful enhancements is completely lacking. I also haven't seen careful comparisons of the various options available and their comparative evaluation on standard collections of test problems.

Smoothing methods. First suggested by STILLINGER [306] and STILLINGER & WEBER [308], smoothing methods are based on the intuition that, in nature, macroscopic features are usually an average effect of microscopic details; averaging smoothes out the details in such a way as to reveal the global picture. A huge valley seen from far away has a well-defined and simple shape; only by looking more closely, the many local minima are visible, more and more at smaller and smaller scales. The hope is that by smoothing rugged potential energy surface, most or all local minima disappear, and the remaining major features of the surface only show a single minimizer. By adding more and more details, the approximations made by the smoothing are undone, and finally one ends up at the global minimizer of the original surface.

While it is quite possible for such a method to miss the global minimum (so that full reliability cannot be guaranteed, and is not achieved in the tests reported by their authors), a proper implementation of this idea at least gives very good local minima with a fraction of the function evaluations needed for the blind annealing and genetic methods. It should be possible to further increase the reliability of the methods by using a limited amount of global search in each optimization stage, though this appears not to have been done so far.

A conceptually attractive smoothing technique is the *diffusion equation method* (PIELA et al., [239], LI & SCHERAGA [189]), where the original potential function $V(x)$ is smeared out by artificial diffusion. The solution $V(x,t)$ of the diffusion equation

$$V_{xx}(x,t) = V_t(x,t)$$

with initial condition $V(x,0) = V(x)$, that can be solved explicitly if $V(x)$ is a linear combination of Gaussians in $\|x_i - x_k\|$, gets smoother and smoother as t gets larger; for large enough t, it is even unimodal. Thus $V(x,t)$ can be minimized by local methods when t is sufficiently large, and using the minimizer at a given t as a starting point for a local optimization at a smaller t, a sequence of local minimizers of $V(x,t)$ for successively smaller t is obtained until finally, with $t = 0$, a minimizer of the original potential is reached. Unfortunately, for general functions, smoothing is very expensive, and at present the methods are practically useful mainly when the potential is a sum of univariate functions of distances between atomic coordinates. With minor errors, the potentials currently used in protein modeling can be approximated in this way.

Positive and negative results for both Lennard-Jones clusters and oligopeptides are reported in KOSTROWICKI et al. [172, 173]. Modifications to take account of the rigid structure of bond lengths and bond angles are discussed in KOSTROWICKI & SCHERAGA [174]. Similar techniques for smoothing have been proposed and applied to potential energy surfaces by COLEMAN et al. [55], MOREÉ & WU [214, 215], SHALLOWAY [283] and WU [349].

A different smoothing technique is proposed in DILL et al. [79]. They construct a surrogate potential surface by fitting an underestimating function to known local optima, hoping that the global optimum lies near the minimizer of this surrogate function. A kind of hybrid method between simulated annealing and smoothing is the quantum mechanical annealing technique of STRAUB et al. [4, 192, 310].

While the above techniques are motivated by nature it is important to remember that processes in nature need not be the most efficient ones; at best it can be assumed to be efficient *given the conditions under which they have to operate.* (Much of our present technology has vastly surpassed natural efficiency, by unnatural means.) Even assuming that nature solves truly *global* optimization problems (a disputable assumption), simple lower estimates for the number of elementary steps – roughly corresponding to function evaluations – available to natural processes to converge are (in chemistry and in biology) in the range of 10^{15} or even more. Thus to be successful on the computers of today or the near future, we must find methods that are much faster, exploring the configuration space in a planned, intelligent way, not with a blind combination of chance and necessity. And the challenge is to devise methods that can be analyzed sufficiently well to guarantee reliablility and success.

Some other recent papers don't fit one of the above general paradigmas. BYRD et al. [39, 40, 41] and VAN DER HOEK [329] employ a mixed stochastic sampling, local optimization and global subset optimization strategy specially adapted to parallel computing. VAJDA & DELISI [327] use dynamic programming. BILLETER [20] works with an ellipsoid algorithm.

Branch and bound methods. Traditionally, see, e.g., NEMHAUSER & WOLSEY [220], branch and bound methods are the method of choice for solving global optimization problems of a combinatorial nature, formulated as mixed integer programs. Since protein folding problems have a combinatorial aspect, branch and bound methods appear to be suitable for this task as well. However, since the coordinates are continuous variables instead of discrete ones, the methods do not immediately extend to potential minimization.

Though at present limited to small oligopeptides, the branch and bound methods developed by MARANAS & FLOUDAS [197, 198, 199, 200] are potentially most interesting since they lead to lower bounds on the minimal energy. They are the only current methods that allow an assessment of the quality of the local minima obtained, and combined with the sufficient conditions for global minima derived in NEUMAIER

[222], they may allow one to actually prove global optimality of the best local optimizers obtained, and thus may remove the aura of ambiguity inherent in all 'global' calculations of the past.

In the worst case, branch and bound methods take an exponential amount of work; but while this might well be realistic for the case of Lennard-Jones clusters, there are signs that the situation is much better than worst case with potentials for proteins. Indeed, a recent study by BUTUROVIĆ et al. [38] abstracted from the continuous global optimization problem a cruder combinatorial version based on a reduction of the Ramachandran plots (explained in the next section) to a discrete set of points, and showed that an exhaustive search for the solution of the resulting combinatorial optimization problem was feasible for several small proteins. With more refined techniques for the construction of underestimating functions, there is hope for repeating the calculations without the need for such drastic simplifications. Many of the heuristic techniques used currently for searching the conformational space of molecules (see LEACH [180] and SAUNDERS et al. [265]) can be adapted to or combined with the branch and bound approach to take advantage of the structural insights of current chemistry.

Branch and bound methods will ultimately allow one not only to calculate the global minimum reliably, but also to find all local minima and saddle points within a certain energy margin of the global minimum. This is essentially the problem of finding all zeros of the nonlinear system $\nabla V(x) = 0$ that satisfy a constraint on the signature of the Hessian and on the value of V. Such problems can already be handled in low dimensions by branch and bound methods, combined with techniques for interval analysis (HANSEN [130], NEUMAIER [221]), and it should be possible to combine these techniques with the underestimation techniques of MARANAS & FLOUDAS [197, 198, 199]. Some results of a branch and bound method on oligopeptides are given in ANDROULAKIS et al. [5].

Since this would provide information about the low-lying transition states and metastable states, developing these branch and bound methods to work for higher-dimensional problems will allow one to study not only the folded equilibrium state but also the final stages of folding and the early stages of unfolding. (However, EKSTEROWICZ & HOUK [88] observed that semiempirical potentials derived from stable molecules may be not accurate enough for the calculation of transition states. See also ANET [6].)

Since no near global minimum will be missed, calculation of the entropy contributions (12), (13) will allow finding the global minimum of the Gibbs free energy, that is more likely to match the true folded state. See SAUNDERS et al. [265] for a performance study of several methods for finding most (or perhaps even all) low-lying local minima of a cycloheptadecane potential. (A huge number of near global minima – and a fortiori transition states, as expected for proteins if they exhibit glassy behavior, would, however, result in exponentially large space and time requirements.)

Constraints. In view of the expected huge number of local minima, techniques that reduce the size of the region in state space that must be searched for the global minimum are very valuable (HEAD-GORDON et al. [134, 135, 133], VAN DER GRAAF & BAAS [328]). Trivial restrictions that can be used are two-sided bounds on bond lengths and bond angles, and lower bounds on contact distances of arbitrary pairs. The bounds can be implemented either as hard bounds (via constraints) or as soft bounds (via penalty terms). A book by MOCKUS [209] discusses stochastic methods for constrained global optimization.

More generally, bounds on distances would allow one to combine the optimization approach to molecular modeling and the distance geometry approach surveyed in CRIPPEN [65], CRIPPEN & HAVEL [66] and BLANEY & DIXON [24]. Steps in this direction are discussed in CRIPPEN [64] and SCARSDALE et al. [266]; they assume that further distance information is known about the molecule, e.g., from nuclear magnetic resonance (NMR) spectroscopy.

In the absence of such detailed experimental information, the most likely candidates for a further significant reduction are constraints for the dihedral angles along the backbone. Indeed, biochemists have known for a long time experimentally that these dihedral angles are severely constrained. They discuss the pertaining information under the concept of *secondary structure* treated in the next section.

A little simpler is the constrained optimization problem known as side chain prediction. The task is to find the positions of the side chains of a protein, given the positions of the backbone atoms. See, e.g., LEE & SUBBIAH [182], TUFFÉRY et al. [324], DUNBACK & KARPLUS [84, 85].

The refinement of crystallographic data of limited (known) accuracy by means of a model potential, cf. BRÜNGER et al. [36], can also be treated as a global constrained optimization problem, and since the given data already strongly restrict the conformation space, this problem may even be solvable in high dimensions. The same holds for the construction of Cartesian coordinates from distance data obtained by nuclear magnetic resonance (NMR) spectroscopy (SCARSDALE et al. [266], TORDA & VAN GUNSTEREN [320]) by finding the global minimum of a weighted sum

$$V_{dc} = \sum_{\substack{i,k \\ i \neq k}} c_{ik}(\|x_i - x_k\|^{-3} - r_{ik}^{-3})^2$$

to find the Cartesian coordinates that best match a set of measured distances r_{ik}.

The related problem of finding the optimal superposition of two geometries for a molecule is discussed in KABSCH [160]. This technique is important since it allows the comparison of experimental coordinates of a molecule with those derived from computational predictions.

9. Simplified models. Various simplifications of the protein folding problem are studied in the literature in order to understand the global optimization process and to simplify the development and testing of optimization algorithms. The Lennard-Jones cluster problem (a problem also of independent interest in applications) has extra symmetry. Hence more specialized techniques can be used (see, e.g., [224, 139, 57, 351]) that allow one to reach rather low lying minimizers even for a huge number of atoms ([350] reports calculations for up to 10^5 atoms). If one adds constraints for fixed bond lengths one gets the models discussed in PARDALOS et al. [229]. Even simpler are the 'toy models' of STILLINGER et al. [307], the bead models of FERGUSON et al. [95], and the lattice models considered by PHILLIPS & ROSEN [238]. A survey of simplified models is given by TROYER & COHEN [321].

Some of the simplified models are accurate enough so that the resulting optimal structures resemble the real native structures, and still simple enough so that the global optimization by one of the methods discussed above appears feasible. Together with a final refinement of the structure by a local optimization, this is believed to give already useful structural information, though structural mistakes cannot be excluded.

These models fall into several classes:

Full energy models. Simplified full energy models have fixed bond lengths, bond angles and some torsion angles (e.g., around the peptide bond). At this level of approximation, many of the difficulties explained in the section on potentials become irrelevant. The only degrees of freedom are an independent set of torsion angles, which limits the number of variables to $\sim 3n - 5n$, where n is the number of residues. Such models are regarded as highly reliable; but function evaluation is expensive due to the required transformations between angles and Cartesian coordinates. The reconstruction of met-enkephalin (5 residuess, 19 torsion angles) is reported in KOSTROWICKI & SCHERAGA [173], using the diffusion equation method; in the latter case, a near-global minimum structurally similar to the native geometry is found.

Statistical backbone potential models. Here only the backbone (or the backbone and a side chain center, or even only the set of C_α atoms) is modeled, with fixed bond lengths, bond angles and peptide bond torsion angles (or fixed distant of neighboring C_α's, respectively). In both cases, the number of variables is reduced to $2n$, and the potential has a simple form, determined by assuming that a set of known structures is an equilibrium ensemble of structures, so that the energy can be calculated from Boltzmann's law and statistics on the known structures. In order to obtain a useful statistics, the protein structures used must be carefully selected; see, e.g., HOBOHM et al. [141]. A more detailed overview can be found in SIPPL [287]. Other statistical force field construction techniques are discussed in BAUER & BEYER [17] and ULRICH et al. [325]. The fact that the potential is now directly derived from geometric data implies that it automatically takes account of solvation and entropy corrections; on the other hand, one only gets a mean potential of less resolution. Reconstructions using such mean potentials are reported by SUN [313] for mellitin (26 residues), APPI (36 residues) and crambin (46 residues) using simulated annealing, by SUN [312] for mellitin and apamin (18 residues) using genetic algorithms, by GUNN et al. [123] for myoglobin (153 residues) using a combination of simulated annealing and genetic algorithms, and by SIPPL et al. [289] for lysozyme, myoglobin and thymosin. In all these papers, results are only 'native-like' when compared with the experimental structures.

Related techniques based on backbone models with terms involving information from Ramachandran plots (see next section) are discussed in DILL et al. [79]. As in earlier models (e.g., HONEYCUTT & THIRUMALAI [146]), the amino acids are treated in a reduced descriprion only, based on their hydrophilic or hydrophobic affinity.

Lattice models with contact potentials. Here the molecule is forced to have its atoms lying on lattice positions, and the potential is a sum of contact energies taken from tables derived again from statistics on databases. Now function evaluation is extremely cheap (addition of table entries for close neighbors only, resulting in a speedup factor of two order of magnitudes), and the problem has become one of combinatiorial optimization. The quality of a lattice model is mainly determined by its *coordination number*, the number of permitted sites for a C_α atom in a residue adjacent to a residue with a C_α atom in a fixed position. Models used to study qualitative questions of statistical mechanics (e.g., ŠALI et al. Sali2) usually use a nearest neighbor cubic lattice (with coordination number 6). While these are occasionally used for structure prediction (e.g., RABOW & SCHERAGA [244], who also report inadequacies of the Miyazawa-Jernigan-Cowell [59] contact potential), a good representation of $C_\alpha - C_\alpha$ distances and angles requires at least a face-centered cubic (FCC) lattice approximation (COVELL & JERNIGAN [59] with coordination number

42). More realistic approximations use a high coordination number of 56 (Kolinski et al. [171]) or 90, with a corresponding increase in combinatorial complexity which partially offsets the gain in evaluation speed. Godzik et al. [114, 115, 116] survey the various lattice models, discuss their merits and deficiencies, and report on some folding predictions. Another recent survey is Skolnick & Kolinski [294]. The broken translational and rotational symmetry causes problems when projecting a molecular geometry into a lattice, see Rykunov et al. [259].

However, Skolnick et al. [295] recover speed and motion invariance by using a three stage approach with an initial coarse lattice search, a fine lattice refinement phase and a final off-lattice optimization using full atom molecular dynamics. They report that compared with direct optimization of a full atom model, a factor of 100 in speed is gained.

All studies mentioned make smaller or larger structural mistakes, at least on some of the test examples. Therefore it is important to have independent ways of checking whether the structure found is likely to be the native structure or an artifact. Two recent studies addressing this with statistical potentials are Maiorov & Crippen [194], Hendlich et al. [136] and Casari & Sippl [45].

Threading. A different approach to ensure natural folding is to try to match known folding structures to amino acid sequences with unknown geometry. This process becomes more interesting as the database of available protein geometries becomes larger and more representative. (Because of its dependence on large databases, this approach is also termed 'knowledge-based'.) It is likely to be particularly effective if a *homologous* protein with a similar amino acid sequence – with mutations only in a few places – has a known fold; then there is a good chance that the new protein folds in a closely related way.

Various folding structures are tried in turn until one is found that makes some measure of fit (usually a statistical potential) small enough. The matching process, done in one of the above approximations followed by local optimization, is called *inverse folding* or *threading*; the latter name derives from intuition for the bead model, where the beads are threaded one by one onto a given wire frame. All reasonable fitting structures are then subjected to a stability test (using molecular dynamics or Monte Carlo simulation) in order to check the correct energetic behavior of the computed structures. (Inverse folding is also used to design protein sequences with a particular folding pattern; see, e.g., Yue & Dill [356].) The decision whether a fit is reasonable must again be based on statistical potentials.

Some papers describing achievements and problems in threading are Fetrow & Bryant [96], Goldstein et al. [117], Bowie & Eisenberg [29], Godzik et al. [114, 115], Johnson et al. [158], Jones & Thornton [159], Lathrop & Smith [179], Sippl et al. [290, 291] and Wodak & Rooman [346]. Several successful predictions have been reported; see, e.g., Crawford et al. [61], Bazan [18], Gerloff et al. [105] and Edwards & Perkins [87].

Lemer et al. [184] gives a critical evaluation of the quality of predictions in general. Zu-Kang & Sippl [360] discuss problems in identifying suitable superpositions of protein structures. For successful inverse folding of general sequences, there are still considerable obstacles to overcome, mainly because the irregular coil regions can have variable length, so that the structural match must reckon with insertions and deletions. Nevertheless, at present, inverse folding seems to be the most efficient way of structural prediction; and it will become more reliable as more and more proteins with known structure become available.

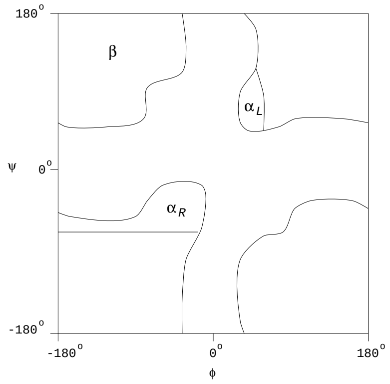

FIG. 7. **A Ramachandran plot.** *The cross-shaped region centered near* $(\varphi, \psi) = 0$ *is forbidden; the regions designated* α_L*,* α_R *and* β *correspond to left-handed helices, right-handed helices and sheets, respectively.*

However, there is a danger of misfolding a structure dissimilar to the known ones by forcing it to attain a known fold. Therefore the quest for finding reliably the global minimum from the potential only will continue. And the improvement of the potentials used to gauge the quality of a proposed fold will remain the key to a reduction of the error rate of current folding procedures.

Competitions. Competitions for (and evaluations of) the best prediction techniques are held every two years in Asilomar, California, USA; see the World Wide Web page `http://iris4.carb.nist.gov/casp2/` of the *Second Meeting on the Critical Assessment of Techniques for Protein Structure Prediction.*

10. Secondary structure. Since bond lengths, bond angles and side chain conformations are reasonably rigid, the dihedral angles determine the rough structure of the folded protein. Since the dihedral angle at the peptide bond is usually around 180^o (except for residues involving proline, where it often is around 0^o), it suffices for an overall view of the protein to know approximations to the sequence of angle pairs (φ_j, ψ_j) along the backbone.

Only a small subset of the set of all possible angle pairs is energetically favorable; most pairs are heavily penalized for steric (i.e., geometric) reasons, since atoms of the side chains are not allowed to come too close. The preferred regions are usually plotted in a so-called *Ramachandran plot*; see Figure 7. Only the smallest amino acid, Glycine, has larger preferred regions since its residue (a single hydrogen atom) is so small that steric effects are much milder.

Modulo 180^0, the amount by which one amino acid is twisted with respect to the previous one is approximately given by the sum $\varphi + \psi$. If several adjacent amino acids repeat roughly the same angle pair – a frequent situation in practice – then a regular local geometrical pattern emerges. These are usually referred to as *sheets* if

$$|\varphi + \psi| < \approx 30^o,$$

as *right-handed helices* if

$$\varphi + \psi < \approx -50^o,$$

and as *left-handed helices* if

$$\varphi + \psi > \approx 50^o$$

(assuming for simplicity that the peptide bond dihedral angle ω is $\approx 180^o$). Actually several types of sheets and helices can be distinguished by a more careful classification (see, e.g., CREIGHTON [62], Table 5-2).

Further geometrically recognizable local structures are *reverse turns* and *bends*, determined by appropriate (wide) ranges of two adjacent angle pairs ([62], Table 6-5). Other, irregular patterns are called *coils*. The assignment of such local geometrical patterns (i.e., one of the labels helix, sheet, turn, coil) to the amino acids in the amino acid sequence of a protein is referred to as the *secondary structure* of the protein. Dynamically, secondary structure is frequently formed in the early stages of folding.

Results from lattice computations by CHAN & DILL [46, 47] suggest that the formation of secondary structure is primarily due to the compact nature of folded proteins and not to the detailed form of the interaction. (But of course the detailed form will depend on it.) For some related results for tertiary structure see COVELL & JERNIGAN [59].

Given the secondary structure, it is still a highly nontrivial task to derive the tertiary structure, i.e., the full protein geometry. See, e.g., MONGE et al. [213].

Pattern recognition. There have been a number of attempts to predict the secondary structure assignment directly from the sequence of amino acids, using pattern recognition techniques (often neural networks) trained on proteins with known geometric structure; see, e.g., [100, 108, 144, 169, 177, 242, 245, 316]. (The many traditional statistical techniques for pattern recognition see, e.g., FUKUNAGA [102], YOUNG & FU [355], have hardly been tried.)

However, success was rather limited; see SCHULZ [282] and STOLORZ et al. [309] for critical evaluations of the state of affairs in 1988 and 1991, but also ROSE & CREAMER [254] for a more (and most likely too) optimistic perspective for the future. Two recent 1995 reviews of secondary structure prediction are BARTON [16] and RUSSELL & STERNBERG [257].

Currently, the best methods predict the correct assignment in only about 70% of the positions (ROST & SANDER [255]). The missing 30% are, according to current interpretation, mainly due to global influences; a simple estimate in [282] shows that most subsequences of 8 or fewer residues must be expected to give rise to more than one possible secondary structure sequence. A way to model some global influence is by minimizing statistical potential functions where the residues are considered as unstructured units (OOBATAKE & CRIPPEN [228], CRIPPEN & SNOW [67]); it should be possible to combine this approach with pattern recognition techniques. Another source of ambiguity is the lack of clear delineations of precisely which sets of geometries are classified to the various secondary structure labels; borderline geometries are interpreted differently by different experts. Also, the core of helices and their ends behave differently, the latter having a more varying range of angles associated with it.

The high error rate of 30% makes it impossible to even roughly determine the shape of the protein in a reliable way; see, e.g., GARNIER [104]. Moreover, even given the backbone positions, the reconstruction of the full geometry of a protein is still a nontrivial global optimization task (LEE & SUBBIAH [182], DUNBACK &

KARPLUS). Secondary structure considerations have been, however, more successful in the converse problem of designing residue sequences that produce tertiary structures of prescribed form; see, e.g., REGAN [249].

But it is possible that the bad results reported are rather due to a too restrictive view of how to model local structure. The typical method studied in the past tries to predict the ith member $S_i \in \{\text{helix, sheet, turn, coil}\}$ of the secondary structure sequence $\ldots S_{i-2}\ S_{i-1}\ S_i\ S_{i+1}\ S_{i+2}\ldots$ from a small local section (typically with 10-25 residues) of the amino acid sequence $\ldots R_{i-2}\ R_{i-1}\ R_i\ R_{i+1}\ R_{i+2}\ldots$ ($R_i \in \{\text{Ala},\ldots,\text{Val}\}$); different methods differ in details of how they modify this basic approach (e.g., by grouping the amino acids into classes of similar ones, by encoding the amino acids in different ways numerically, by a posteriori adaptation of the sequence of secondary structure labels, etc.).

Local statistical dependencies. However, if one looks at typical problems in physics with local and global aspects, e.g., finding the shape of a string fixed at both ends and loaded with different weights (an analogy to the different amino acids), one finds that the shape of the string is only very loosely determined by the sequence of nearby weights, since the location of the end points has a large influence on the shape. On the other hand, some *relations* between neighboring positions are determined completely by local information.

Of course, this analogy is very superficial, but it suggests the following natural way of capturing some of the secondary structure information. Indeed, there seem to be enough data available in the Brookhaven Protein Data Bank to be able to look for local statistical dependencies of the form

$$\cos\varphi_i \ (\text{or } \sin\varphi_i) \approx \Phi_{R_{i-1},R_i}(\psi_{i-1},\psi_i) \tag{14}$$

and

$$\cos\psi_i \ (\text{or } \sin\psi_i) \approx \Psi_{R_{i+1},R_i}(\varphi_{i+1},\varphi_i) \tag{15}$$

with suitable trigonometric polynomials Φ, Ψ in two variables depending on two adjacent residues.

While this is not enough to fix the secondary structure, it would be very useful information that, together with suitable boundary conditions, has a chance to fix the rough global structure of the protein. It also avoids the problem of ambiguity in the classification of borderline geometries mentioned above. A step in this direction is taken by KANG et al. [161]

Moreover, and independently of this, conditions like (14) and (15), coupled with realistic error bounds, would be very useful as constraints in global optimization routines, since they would drastically reduce the size of the region in state space that must be searched for the global minimum. Evidence for the potential effectiveness of such an approach is the paper by MONGE et al [213]. They freeze assumed secondary structure to limit the number of degrees of freedom, and they discretize the dihedral angles along the backbone turns by selecting one of six particular angle pairs. Then a Monte Carlo search method is used to find good minimizers which resemble the true folded geometry. A constrained approach would yield a more realistic version in place of the frozen secondary structure, and a branch and bound methodology would yield a more satisfactory and adaptive way of handling the unconstrained turns.

11. Conclusions. We discussed the physical and chemical background of protein folding, focussed on the mathematical models used, surveyed the prospects for global optimization of potential energy functions with many local minima, and looked at the various approximations needed to make the problem tractable.

The discussion of the various problems involved in protein folding revealed a number of weaknesses in the current models and in the schemes used for their analysis. Our presentation also indicated a number of new ideas that might overcome some of these weaknesses and ultimately lead to quantitatively predictive models within the limits of present-day computational power.

The importance of the topic, the many open questions, the intricacies of modeling, and the challenging computational aspects of protein folding make the subject a paradise not only for researchers in biochemistry but, we hope, also for applied mathematicians and numerical analysts.

Appendix 1: On the choice of the dynamic energy function. We present an improved formulation of the dynamic energy minimization approach of ZHANG & SCHLICK [358, 359] to the solution of the dynamical system

$$M\ddot{x} + C\dot{x} + \nabla V(x,t) = 0 \tag{16}$$

with initial conditions

$$x(t_0) = x^0, \quad \dot{x}(t_0) = \dot{x}^0.$$

We assume that M and C are positive semidefinite, and, for each t, $V(\cdot,t)$ is twice continuously differentiable and bounded from below. (The time-dependent character of the potential allows us later to include the stochastic case as well.)

Each time step consists in the calculation of an approximation $\hat{x}$ to $x(t)$ at $t = t_{l+1} := t_l + h_l$, but, clearly, it suffices to look at the case $t = t_0 + h$, thus saving many indices. The key idea of the new approach is to determine $\hat{x}$ by minimizing

$$\tilde{V}(x) := V(x,t) + \frac{1}{2h^2}(x-z)^T N(x-z) \tag{17}$$

for a suitable symmetric matrix N and a simpler approximation z to x. In the time independent case, (17) is motivated by the fact that for small h the second term dominates and forces $\hat{x} = z + O(h)$, while for $h \to \infty$ the second term drops away and a local minimizer of V is approached. Thus (17) captures the qualitative behavior expected from a solution to (16).

N and z will be chosen to make the resulting approximation $\hat{x}$ accurate of high order when h is small. Following ZHANG & SCHLICK, we first solve a simpler approximate problem

$$M\ddot{y} + C\dot{y} + \nabla W(y,t) = 0, \tag{18}$$

$$y(t_0) = x^0, \quad \dot{y}(t_0) = \dot{x}^0, \tag{19}$$

where W is an approximate potential satisfying the consistency condition

$$\nabla W(x^0, t^0) = \nabla V(x^0, t^0). \tag{20}$$

A typical choice is

$$W(y,t) = \nabla V(x^0,t^0)^T(y-x^0) + \frac{1}{2}(y-x^0)^T H(y-x^0)$$

for a suitable simplified Hessian H. Since ∇W is linear, (18)-(19) can be solved exactly; see below.

For computational reasons, H is chosen as block-diagonal or banded matrix (i.e., $H_{ik} = 0$ except when $|i-k|$ is small), so that the spectral factorization of H is easy to obtain. However, to get a sufficiently good approximate problem, we must choose H in such a way that it captures the dominant entries from the Hessian $G = \nabla^2 V(x^0, t^0)$, and hence the high-frequency behavior of the system (16). More precisely, $M^{-1/2}HM^{-1/2}$ should approximate $M^{-1/2}GM^{-1/2}$ since the spectrum of the latter determines the short term motion.

Noting that if (17) is minimized at $\hat{x}$ then $\nabla V(\hat{x}, t) + h^{-2}N(\hat{x} - z) = 0$, or

$$z = \hat{x} + h^2 N^{-1} \nabla V(\hat{x}, t), \tag{21}$$

we see that we would get $\hat{x} = x$ for the choice $z = x + h^2 N^{-1} \nabla V(x, t)$. Since x is unknown, we replace in this equation the true potential V by the approximation W and the unknown x by the known $y = y(t)$. Thus we define

$$z := y + h^2 N^{-1} \nabla W(y, t), \tag{22}$$

where

$$t = t_0 + h, \quad y = y(t). \tag{23}$$

The choice $N = M + hC$ corresponds to the method advocated in ZHANG & SCHLICK [358, 359] (who consider only the case $C = \gamma M$, M definite) derived form the implicit Euler step and hence gives an approximation $\hat{x} = x + O(h^3)$ (but only $x + O(h^2)$ if $M = 0$). However a different choice gives better accuracy:

THEOREM 11.1. *(i) If M is positive definite then any choice*

$$N = 6M + O(h) \tag{24}$$

gives an approximation

$$\hat{x} = x + O(h^4) \tag{25}$$

and hence a global error of $O(h^3)$ over long time intervals.

(ii) If $M = 0$ and C is positive definite then any choice

$$N = 2hC + O(h^2) \tag{26}$$

gives an approximation

$$\hat{x} = x + O(h^3) \tag{27}$$

and hence a global error of $O(h^2)$ over long time intervals.

Proof. (i) By (19) and (20), a comparison of (16) and (18) shows that $\ddot{y}(t_0) = \ddot{x}(t_0)$, whence

$$y - x = \frac{h^3}{6}(\dddot{y}\,(t_0) - \dddot{x}\,(t_0)) + O(h^4).$$

Using (24), we deduce

$$\begin{aligned} N(y-x) &= 6M(y-x) + O(h^4) = h^3(M\,\dddot{y}\,(t_0) - M\,\dddot{x}\,(t_0)) + O(h^4) \\ &= h^3 \frac{d}{dt}(M\ddot{y} - M\ddot{x})\Big|_{t=t_0} = h^3 \frac{d}{dt}(-\nabla W + \nabla V)\Big|_{t=t_0}. \end{aligned}$$

Now (21) gives

$$\begin{aligned}
\hat{x} &= z - h^2N^{-1}\nabla V(\hat{x},t) = y - h^2N^{-1}(\nabla V(\hat{x},t) - \nabla W(y,t)) \\
&= y - h^2N^{-1}\left(h\frac{d}{dt}(\nabla V - \nabla W)\Big|_{t=t_0} + O(h^2)\right) \\
&= y - (-(y-x) + O(h^4)) + O(h^4) = x + O(h^4).
\end{aligned}$$

(ii) If $M = 0$ then (16), (18) and (19) force $\dot{y}(t_0) = \dot{x}(t_0)$ (that can now no longer be prescribed freely), whence

$$y - x = \frac{h^2}{2}(\ddot{y}(t_0) - \ddot{x}(t_0)) + O(h^3).$$

Using (26), we deduce

$$\begin{aligned}
N(y-x) &= h^3C(\ddot{y}(t_0) - \ddot{x}(t_0)) + O(h^4) \\
= h^3\frac{d}{dt}(C\dot{y} - C\dot{x})\Big|_{t=t_0} + O(h^4) & \\
&= h^3\frac{d}{dt}(-\nabla w + \nabla v)\Big|_{t=t_0} + O(h^4),
\end{aligned}$$

and as before we get $\hat{x} = x + O(h^3)$. A power of h is lost since $N = O(h)$. Finally, the statements about the global errors follow along standard lines. □

We conjecture that the choice

$$N = 6M + 2hC \tag{28}$$

which gives the approximation (25) if M is definite and (27) if $M = 0$ will be a good choice that gives good long time results even in the case of strong damping $C \gg M$ (when, after rescaling the time, a singularly perturbed version of (ii) results, and the choice of an arbitrary N with (24) would be accurate only for times of the order of $C^{-1}M$).

Note that the theorem is valid for *all* consistent choices of W; however, this choice affects the time scale on which the asymptotic result is valid. In order to be able to use large time steps, $\nabla^2 W$ should account well for all absolutely large eigenvalues of $\nabla^2 V$, thus eliminating them from the corrective dynamics handled by the minimization.

We now look at the solution of (18), (19) for the choice

$$W(y,t) = g^T(y - x^0) + \frac{1}{2}(y - x^0)^TH(y - x^0), \tag{29}$$

with a symmetric matrix H, in principle arbitrary but in practice an easily factorizable approximation to $\nabla^2 V(x_0, t_0)$. Using a spectral factorization

$$M^{-\frac{1}{2}}HM^{-\frac{1}{2}} = Q\Lambda Q^T, \quad Q^TQ = I, \quad \Lambda \text{ diagonal} \tag{30}$$

and

$$S := M^{-\frac{1}{2}}Q, \tag{31}$$

we have

$$HS = MS\Lambda, \ (MS)^{-1} = S^T. \tag{32}$$

Hence multiplication of (29) with S^T and use of the new variable u with

$$y = x^0 + Su$$

reduces the system (18) to the form

$$\ddot{u} + \Gamma\dot{u} + \Lambda u + h = 0, \quad h = S^T g,$$

where $\Gamma = S^T CS$. In molecular dynamics simulations, where we have little information about C anyway, the choice

$$C = \gamma M \tag{33}$$

has been repeatedly used in the literature. For this choice, $\Gamma = \gamma I$, and we get the decoupled system

$$\ddot{u} + \gamma\dot{u} + \Lambda u + h = 0, \quad h = S^T g \tag{34}$$

that is easy to solve. (For other choices of C, one would need to solve a quadratic eigenvalue problem in place of (30) in order to diagonalize the system, recast in first order form.)

In the case of a stochastic differential equation we have

$$V(x,t) = V(x) - \varepsilon(t)^T(x - x_0),$$

where

$$\langle \varepsilon(t)\varepsilon(t)^T dt\rangle = 2k_B T\ C.$$

It is then natural to keep this stochastic linear term also in W, thus using $g - \varepsilon(t)$ in place of g. We find in place of h the transformed forcing term

$$S^T(g - \varepsilon(t)) = h - \eta(t)$$

with a random term η satisfying the relation

$$\langle \eta(t)\eta(t)^T dt\rangle = \langle S^T\varepsilon(t)\varepsilon(t)^T S dt\rangle = 2k_B T\ S^T CS.$$

Hence, assuming again (33), we find

$$\ddot{u} + \gamma\dot{u} + \Lambda u + h = \eta(t), \quad h = S^T g, \tag{35}$$

where

$$\langle \eta(t)\eta(t)^T dt\rangle = 2\gamma k_B T\ I.$$

Thus the system is again fully decoupled and is easy to solve; for the details see Zhang & Schlick [359].

Using (22) and (23), we find from $u = u(t)$ the vector

$$z = y + h^2 N^{-1}(g + H(y - x^0)) = y + \frac{h^2}{6 + 2h\gamma} M^{-1}(g + HSu).$$

Hence, using (31) and (32),

$$z = y + \frac{h^2}{6 + 2h\gamma} S(\Lambda u + h), \tag{36}$$

and in the stochastic case (35)

$$z = y + \frac{h^2}{6 + 2h\gamma} S(\Lambda u + h - \eta(t)). \tag{37}$$

Since the formulas (34), (36) (or (35), (37) in the stochastic case) no longer depend on H, they remain valid even when only an approximate spectral factorization of $M^{-1}HM^{-1}$ (but with orthogonal Q) is used. This is equivalent to using instead of the original H the matrix $H = MS\Lambda(MS)^T$ for which the spectral factorization (30) is valid.

Moreover, due to the fact that the matrix S is only used to calculate two matrix vector products, the spectral matrix Q need not be available explicitly. Instead it can be held as a product of reflections and rotations, as obtained from the cheaper eigenvalue calculation. (This is usually done by tridiagonalization and subsequent use of the implicit QR algorithm; see e.g. PARLETT [231]. There is a trade-off between storing the rotations from the QR iteration or recomputing them at the time of calculating S^Tg and later again for $S(...)$.) This also assures that the linear mapping Q determined in this way remains essentially orthogonal even in finite precision arithmetic.

Appendix 2: Solvation energy and combination rules. We give here some arguments suggesting that by modifying the combination rules for the pair potential parameters, a large part of the solvation effects can be taken into account automatically.

Suppose we have a molecule with coordinate vector x in a solvent whose molecular position coordinates are part of the vector x_{solv}. The combined system is governed by a potential $V_{tot}(x, x_{solv})$, and a natural way to eliminate the solvent from consideration is to look at the reduced potential

$$V(x) := \min_{x_{solv}} V_{tot}(x, x_{solv}). \tag{38}$$

Thus we assume that at given molecule position, the solvent molecules take the positions at the global minimum of the total energy, which amounts to looking at the molecule in a completely frozen solvent. If we could find a phenomenological description of the reduced potential $V(x)$, the solvent would not need to be considered explicitly.

WESSON & EISENBERG [343] suggest that the solvation energy could be approximately modeled by adding to the potential $V_{mol}(x)$ of the molecule in isolation additional terms proportional to the surface area exposed to the solvent. We modify his approach and argue that it might be feasible to account for much of the solvation energy by means of suitable corrections to the pair potentials in the force field, and by suitable adjustments of the combination rules for their parameters.

Indeed, suppose we have two atoms i and k with atomic radii R_i and R_k at distance r_{ik}. Seen from the nucleus of atom i, atom k appears under an angle α given by $\sin(\alpha/2) = R_k/r_{ik}$, and the corresponding cone cuts out on the surface of atom i

a cap of area

$$A_{ik} = 2\pi R_i^2 \left(1 - \frac{1}{\sqrt{1 + (R_k/r_{ik})^2}}\right).$$

Since the total surface area is $A_i = 4\pi R_i^2$, we find

$$A_{ik}/A_i = \frac{1}{2}\left(1 - \frac{1}{\sqrt{1 + (R_k/r_{ik})^2}}\right) =: \mu(r_{ik}/R_k)$$

as an approximation to the surface fraction of atom i excluded by atom k, at least when r_{ik} is not large. For large r_{ik}, one would have to change μ so that it decays at least like the inverse 6th power since particles are not expected to have a larger influence over large distances.

The total excluded fraction can now be approximated by summing over all atoms $k \neq i$, and the solvation energy contributed by atom i could be modeled by a multiple of the difference between this sum and 1, hence by

$$E_{solv,i} = \sigma_i \left(1 - \sum_{k\neq i} \mu\left(\frac{r_{ik}}{R_k}\right)\right),$$

where σ_i is the solvation energy of an isolated atom i. Apart from constant terms that don't affect forces and energy differences, a natural form for a phenomenological total solvation energy term is therefore

$$E_{solv} = -\sum_i \sigma_i \sum_{k\neq i} \mu(r_{ik}/R_k), \tag{39}$$

for suitable constants σ_i and a suitable quickly decaying function μ.

Hydrophilic and hydrophobic behavior. We now show that for a simple model situation, a solvation term (39) is indeed sufficient to model hydrophilic and hydrophobic effects. We consider a hypothetical situation where we have atoms of two kinds A and B that interact by two-particle forces only, given by a pair potential $W(r)$. The particles are assumed to have the same unit radius and identical behavior with respect to each other, irrespective of their kind. However, their behavior with respect to the solvent (water) is assumed to be opposite: The A's are *hydrophobic*, i.e., a larger excluded surface (less surface exposed to water) decreases the solvation energy. On the other hand, the B's are *hydrophilic*, i.e., a smaller excluded surface (more surface exposed to water) decreases the solvation energy. In our model (39), this is achieved by the choice $\sigma_i = \sigma_A > 0$ for hydrophobic atoms and $\sigma_i = \sigma_B < 0$ for hydrophilic ones. Thus the potential energy (including the solvation terms) is

$$V = \sum_{x\in A} V_A(x) + \sum_{x\in B} V_B(x), \tag{40}$$

where, for simplicity, A and B now also denote the set of position vectors of atoms of type A and B, respectively, and

$$V_A(x) = \sum_{y\neq x} \left(\frac{1}{2}W(\|y-x\|) - \sigma_A\mu(\|y-x\|)\right), \tag{41}$$

(42) $$V_B(x) = \sum_{y\neq x} \left(\frac{1}{2}W(\|y-x\|) - \sigma_B\mu(\|y-x\|)\right).$$

THEOREM 11.2. *In any global minimum configuration,*

(43) $$\min_{x\in A} D(x) \geq \max_{x\in B} D(x),$$

where

$$D(x) = \sum_{y\neq x} \mu(\|y-x\|).$$

In other words, the hydrophobic atoms occupy the positions with largest values of $D(x)$.

Proof. Consider a global minimum configuration, and suppose we swap the places of a single pair of atoms in positions $x_1 \in A$ and $x_2 \in B$. Then the potential (40) changes into a potential

$$\begin{aligned} V' &= V - V_A(x_1) - V_B(x_2) + V_A(x_2) + V_B(x_1) \\ &= V + (V_A(x_2) - V_B(x_2)) - (V_A(x_1) - V_B(x_1)) \\ &= V - (\sigma_A - \sigma_B)(D(x_2) - D(x_1)). \end{aligned}$$

Now $\sigma_A > 0 > \sigma_B$ and $V' \geq V$ since we started at a global minimum. The formula for V' therefore implies that $D(x_2) \leq D(x_1)$. Since $x_1 \in A$ and $x_2 \in B$ were arbitrary, the theorem follows. □

Now if $\mu(r)$ is a sufficiently fast decaying function, the dominant contributions to $D(x)$ are those by the atoms closest to the atom at x. Therefore, in a large molecular cluster, $D(x)$ is largest far away from the surface of the molecule and smallest at the surface. Thus the hydrophobic atoms occupy the interior of the cluster, and the hydrophilic atoms are found at the surface. Thus the simple model correctly predicts the structure expected from purely qualitative reasoning, and the result still leaves much room for the precise form of the solvation potential. Of course, in more complex situations, various forces compete for their influence on the shape of the cluster, and may cause modifications of this qualitative picture.

We can give the potential (40) a different interpretation by noting that the solvation terms just act as asymmetric corrections to the pair interactions: If we introduce the pair potentials

$$V_{AA}(r) = W(r) - 2\sigma_A\mu(r),$$

$$V_{BB}(r) = W(r) - 2\sigma_B\mu(r),$$

$$V_{AB}(r) = W(r) - (\sigma_A + \sigma_B)\mu(r),$$

we can write the total potential in the traditional form

$$V(x) = \sum_{i<k} V_{ik}(\|x_i - x_k\|),$$

where

$$V_{ik} = \begin{cases} V_{AA} & \text{if both atoms } i \text{ and } k \text{ are of type } A, \\ V_{BB} & \text{if both atoms } i \text{ and } k \text{ are of type } B, \\ V_{AB} & \text{otherwise.} \end{cases}$$

Moreover, from the construction, we see that the new pair potentials satisfy the combination rule

$$V_{AB} = \frac{1}{2}(V_{AA} + V_{BB}). \tag{44}$$

In particular, if we think of $W(r)$ as a Lennard-Jones potential (11) and of $\mu(r) = \text{const} \cdot r^{-6}$ (a cheap choice for μ; the small-distance singularity does not matter since the repulsion term dominates there), then the V_{ik} are also Lennard-Jones potentials, but with modified radii, and with a combination rule different from what we would expect from the unsolvated case. However, this suggests that, by fitting the radii and the combination rules to experimental data instead of deriving them from geometric considerations, we might be able to catch a large part of the solvation energy without the need for explicit solvation energy terms. Essentially we are fitting directly the reduced potential (38) with the data at hand, and for this we need more flexible combination rules.

REFERENCES

[1] R. A. Abagyan, *Towards protein folding by global energy optimization*, FEBS Letters 325 (1993), pp. 17–22.

[2] V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, *Free energy landscapes for protein folding kinetics: intermediates, traps, and multiple pathways in theory and lattice model simulations*, J. Chem. Phys. 101 (1994), pp. 6052–6062.

[3] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford Univ. Press, New York 1990.

[4] P. Amara, J. Ma and J. E. Straub, *Global minimization on rugged energy landscapes*, in Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos et al., eds., Amer. Math. Soc., Providence, RI, 1996, pp. 1–15 .

[5] I. P. Androulakis, C. D. Maranas and C. A. Floudas, *Global minimum potential energy conformations of oligopeptides*, Manuscript (1995). (floudas@zeus.princeton.edu)

[6] F. L. Anet, *Inflection points and chaotic behavior in searching the conformation space of cyclononane*, J. Am. Chem. Soc. 112 (1990), pp. 7172–7178.

[7] N. L. Allinger, Y. H. Yuh and J.-H. Lii, *Molecular mechanics. The* MM3 *force field for hydrocarbons. 1–3*, J. Amer. Chem. Soc. 111 (1989), pp. 8551–8566, 8566-8575, 8576-8582.

[8] A. W. Appel, *An efficient program for many-body simulation*, SIAM J. Sci. Stat. Comp. 6 (1985), pp. 85–103.

[9] A. C. Atkinson, *Developments in the design of experiments*, Internat. Statist. Rev., 50 (1982), pp. 161–177.

[10] C. L. Atwood, *Optimal and efficient designs of experiments*, Ann. Math. Statist., 40 (1969), pp. 1570–1602.

[11] B. M. Axilrod and E. Teller, *Interaction of the van der Waals' type between three atoms*, J. Chem. Phys. 11 (1943), pp. 299–300.

[12] L. M. Balbes, S. W. Mascarella and D. B. Boyd, *A perspective of modern methods in computer-aided drug design*, in Reviews in Computational Physics, Vol. V, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1994, pp. 337–379.

[13] R. E. Bank, R. Bulirsch, H. Gajewski and K. Merten, eds., *Mathematical Modelling and Simulation of Electrical Circuits and Semiconductor Devices*, Birkhäuser, Basel 1994.

[14] E. Barth, K. Kuczera, B. Leimkuhler and R. D. Skeel, *Algorithms for constrained molecular dynamics*, J. Comp. Chem. 16 (1995), pp. 1192—1209.

[15] E. Barth, M. Mandziuk and T. Schlick, *A separating framework for increasing the time step in molecular dynamics*, in Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications, W. F. van Gunsteren et al., eds., ESCOM, Leiden, The Netherlands, 1996 (in press).

[16] G. J. BARTON, *Protein secondary structure prediction*, Curr. Opinion Struct. Biol. 5 (1995), pp. 372–376.
[17] A. BAUER AND A. BEYER, *An improved pair potential to recognize native protein folds*, Proteins: Struct. Funct. Gen. 18 (1994), pp. 254–261.
[18] J. F. BAZAN, *Structural Design and Molecular Evolution of a Cytokine Receptor Superfamily*, Proc. Natl. Acad. Sci. USA 87 (1990), pp. 6934–6938.
[19] F. C. BERNSTEIN, T. F. KOETZLE, G. J. WILLIAMS, E. MEYER, M. D. BRYCE, J. R. ROGERS, O. KENNARD, T. SHIKANOUCHI AND M. TASUMI, *The protein data bank: A computer-based archival file for macromolecular structures*, J. Mol. Biol. 112 (1977), pp. 535–542.
[20] M. BILLETER, T. F. HAVEL AND K. WÜTRICH, *The ellipsoid algorithm as a method for the determination of polypeptide conformations from experimental distance constraints and energy minimization*, J. Comp. Chem. 8 (1986), pp. 132–141.
[21] M. BISHOP AND S. FRINKS, *Error analysis in computer simulations*, J. Chem. Phys. 87 (1987), pp. 3675–3676.
[22] A. BJÖRCK, *Solution of equations in* $\mathbb{R}^n$, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., North Holland, Amsterdam 1990, pp. 465–652.
[23] A. BJÖRCK, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia 1996.
[24] J. M. BLANEY AND J. S. DIXON, *Distance geometry in molecular modeling*, in Reviews in Computational Physics, Vol V, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1994.
[25] J. A. BOARD, J. W. CAUSEY, T. F. LEATHRUM JR., A. WINDEMUTH AND K. SCHULTEN, *Accelerated molecular dynamics simulation with the parallel fast multipole algorithm*, Chem. Phys. Lett. 198 (1992), pp. 89–94.
[26] B. BORŠTNIK, D. PUMPERNIK, D. JANEŽIČ AND A. AŽMAN, *Molecular dynamics studies of molecular interactions*, Chapter 7 in Molecular Interaction, H. Ratajczak and W. J. Orville-Thomas, eds., Wiley, New York 1980.
[27] G. BOSSIS, B. QUENTREC AND J. P. BOON, *Brownian dynamics and the fluctuation-dissipation theorem*, Mol. Phys. 45 (1982), pp. 191–196.
[28] J. P. BOWEN AND N. L. ALLINGER, *Simplified models for understanding and predicting protein structure*, in Reviews in Computational Chemistry, Vol. II, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1991, pp. 57–80.
[29] J. U. BOWIE AND D. EISENBERG, *Inverted protein structure prediction*, Curr. Opinion Struct. Biol. 3 (1993), pp. 437–444.
[30] C. BRANDEN AND J. TOOZE, *Introduction to Protein Structure*, Garland, New York 1991.
[31] K. E. BRENAN, S. L. CAMPBELL AND L. R. PETZOLD, *Numerical Solution of Initial Value Problems in Differential-Algebraic Equations*, North-Holland, New York 1989.
[32] T. BRODMEIER AND E. PRETSCH, *Application of genetic algorithms in molecular modeling*, J. Comp. Chem. 15 (1994), pp. 588–595.
[33] B. R. BROOKS, R. BRUCCOLERI, B. OLAFSON, D. STATES, S. SWAMINATHAN AND M. KARPLUS, CHARMM: *a program for macromolecular energy, minimization, and dynamics calculations*, J. Comp. Chem. 4 (1983), pp. 187–217.
[34] C. L. BROOKS, M. KARPLUS AND B. M. PETTITT, *Proteins: a Theoretical Perspective of Dynamics, Structure, and Thermodynamics*, Adv. Chem. Phys. 71, Wiley, New York 1988.
[35] A. BRÜNGER, C. B. BROOKS AND M. KARPLUS, *Stochastic boundary conditions for molecular dynamics simulations of* ST2 *water*, Chem. Phys. Lett. 105 (1982), pp. 495–500.
[36] A. T. BRÜNGER, J. KURIYAN AND M. KARPLUS, *Crystallographic R factor refinement by molecular dynamics*, Science 235 (1987), pp. 458.
[37] U. BURKERT AND N. L. ALLINGER, *Molecular Mechanics*, Amer. Chem. Soc., Washington, D. C. 1982.
[38] L. J. BUTUROVIĆ, T. F. SMITH AND S. VAJDA, *Finite-state and reduced-parameter representations of protein backbone conformations*, J. Comp. Chem. 15(1994), pp. 300–312.
[39] R. H. BYRD, E. ESKOW, R. B. SCHNABEL AND S. L. SMITH, *Parallel global optimization: numerical methods, dynamic scheduling methods, and applications to molecular configuration*, in Parallel Computation, B. Ford and A. Fincham (eds.), Oxford University Press, 1993, pp. 187–207.
[40] R. H. BYRD, E. ESKOW, A. VAN DER HOEK, R. B. SCHNABEL AND K. B. OLDENKAMP, *A parallel global optimization method for solving molecular cluster and polymer conformation problems*, in Proc. 7th SIAM Conf. Parallel Processing Sci. Comput., SIAM, 1995, pp. 72–77.
[41] R. H. BYRD, E. ESKOW, A. VAN DER HOEK, R. B. SCHNABEL, C.-S. SHAO AND Z. ZOU, *Global optimization methods for protein folding problems*, in Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos et al.,

eds., Amer. Math. Soc., Providence, RI, 1996, pp. 29–39.

[42] A. Caflisch and M Karplus, *Molecular dynamics studies of protein and peptide folding and unfolding*, Chapter 7 in The Protein Folding Problem and Tetiary Structure Prediction, K. Merz et al., eds, Birkhäuser, Boston 1994.

[43] C. J. Camacho and D. Thirumalai, *Minimum energy compact structures of random sequences of heteropolymers*, Phys. Rev. Lett. 71 (1993), pp. 2505–2508.

[44] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry*, Freeman, New York 180.

[45] G. Casari and M. J. Sippl, *Structure-derived hydrophobic potential*, J. Mol. Biol. 224 (1992), pp. 725–732.

[46] H. F. Chan and K. A. Dill, *Origins of structure in globular proteins*, Proc. Natl. Acad. Sci. USA 87 (1990), pp. 6388–6392.

[47] H. F. Chan and K. A. Dill, *Polymer principles in protein structure and stability*, Ann. Rev. Biophys. Chem. 20 (1991), pp. 447–490.

[48] H. F. Chan and K. A. Dill, *The protein folding problem*, Physics Today (February 1993), pp. 24–32.

[49] C. Chen, Y. Zhu, J. A. King and L. B. Evans, *A molecular thermodynamic approach to predict the secondary structure of homopolypeptides in aqueous systems*, Biopolymers 32 (1992), pp. 1375–1392.

[50] G. Ciccotti, D. Frenkel and I. R. McDonald, *Simulation of Liquids and Solids: Molecular Dynamics and Monte Carlo Methods in Statistical Mechanics*, North-Holland Amsterdam 1987.

[51] G. Ciccotti, and J. P. Ryckaert, *On the derivation of the generalized Langevin equation for interacting Brownian particles*, J. Stat. Phys. 26 (1981), pp. 73–82.

[52] M. Clark, R. D. Cramer III and N. van Opdenbosch, *Validation of the general purpose Tripos 5.2 force field*, J. Comput. Chem. 10 (1989), pp. 982–1012.

[53] T. Clark, *A Handbook of Computational Chemistry: A Practical Guide to Chemical Structure and Energy Calculations*, Wiley-Interscience, New York 1985.

[54] F. E. Cohen, L. M. Gregoret, S. R. Presnell and I. D. Kuntz, *Theoretical approaches to protein structure prediction*, in Protein Folding, L. M. Gierasch and J. King, eds., Amer. Ass. Adv. Sci., Washington 1990, pp. 251–258.

[55] T. Coleman, D. Shalloway and Z. Wu, *A parallel build-up algorithm for global energy minimizations of molecular clusters using effective energy simulated annealing*, J. Global Optim. 4 (1994), pp. 171–186.

[56] M. L. Conolly, *Molecular surfaces: a review*, Network Science, April 1996. (http://www.awod.com/netsci/Issues/Apr96/feature1.html)

[57] J. H. Conway, T. D. Duff, R. H. Hardin and N. A. Sloane, *Minimal-energy clusters of hard spheres*, Discr. Comput. Geom., to appear.

[58] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollmann, *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*, J. Amer. Chem. Soc. 117 (1995), pp. 5179–5197.

[59] D. G. Covell and R. L. Jernigan, *Conformations of folded proteins in restricted spaces*, Biochemistry 29 (1990), pp. 3287–3294.

[60] C. J. Cramer and D. G. Truhlar, *Continuum solvation models: classical and quantum mechanical implementations*, in Reviews in Computational Chemistry, Vol. VI, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1995, pp. 1–72.

[61] I. P. Crawford, T. Niermann and K. Kirschner, *Predictions of secondary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthase*, Proteins: Struct. Funct. Gen. 1 (1987), pp. 118–129.

[62] T. E. Creighton, *Proteins. Structure and Molecular Principles*, Freeman, New York 1984.

[63] T. E. Creighton, *Understanding protein folding pathways and mechanisms*, in Protein Folding, L. M. Gierasch and J. King, eds., Amer. Ass. Adv. Sci., Washington 1990, pp. 157–170.

[64] G. M. Crippen, *Chemical distance geometry: current realization and future projection*, J. Math. Chem. 6 (1991), pp. 307–324.

[65] G. M. Crippen, *Distance Geometry and Conformational Calculations*, Wiley, New York 1981.

[66] G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformation*, Wiley, New York 1988.

[67] G. M. Crippen and M. E. Snow, *A 1.8 Å resolution potential function for protein folding*, Biopolymers 29 (1990), pp. 1479–1489.

[68] P. Culot, G. Dive, V. H. Nguyen and J. M. Ghuysen, *A quasi-Newton algorithm for first order saddle point location*, Theor. Chim. Acta 82 (1992), pp. 189–205.

[69] V. Daggett and M. Levitt, *Protein unfolding pathways explored through molecular dynam-*

ics simulations, J. Mol. Biol. 232 (1993), pp. 600–619.

[70] T. Darden, D. York and L. Pedersen, *Particle mesh Ewald: An* $N \cdot \log(N)$ *method for Ewald sums in large systems*, J. Chem Phys. 98 (1993), pp. 10089–10092.

[71] P. Dauber-Osguthorpe and D. J. Osguthorpe, *Partitioning the motion in molecular dynamics simulations into characteristic modes of motion*, J. Comp. Chem. 14 (1993), pp. 1259–1271.

[72] P. Dauber-Osguthorpe, V. A. Roberts, D. J. Osguthorpe, J. Wolff, M. Genest and A. T. Hagler, Proteins: Struct. Funct. Gen. 4 (1988), pp. 31.

[73] L. Davis, ed., *Handbook of Genetic Algorithms*, van Nostrand Reinhold, New York 1991.

[74] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics in Applied Math. 16, SIAM, Philadelphia 1996.

[75] P. Derreumaux, and G. Vergoten, *Influence of the Spectroscopic Potential Energy Function* SPASIBA *on Molecular Dynamics of Proteins: Comparison with the* AMBER *Potential*, J. Mol. Struct. 286 (1993), pp. 55–64.

[76] P. Derreumaux, G. Zhang, T. Schlick and B. Brooks, *A truncated Newton minimizer adapted for* CHARMM *and biomolecular applications*, J. Comp. Chem. 15 (1994), pp. 532–552.

[77] J. M. Deutch and I. Oppenheim, *Molecular theory of Brownian motion for several particles*, J. Chem. Phys. 54 (1971), pp. 3547–3555.

[78] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas and H. S. Chan, *Principles of protein folding - a perspective from simple exact models*, Protein Science 4 (1995), pp. 561–602.

[79] K. A. Dill, A. T. Phillips and J. B. Rosen, *Molecular structure prediction by global optimization*, Manuscript (1996). (phillips@nadn.navy.mil)

[80] U. Dinur and A. T. Hagler, *New approaches to empirical force fields*, in Reviews in Computational Chemistry, Vol. II, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1991, pp. 99–164.

[81] C. M. Dobson, *Unfolded proteins, compact states and molten globules*, Curr. Opinion Struct. Biol. 2 (1992), pp. 6–12.

[82] S. Dosanjh and W. J. Camp, *Computational design of materials: Part II*, SIAM News, May/June 1994, 10–11.

[83] N. Draper and H. Smith, *Applied Regression Analysis*, 2nd ed., Wiley, New York 1981.

[84] R. L. Dunback and M. Karplus, *Backbone-dependent rotamer library for proteins. Application to side-chain prediction*, J. Mol. Biol. 230 (1993), pp. 543–574.

[85] R. L. Dunback and M. Karplus, *Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains*, Struct. Biol. 1 (1994), pp. 334–340.

[86] J. D. Dunitz, H. Eser, M. Bixon and S. Lifson, *Die Strukturen der mittleren Ringverbindungen XII - XIII*, Helvetica Chimica Acta 50 (1967), pp. 1565–1572; 1572-1583.

[87] Y. K. Edwards and S. J. Perkins, *The protein fold of the von Willebrand factor type A is predicted to be similar to the open twisted beta-sheet flanked by alpha-helices found in human ras-p21*, FEBS Letters 358 (1995), pp. 283–286.

[88] J. E. Eksterowicz and K. N. Houk, *Transition state modeling with empirical force fields*, Chem. Rev. 93 (1993),2439–2461.

[89] R. Elber and M. Karplus, *Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin*, Science 235 (1987), pp. 318–321.

[90] R. A. Engh and R. Huber, *Accurate bond and angle parameters for X-ray protein structure refinement*, Acta Cryst. A47 (1991), pp. 392–400.

[91] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *A smooth particle mesh Ewald method*, J. Chem. Phys., to appear.

[92] P. Ewald, Ann. Phys. 64 (1921), pp. 253.

[93] L. Farnell, W. G. Richards and C. R. Ganellin, *Calculation of conformational free energy of histamine*, J. Theor. Biol. 43 (1974), pp. 389–392.

[94] V. V. Fedorov, *Theory of Optimal Experiments*, Acad. Press, London 1972.

[95] D. M. Ferguson, A. Marsh, T. Metzger, D. Garrett and K. Kastella, *Conformational searches for the global minimum of protein models*, J. Global Optim. 4 (1994), pp. 209–227.

[96] J. S. Fetrow and S. H. Bryant, *New programs for protein tertiary structure prediction*, Biotechnology 11 (1993), pp. 479–484.

[97] R. Fletcher, *Practical Methods of Optimization*, Wiley, New York 1987.

[98] G. Fogarasi and P. Pulay, *Ab initio calculation of force fields and vibrational spectra*, in Vibrational Spectra and Structure, Vol. 14, J. R. Durig, ed., Elsevier, Amsterdam 1985, pp. 125–219.

[99] S. Fortier, I. Castleden, J. Glasgow, D. Conklin, C. Walmsley, L. Leherte and F. A. Allen, *Molecular scene analysis: the integration of direct-method and artificial intelligence strategies for solving protein crystal structures*, Acta Cryst. D49 (1993), pp. 168–178.

[100] M. S. Friedrichs, R. A. Goldstein and P. G. Wolynes, *Generalized protein tertiary structure recognition using associative memory Hamiltonians*, J. Mol. Biol. 222 (1991), pp. 1013–1034.

[101] H. Fröhlich, *Theory of Dielectrics*, Clarendon Press, Oxford 1958.

[102] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, London 1990.

[103] C. W. Gardiner, *Handbook of Stochastic Methods*, Springer, Berlin 1985.

[104] J. Garnier, *Protein structure prediction*, Biochimie 72 (1990), pp. 513–524.

[105] D. L. Gerloff, G. Chelvanayagam and S. A. Benner, *A predicted consensus structure for the protein-kinase c2 homology (c2h) domain, the repeating unit of synaptotagmin*, Proteins: Struct. Funct. Gen. 22 (1995), pp. 299–310.

[106] C. Giacovazzo, ed., *Fundamendals of Crystallography*, Oxford Univ. Press, Oxford 1985.

[107] J. Gibrat, J. Garnier and N. Go, *Normal mode analysis of oligomeric proteins: reduction of the memory requirements by consideration of rigid geometry and molecular symmetry*, J. Comp. Chem. 15 (1994), pp. 820–837.

[108] J. Gibrat, J. Garnier and B. Robson, *Further developments of secondary structure predictions using information theory: new parameters and consideration of residue pairs*, J. Mol. Biol. 198 (1987), pp. 425–443.

[109] K. D. Gibson and H.. A. Scheraga, *Desicions in force field development. Reply to Kollman and Dill*, J. Biomol. Struct. Dyn. 8 (1991), pp. 1109–1111.

[110] L. M. Gierasch and J. King, *Protein Folding: Deciphering the Second Half of the Genetic Code*, Amer. Ass. Advancement Sci., 1990.

[111] T. L. Gilbert, *Soft-sphere model for closed-shell atoms and ions*, J. Chem. Phys. 49 (1968), pp. 2640–2642.

[112] P. E. Gill, W. Murray and M. H. Wright, *Practical Optimization*, Acad. Press, London 1981.

[113] M. K. Gilson, K. A. Sharp and B. H. Honig, *Calculating the electrostatic potential of molecules in solution: methods and error assessment*, J. Comput. Chem. 9 (1987), pp. 327–335.

[114] A. Godzik, A. Kolinski and J. Skolnick, *Topology fingerprint approach to the inverse folding problem*, J. Mol. Biol. 227 (1992), pp. 227–238.

[115] A. Godzik, A. Kolinski and J. Skolnick, *De novo and inverse folding predictions of protein structure and dynamics*, J. Computer-Aided Mol. Des. 7 (1993), pp. 397–438.

[116] A. Godzik, A. Kolinski and J. Skolnick, *Lattice representations of globular proteins: How good are they?*, J. Comp. Chem. 14 (1993), pp. 1194–1202.

[117] R. A. Goldstein, Z. A. Luthey-Schulten and P. G. Wolynes, *The statistical mechanical basis of sequence alignment algorithms for protein structure recognition*, Manuscript, 1993.

[118] N. Go and H. A. Scheraga, *Analysis of the contribution of internal vibrations to the statistical weights of equilibrium conformations of macromolecules*, J. Chem. Phys. 51 (1969), pp. 4751–4767.

[119] S. K. Gray, D. W. Noid and B. G. Sumpter, *Symplectic integrators for large-scale molecular dynamics simulations: A comparison of several explicit methods*, J. Chem. Phys. 101 (1994), pp. 4062–4072.

[120] L. Greengard and V. Rokhlin, *The rapid evaluation of potential fields in three dimensions*, in Vortex Methods, C. Anderson and C. Greengaard, eds., Lecture Notes in Math., Springer 1988, pp. 121–141.

[121] A. Greiner, W. Strittmatter and J. Honerkamp, *Numerical integration of stochastic differential equations*, J. Statist. Phys. 51 (1988), pp. 95–108.

[122] T. Gund and P. Gund, *Three-dimensional molecular modeling by computer*, Chapter 10 in Molecular Structure and Energetics, J. F. Liebman and A. Greenberg, eds., Vol. 4 (1987), pp. 319–340.

[123] J. R. Gunn, A. Monge, R. A. Friesner and C. H. Marshall, *Hierarchical algorithm for computer modeling of protein tertiary structure: folding of myoglobin to 6.2A resolution*, J. Phys. Chem. 98 (1994), pp. 702–711.

[124] Z. Guo, D. Thirumalai and J. D. Honeycutt, *Folding kinetics of proteins: A model study*, J. Chem. Phys. 97 (1992), pp. 525–535.

[125] P. Hänggi, P. Talkner and M. Borkovec, *Reaction-rate theory: fifty years after Kramers*, Rev. Mod. Phys. 62 (1990), pp. 251–341.

[126] A. T. Hagler, D. J. Osguthorpe, P. Dauber-Osguthorpe and J. C. Hempel, *Dynamics and conformational energetics of a peptide hormone: Vasopressin*, Science 227 (1985), pp. 1309–1315.
[127] E. Hairer, C. Lubich and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, Lecture Notes in Math. 1409, Springer, Berlin 1989.
[128] E. Hairer, S. P. Norsett and G. Wanner, *Solving Ordinary Differential Equations I*, 2nd rev. ed., Springer, Berlin 1993.
[129] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II*, Springer, Berlin 1991.
[130] E. Hansen, *Global Optimization Using Interval Analysis*, Dekker, New York 1992.
[131] M.-H. Hao and S. C. Harvey, *Analyzing the normal mode dynamics of macromolecules by the component synthesis method*, Biopolymers 32 (1992), pp. 1395–1405.
[132] M.-H. Hao and H. A. Scheraga, *Statistical thermodynamics of protein folding: sequence dependence*, J. Phys. Chem. 98 (1994), 9882-9893.
[133] T. Head-Gordon and F. H. Stillinger, *Predicting polypeptide and protein structures from amino acid sequence: antlion method applied to melittin*, Biopolymers 33 (1993), pp. 293–303.
[134] T. Head-Gordon, F. H. Stillinger and J. Arrecis, *A strategy for finding classes of minima on a hypersurface: implications for approaches to the protein folding problem*, Proc. Natl. Acad. Sci. USA 88 (1991), pp. 11076–11080.
[135] T. Head-Gordon, F. H. Stillinger, M. H. Wright and D. M. Gay, Poly(L-alanine) as a universal reference material for understanding protein energies and structures, Proc. Natl. Acad. Sci. USA 89 (1992), pp. 11513–11517.
[136] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari and M. J. Sippl, *Identification of native protein folds amongst a large number of incorrect models*, J. Mol. Biol. 216 (1990), pp. 167–180.
[137] W. A. Hendrickson, Methods Encymol. 115 (1985), pp. 252–270.
[138] D. M. Hirst, *A Computational Approach to Chemistry*, Blackwell Sci. Publ., Oxford 1990.
[139] M. R. Hoare, *Structure and dynamics of simple microclusters*, Adv. Chem. Phys. 40 (1979), pp. 49–135.
[140] M. R. Hoare and J. A. McInnes, *Morphology and statistical statics of simple microclusters*, Adv. Phys. 32 (1983), pp. 791–821.
[141] U. Hobohm, M. Scharf, R. Schneider and C. Sander, *Selection of representative protein data sets*, Protein Sci. 1 (1992), pp. 409–417.
[142] C. Hoheisel, *Theoretical Treatment of Liquids and Liquid Mixtures*, Elsevier, Amsterdam 1993.
[143] J. Holland, *Genetic algorithms and the optimal allocation of trials*, SIAM J. Computing 2 (1973), pp. 88–105.
[144] L. Holley and M. Karplus, *Protein secondary structure prediction with a neural network*, Proc. Natl. Acad. Sci. USA 86 (1989), pp. 152–156.
[145] J. Honerkamp, *Stochastic Dynamical Systems: Concepts, Numerical Methods, Data Analysis*, VCH, New York 1994.
[146] J. D. Honeycutt and D. Thirumalai, *The nature of folded states of globular proteins*, Biopolymers 32 (1992), pp. 695–709.
[147] B. Honig and A. Nicholls, *Classical electrostatics in biology and chemistry*, Science 268 (1995), pp. 1144–1149.
[148] A. J. Hopfinger and R. A. Pearlstein, *Molecular mechanics force-field parametrization procedures*, J. Comput. Chem. 5 (1984), pp. 486–499.
[149] Q. X. Hua, M. Kochoyan and M. A. Weiss, *Structure and dynamics of despentapeptide-insulin in solution: the molten globule hypothesis*, Proc. Natl. Acad. Sci. USA 89 (1992), pp. 2379–2383.
[150] Q. X. Hua, J. E. Ladbury and M. A. Weiss, *Dynamics of a monomeric insulin analogue: testing the molten globule hypothesis*, Biochemistry 32 (1993), pp. 1433–1442.
[151] A. Iniesta and J. G. de la Torre, *A second-order algorithm for the simulation of the Brownian dynamics of macromolecular models*, J. Chem. Phys. 92 (1990), pp. 2015–2018.
[152] G. Iori, E. Marinari and G. Parisi, *Heteropolymer folding on a* APE-100 *supercomputer*, Int. J. Mod. Phys. C 4 (1993), pp. 1333–1341.
[153] A. Irbäck, C. Peterson and F. Potthast, *Identification of amino acid sequences with good folding properties*, Phys. Rev. E, submitted, 1996.
[154] A. Irbäck and F. Potthast, *Studies of an off-lattice model for protein folding: sequence dependence and improved sampling at finite temperature*, J. Chem. Phys. 103 (1995), pp. 10298

[155] IUPAC-IUP Commission on Biochemical Nomenclature, *Abbreviations and symbols for the description of the conformation of polypeptide chains*, Biochemistry 9 (1970), pp. 3471–3479.

[156] R. Jaenicke, *Protein folding: local structures, domains, subunits, and assemblies*, Biochemistry 30 (1991), pp. 3147–3161.

[157] O. Jardetzky and A. N. Lane, *Determination of the solution structure of proteins from* NMR, in Physics of NMR Spectroscopy in Biology and Medicine, B. Maraviglia, ed., North Holland, Amsterdam 1988, pp. 267–301.

[158] M. S. Johnson, N. Srinivasan, R. Sowdhamini and T. L. Blundell, *Knowledge-based protein modeling*, Crit. Rev. Biochem. Mol. Biol. 29 (1994), pp. 1–68.

[159] D. Jones and J. Thornton, *Protein fold recognition*, J. Comput. Aided Mol. Design 7 (1993) 439–456.

[160] W. Kabsch, *A discussion of the solution for the best rotation to relate two sets of vectors*, Acta Cryst. A34 (1978), pp. 827–828.

[161] H. S. Kang, N. A. Kurochkina and B. Lee, *Estimation and use of protein backbone angle probablilities*, J. Mol. Biol. 229 (1993), pp. 448–460.

[162] I. G. Kaplan, *Theory of Molecular Interactions*, Elsevier, Amsterdam 1986.

[163] M. Karplus, A. Caflish, A. Šali and E. Shaknovich, *Protein dynamics: from the native to the unfolded state and back again*, in Proc. Int. Conf. Mol. Struct. Biol. Vienna, September 17–20, 1995, A. J. Kungl. et al., eds., Gesellschaft Österreichischer Chemiker, Wien 1995, pp. 129–155.

[164] H.Kawai, T.Kikuchi and Y. Okamoto, *A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method*, Protein Eng. 3 (1989), pp. 85–94.

[165] T. Kihara and S. Ichimaru, *Intermolecular Forces*, Wiley, Chichester 1978.

[166] S. Kirkpatrick, C. D. Geddat, Jr., and M. P. Vecchi, *Optimization by simulated annealing*, Science 220 (1983), pp. 671–680.

[167] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin 1992.

[168] P. E. Kloeden, E. Platen and H Schurz, *Numerical Solution of* SDE *Through Computer Experiments*, Springer, Berlin 1994.

[169] D. Kneller, F. Cohen and R. Langridge, *Improvements in protein secondary structure predictions by an enhanced neural network*, J. Mol. Biol. 214 (1990), pp. 171–182.

[170] H. Köppen, *The use of supercomputers in medical chemistry. Examples from peptide and protein projects*, in Supercomputer and Chemistry, U. Harms, ed., Springer, Berlin 1990, pp. 99–113.

[171] A. Kolinski, A. Godzik and J. Skolnick, *A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: application to designed helical proteins*, J. Chem. Phys. 98 (1993), pp. 7420–7433.

[172] J. Kostrowicki, L. Piela, B. J. Cherayil and H. Scheraga, *Performance of the diffusion equation method in searches for optimum structures of clusters of Lennard-Jones atoms*, J. Phys. Chem. 95 (1991), pp. 4113–4119.

[173] J. Kostrowicki and H. A. Scheraga, *Application of the diffusion equation method for global optimization to oligopeptides*, J. Phys. Chem. 96 (1992), pp. 7442–7449.

[174] J. Kostrowicki and H. A. Scheraga, *Some approaches to the multiple-minima problem in protein folding*, in Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos et al., eds., Amer. Math. Soc., Providence, RI, 1996, pp. 123–132.

[175] S. Krimm, *Spectra and structure of polypeptides*, in Vibrational Spectra and Structure, Vol. 16, J. R. Durig, ed., Elsevier, Amsterdam 1987, pp. 1–72.

[176] B. M. Ladanyi and M. S. Skaf, *Computer simulation of hydrogen-bonding liquids*, Ann. Rev. Phys. Chem. 44 (1993), pp. 335–368.

[177] M. Lambert and H. Scheraga, *Pattern recognition in the prediction of protein structure, I-III*, J. Comp. Chem. 10 (1989), pp. 770–797; 798-816; 817-831.

[178] E. S. Lander and M. S. Waterman, *Calculating the Secrets of Life: Contributions of the Mathematical Sciences to Molecular Biology*, National Academic Press, Washington, D. C. , 1995.

[179] R. H. Lathrop and T. F. Smith, *Global optimum protein threading with gapped alignment and empirical pair score functions*, J. Mol. Biol., 1996 (in press).

[180] A. R. Leach, *A survey of methods for searching the conformational space of small and medium-sized molecules*, in Reviews in Computational Chemistry, Vol. II, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1991, pp. 1–55.

[181] T. Lee, D. M. York and W. Yang, *A new definition of atomic charges based on a variational*

principle for the electrostatic potential energy, J. Chem. Phys. 102 (1995), pp. 7549–7556.
[182] C. Lee and S. Subbiah, *Prediction of protein side-chain conformation by packing optimization*, J. Mol. Biol. 217 (1991), pp. 373–388.
[183] S. L. Le Grand and K. M. Merz jr., *The application of the genetic algorithm to the minimization of potential energy functions*, J. Global Optim. 3 (1993), pp. 49–66.
[184] C. M. -R. Lemer, M. J. Rooman and S. J. Wodak, *Protein structure prediction by threading methods: evaluation of current techniques*, Proteins: Struct. Funct. Gen. 23 (1995), pp. 337–355.
[185] P. E. Leopold, M. Montal and J. N. Onuchic, *Protein folding funnels: a kinetic approach to the sequence-structure relationship*, Proc. Natl. Acad. Sci. USA 89 (1992), pp. 8721–8725.
[186] C. Levinthal, *Are there pathways to protein folding?* J. Chim. Phys. 65 (1968), pp. 44–45.
[187] M. Levitt, C. Sander and P. S. Stern, *Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme*, J. Mol. Biol. 181 (1985), pp. 423–447.
[188] R. M. Levy, R. P. Sheridan, J. W. Keepers, G. S. Dubey, S. Swaminathan and M. Karplus, *Molecular dynamics of myoglobin at* $298^o K$, Biophys. J. 48 (1985), pp. 509–518.
[189] Z. Li and H. A. Scheraga, *Monte Carlo approach to the multiple-minima problem in protein folding*, Proc. Natl. Acad. Sci. USA 84 (1987), pp. 15–29.
[190] S. Lifson, *Potential energy functions for structural molecular biology*, in Supramolecular Structure and Function, G. Pifat and J. N. Herak, eds., Plenum, New York 1983, pp. 1–44.
[191] K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. I-VII, VCH Publ., New York 1990–1996.
[192] J. Ma and J. E. Straub, *Simulated annealing using the classical density distribution*, J. Chem. Phys. 101 (1994), pp. 533–541.
[193] R. März, *Numerical methods for differential-algebraic equations*, in Acta Numerica 1992, A. Iserles, ed., Cambridge Univ. Press 1992, pp. 141–198.
[194] V. N. Maiorov and G. M. Crippen, *Contact potential that recognizes the correct folding of globular proteins*, J. Mol. Biol. 227 (1992), pp. 876–888.
[195] G. C. Maitland, M. Rigby, E. B. Smith and W. A. Wakeham, *Intermolecular Forces. Their Origin and Determination*, Clarendon Press, Oxford 1981.
[196] J. R. Maple, M.-J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewing and A. T. Hagler, *Derivation of Class II Force Fields. I. Methodology and Quantum Force Field for the Alkyl Functional Group and Alkane Molecules*, J. Comput. Chem. 15 (1994), pp. 162–182.
[197] C. D. Maranas and C. A. Floudas, *A global optimization approach for Lennard-Jones microclusters*, J. Chem. Phys. 97 (1992), pp. 7667–7678.
[198] C. D. Maranas and C. A. Floudas, *Global minimum potential energy conformations of small molecules*, J. Global Optim. 4 (1994), pp. 135–170.
[199] C. D. Maranas and C. A. Floudas, *A deterministic global optimization approach for molecular structure determination*, J. Chem. Phys. 100 (1994), pp. 1247–1261.
[200] C. D. Maranas, I. P. Androulakis and C. A. Floudas, *A deterministic global optimization approach for the protein folding problem*, in Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos et al., eds., Amer. Math. Soc., Providence, RI, 1996, pp. 133–150.
[201] P. Matstoms, *Sparse* QR *Factorization in* MATLAB, ACM Trans. Math. Software 20 (1994), pp. 136–159.
[202] J. A. McCammon and S. C. Harvey, *Dynamics of Proteins and Nuclein Acids*, Cambridge Univ. Press, Cambridge 1987.
[203] J. A. McCammon and M. Karplus, *Simulation of protein dynamics*, Ann. Rev. Phys. Chem. 31 (1980), pp. 29–45.
[204] M. L. McKee and M. Page, *Computing reaction pathways on molecular potential energy surfaces*, in Reviews in Computational Physics, Vol. IV, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1993, pp. 337–379.
[205] P. G. Mezey, *Potential Energy Hypersurfaces*, Elsevier, Amsterdam 1987.
[206] S. Miyazawa and R. L. Jernigan, Protein Eng. 6 (1993), pp. 267–278.
[207] A. D. Miranker and C. M. Dobson, *Collapse and cooperativity in protein folding*, Curr. Opinion Struct. Biol. 6 (1996), pp. 31–42.
[208] S. Miyamoto and P. A. Kollman, SETTLE: *An analytic version of the* SHAKE *and* RATTLE *algorithm for rigid water models*, J. Comp. Chem. 13 (1992), pp. 952–962.
[209] J. Mockus, *Bayesian Approach to Global Optimization*, Kluwer, Dordrecht 1989.
[210] F. A. Momany, R. F. McGuire, A. W. Burgess and H. A. Scheraga, *Energy parameters in polypeptides. VIII*, J. Phys. Chem. 79 (1975), pp. 2361–2381.
[211] F. A. Momany, V. J. Klimkowski and L. Schäfer, *On the use of conformationally depen-*

dent geometry trends from ab initio dipeptide studies to refine potentials for the empirical force field CHARMM, J. Comp. Chem. 11 (1990), pp. 654–662.

[212] F. Momany, R. McGuire, A. Burgess and H. Scheraga, *Energy parameters in polypeptides VII*, J. Phys. Chem. 79 (1975), pp. 2361–2381.

[213] A. Monge, R. A. Friesner and B. Honig, *An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure*, Proc. Natl. Acad. Sci. USA 91 (1994), pp. 5027–5029.

[214] J. J. Moré and Z. Wu, *Global continuation for distance geometry problems*, Manuscript (1995). (more@mcs.anl.gov)

[215] J. J. Moré and Z. Wu, *Smoothing techniques for macromolecular global optimization*, Manuscript (1995). (more@mcs.anl.gov)

[216] B. T. Nall and K. A. Dill, *Conformations and Forces in Protein Folding*, Amer. Ass. Adv. Sci., Washington 1991.

[217] S. G. Nash, *Preconditioning of truncated Newton methods*, SIAM J. Sci. Statist. Comp. 6 (1985), pp. 599–616.

[218] National Research Council, *Mathematical Challenges From Theoretical/Computational Chemistry*, National Academy Press, Washington, D. C. 1995. (available via the World Wide Web, http://www.nas.edu)

[219] G. Némethy, K. D. Gibson, K. A. Palmer, C. N. YoonG. Paterlini, A. Zagari, S. Rumsey and H. A. Scheraga, *Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the* ECEPP/3 *algorithm, with application to proline-containing peptides*, J. Phys. Chem. 96 (1992), pp. 6472–6484.

[220] G. L. Nemhauser and L. A. Wolsey, *Integer Programming*, Chapter VI in Optimization, G. L. Nemhauser et al., eds., Handbooks in Operations Research and Management Science, Vol. 1, North Holland, Amsterdam 1989, pp. 447–527.

[221] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge Univ. Press, Cambridge 1990.

[222] A. Neumaier, *Second-order sufficient optimality conditions for local and global nonlinear programming*, Manuscript, 1994.

[223] A. Nicholls and B. Honig, *A rapid finite difference algorithm, utilizing successive overrelaxation to solve the Poisson-Boltzmann equation*, J. Comput. Chem. 12 (1991), pp. 435–445.

[224] J. A. Northby, *Structure and binding of Lennard-Jones clusters*: $13 \leq N \leq 147$, J. Chem. Phys. 87 (1987),6166–6177.

[225] J. Novotný, R. Bruccoleri and M. Karplus, *An analysis of incorrectly folded protein models*, J. Mol. Biol. 177 (1984), pp. 787–818.

[226] W. K. Olson, *How flexible is the furanose ring? An updated potential energy estimate*, J. Am. Chem. Soc. 104 (1982), pp. 278–286.

[227] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten and N. D. Socci, *Toward an outline of the topography of a realistic protein-folding funnel*, Proc. Natl. Acad. Sci. USA 92 (1995), pp. 3626–3630.

[228] M. Oobatake and G. M. Crippen, *Residue-residue potential function for conformational analysis of proteins*, J. Phys. Chem. 85 (1981), pp. 1187–1197.

[229] P. M. Pardalos, D. Shalloway and G. Xue, *Optimization methods for computing global minima of nonconvex potential energy functions*, J. Global Optim. 4 (1994), pp. 117–133.

[230] P. M. Pardalos, D. Shalloway and G. Xue, eds., *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, Amer. Math. Soc., Providence, RI, 1996.

[231] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, N. J. , 1980.

[232] R. W. Pastor, *Techniques and applications of Langevin dynamics simulations*, in The Molecular Dynamics of Liquid Crystals, G. R. Luckhurst and C. A. Veracini, eds., Kluwer, Dordrecht 1994, pp. 85–138.

[233] R. W. Pastor, B. R. Brooks and A. Szabo, *An analysis of the accuracy of Langevin and molecular dynamics algorithm*, Mol. Phys. 65 (1988), pp. 1409–1419.

[234] D. A. Pearlman and P. A. Kollman, *Evaluating the assumptions underlying force field development and application using free energy conformational maps for nucleotides*, J. Am. Chem. Soc. 113 (1991), pp. 7167–7177.

[235] G. Perrot, B. Cheng, K. D. Gibson, J. Vila, K. A. Palmer, A. Nayeem, B. Maigret and H. A. Scheraga, MSEED: *a program for the rapid analytic determination of accessible surface areas and their derivatives*, J. Comput. Chem. 13 (1992), pp. 1–11.

[236] W. B. Person and K. Szcepaniak, *Calculated and experimental vibrational spectra and*

force fields for isolated pyrimidine bases, in Vibrational Spectra and Structure, Vol. 20, J. R. Durig, ed., Elsevier, Amsterdam 1993, pp. 240–325.

[237] C. S. Peskin and T. Schlick, *Molecular dynamics by the backward-Euler method*, Comm. Pure Appl. Math. 42 (1989), pp. 1001–1031.

[238] A. T. Phillips and J. B. Rosen, *A quadratic assignment formulation of the molecular conformation problem*, J. Global Optim. 4 (1994), pp. 229–241.

[239] L. Piela, J. Kostrowicki and H. A. Scheraga, *The multiple-minima problem in the conformational analysis of molecules. Deformation of the protein energy hypersurface by the diffusion equation method*, J. Phys. Chem. 93 (1989), pp. 3339–3346.

[240] S. B. Prusiner, *The prion disease*, Scientific American, January 1995, 30–37.

[241] F. Pukelsheim, *On linear regression designs which maximize information*, J. Statist. Plann. Inference, 4 (1980), pp. 339–364.

[242] N. Qian and T. Sejnowski, *Predicting the secondary structure of globular proteins using neural network models*, J. Mol. Biol. 202 (1988), pp. 865–884.

[243] H. Rabitz, *Systems sensitivity analysis at the molecular scale*, Science 246 (1989), pp. 221–226.

[244] A. A. Rabow and H. A. Scheraga, *Lattice neural network minimization*, J. Mol. Biol. 232 (1993), pp. 1157–1168.

[245] S. Rackovsky, *On the nature of the protein folding code*, Proc. Natl. Acad. Sci. USA 90 (1993), pp. 644–648.

[246] S. E. Radford and C. M. Dobson, *Insight into protein folding using physical techniques: studies of lysozome and α-lactalbumin*, Phil. Trans. R. Soc. Lond. B 348 (1995), pp. 17–25.

[247] C. Radin and L. S. Schulmann, *Periodicity of classical ground states*, Phys. Rev. Lett. 51 (1983), pp. 621–622.

[248] G. Ramachandran and T. Schlick, *Beyond optimization: simulating the dynamics of supercoiled* DNA *by a macroscopic mode*, in Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos et al., eds., Amer. Math. Soc., Providence, RI, 1996, pp. 215—231.

[249] L. Regan, S. P. Ho, Z. Wasserman and W. F. De Grado, *Helical proteins. De novo designs*, in Protein Folding, L. M. Gierasch and J. King, eds., Amer. Ass. Adv. Sci., Washington 1990, pp. 171–176.

[250] W. G. Richards, *Calculation of conformational free energy of histamine*, J. Theor. Biol. 43 (1974), pp. 389.

[251] T. J. Richmond, *Solvent accessible surface area and excluded volume in proteins*, J. Mol. Biol. 178 (1984), pp. 63–89.

[252] F. Richards, *The protein folding problem*, Scientific American 264 (January 1991), pp. 54–63.

[253] B. Robson and J. Garnier, *Introduction to Proteins and Protein Engineering*, Elsevier, New York 1986.

[254] G. D. Rose and T. P. Creamer, *Protein folding: predicting predicting*, Proteins: Struct. Funct. Gen. 19 (1994), pp. 1–3.

[255] B. Rost and C. Sander, *Prediction of protein secondary structure at better than* 70% *accuracy*, J. Mol. Biol. 232 (1993), pp. 584–599.

[256] I. K. Roterman, M. H. Lambert, K. D. Gibson and H. A. Scheraga, *A comparison of the* CHARMM, AMBER *and* ECEPP *potentials for peptides. II. $\varphi - \psi$ maps for N-acetyl alanine N'-methyl amide: Comparisons, contrasts and simple experimental tests*, J. Biomol. Struct. Dyn. 7 (1989), pp. 421–453.

[257] R. B. Russell and M. E. Sternberg, *Protein structure prediction: how good are we?*, Current Biology 5 (1995), pp. 488–490.

[258] J. P. Ryckaert, G. Ciccotti and H. C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alcanes*, J. Comp. Phys. 23 (1977), pp. 327–341.

[259] D. S. Rykunov, B. A. Reva and A. V. Finkelstein, *Accurate general method for lattice approximation of three-dimensional structure of a chain molecule*, Proteins: Struct. Func. Gen. 22 (1995), pp. 100–109.

[260] A. Šali, E. Shakhnovich and M. Karplus, *How does a protein fold?*, Nature 369 (1994), pp. 248–251.

[261] A. Šali, E. Shakhnovich and M. Karplus, *Kinetics of protein folding. A lattice model study of the requirements for folding to the native state*, J. Mol. Biol. 235 (1994), pp. 1614–1636.

[262] A. Šali, E. Shakhnovich and M. Karplus, *Thermodynamics and kinetics of protein folding*, in Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos et al., eds., Amer. Math. Soc., Providence, RI, 1996, pp. 199–213.

[263] J. M. Sanz-Serna, *Symplectic integrators for Hamiltonian problems: an overview*, Acta Numerica 1 (1992), pp. 243–286.
[264] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London 1994.
[265] M. Saunders, K. N. Houk, Y.-D. Wu, W. C. Still, M. Lipton, G. Chang and W. C. Guida, *Conformations of cycloheptadecane. A comparison of methods for conformational searching*, J. Am. Chem. Soc. 112 (1990), pp. 1419–1427.
[266] J. N. Scarsdale, P. Ram, J. H. Prestegard and R. K. Yu, *A molecular mechanics*-NMR *pseudoenergy approach to the solution conformation of glycolipids*, J. Comput. Chem. 9 (1988), pp. 133–147.
[267] S. Scheiner, *Calculating the properties of hydrogen bonds by ab initio methods*, in Reviews in Computational Chemistry, Vol. II, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1991, pp. 165–218.
[268] H. A. Scheraga, *Predicting three-dimensional structures of oligopeptides*, in Reviews in Computational Chemistry, Vol. III, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1992, pp. 73–142.
[269] H. A. Scheraga, *Treatment of hydration in conformational energy calculations on polypeptides and proteins*, Manuscript, 1994.
[270] C. A. Schiffer, J. W. Caldwell, P. A. Kollman and R. M. Stroud, *Protein structure prediction with a combined solvation free energy-molecular mechanics force field*, Molecular Simulation 10 (1993), pp. 121–149.
[271] T. Schlick, *Modeling and minimization techniques for predicting three-dimensional structures of large biological molecules*, Ph. D. Thesis, Courant Institute of Math. Sciences, New York 1987.
[272] T. Schlick, *New approaches to potential energy minimization and molecular dynamics algorithms*, Computers Chem. 15 (1991), pp. 251–260.
[273] T. Schlick, *Optimization methods in computational chemistry*, in Reviews in Computational Chemistry, Vol. III, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1992, pp. 1–71.
[274] T. Schlick and A. Fogelson, TNPACK – *A truncated Newton minimization package for large scale problems*, ACM Trans. Math. Softw. 18 (1992), pp. 46–70, 71-111.
[275] T. Schlick and W. K. Olson, *Supercoiled* DNA *Energetics and Dynamics by Computer Simulation*, J. Mol. Biol. 223 (1992), pp. 1089–1119.
[276] T. Schlick and M. Overton, *A powerful truncated Newton method for potential energy minimization*, J. Comp. Chem. 8 (1987), pp. 1025–1039.
[277] T. Schlick, C. Peskin, S. Broyde and M. Overton, *An analysis of the structural and energetic properties of deoxyribose by potential energy methods*, J. Comp. Chem. 8 (1987), pp. 1199–1224.
[278] H. Schreiber and O. Steinhauser, *Taming cut-off induced artifacts in molecular dynamics studies of solvated polypeptides*, J. Mol. Biol. 228 (1992), pp. 909–923.
[279] H. Schreiber and O. Steinhauser, *Molecular dynamics studies of solvated polypeptides: why the cut-off scheme does not work*, Chem. Phys. 168 (1992), pp. 75–89.
[280] H. Schreiber and O. Steinhauser, *Cutoff size does strongly influence molecular dynamics results on solvated polypeptides*, Biochemistry 31 (1992), pp. 5856–5860.
[281] H. Schreiber, O. Steinhauser and P. Schuster, *Parallel molecular dynamics of biomolecules*, Parallel Computing 18 (1992), pp. 557–573.
[282] G. E. Schulz, *A critical evaluation of methods for prediction of protein secondary structures*, Ann. Rev. Biophys. Biophys. Chem. 17 (1988), pp. 1–21.
[283] D. Shalloway, *Packet annealing: a deterministic method for global minimization. Application to molecular conformation*, in Recent Advances in Global Optimization, C. A. Floudas and P. M. Pardalos, eds., Princeton Univ. Press, Princeton 1992, pp. 433–477.
[284] J. Shimada, H. Kaneko and T. Takada, *Performance of fast multipole methods for calculating electrostatic interactions in biomacromolecular simulations*, J. Comp. Chem. 15 (1994), pp. 28–43.
[285] J. K. Shin and M. S. Jhon, *High directional Monte Carlo procedure coupled with the temperature heating and annealing method to obtain the global energy minimum structure of polypeptides and proteins*, Biopolymers 31 (1991), pp. 177–185.
[286] S. D. Silvey, *Optimal Design*, Chapman and Hall, London 1980.
[287] M. J. Sippl, *Boltzmann's principle, knowledge based mean fields and protein folding*, J. Comp. Aided Mol. Design 7 (1993), pp. 473–501.
[288] M. J. Sippl, *Knowledge-based potentials for proteins*, Curr. Opinion Str. Biol. 5 (1995), pp. 229–235.

[289] M. J. SIPPL, M. HENDLICH AND P. LACKNER, *Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments*, Protein Sci. 1 (1992) 625–640.

[290] M. J. SIPPL, S. WEITCKUS AND H. FLÖCKNER, *In search of protein folds*, Chapter 12 in The Protein Folding Problem and Tertiary Structure Prediction, K. Merz and S. Legrand, eds., Birkhäuser, Boston 1994.

[291] M. J. SIPPL, S. WEITCKUS AND H. FLÖCKNER, *Fold recognition*, in Modelling of Biomolecular Structures and Mechanisms, A. Pullman et al., eds., Kluwer, Dordrecht 1995, pp. 107–118.

[292] R. D. SKEEL, J. J. BIESIADECKI AND D. OKUNBOR, *Symplectic integration for macromolecular dynamics*, Manuscript (1994). (skeel@cs.uiuc.edu)

[293] J. SKOLNICK AND A. KOLINSKI, *Simulations of the folding of a globular protein*, Science 250 (1990), pp. 1121–1125.

[294] J. SKOLNICK AND A. KOLINSKI, *Monte Carlo lattice dynamics and the prediction of protein folds*, Manuscript (1996). (skolnick@scripps.edu)

[295] J. SKOLNICK, A. KOLINSKI, C. L. BROOKS, A. GODZIK AND A. REY, *A method for predicting protein structure from sequence*, Current Biology 3 (1993), pp. 414–422.

[296] Z. SLANINA, *Does the global energy minimum always also mean the thermodynamically most stable structure?*, J. Mol. Str. 206 (1990),143–151.

[297] J. C. SLATER AND J. G. KIRKWOOD, *The van der Waals forces in gases*, Phys. Rev. 37 (1931), pp. 682–697.

[298] P. E. SMITH, B. M. PETTITT AND M. KARPLUS, *Stochastic dynamics simulations of the alanine dipeptide using a solvent-modified potential energy surface*, J. Phys. Chem. 97 (1993), pp. 6907–6913.

[299] P. E. SMITH AND B. M. PETTITT, *Peptides in ionic solutions: A comparison of the Ewald and switching function techniques*, J. Chem. Phys. 95 (1991), pp. 8430–8441.

[300] K. SOBCZYK, *Stochastic Differential Equations*, Kluwer, Dordrecht 1991.

[301] K. V. SOMAN, A. KARIMI AND D. A. CASE, *Unfolding of an α-helix in water*, Biopolymers 31 (1991), pp. 1351–1361.

[302] D. R. STAMPF, C. E. FELSER AND J. L. SUSSMAN, PDBBrowse – *a graphics interface to the Brookhaven protein data bank*, Nature 374 (1995), pp. 572–574.

[303] S. STEELE, E. R. LIPPINCOTT AND J. T. VANDERSLICE, *Comparative study of empirical internuclear potential functions*, Rev. Mod. Phys. 34 (1962), pp. 239–251.

[304] P. J. STEINBACH AND B. R. BROOKS, *New spherical-cutoff methods for long-range forces in macromolecular simulation*, J. Comp. Chem. 15 (1994), pp. 667–683.

[305] F. H. STILLINGER, *Theory and molecular models for water*, Adv. Chem. Phys. 31 (1975), pp. 1–101.

[306] F. H. STILLINGER, *Role of potential-energy scaling in the low-temperature relaxation behavior of amorphous materials*, Phys. Rev. B 32 (1985), pp. 3134–3141.

[307] F. H. STILLINGER, T. HEAD-GORDON AND C. L. HIRSHFELD, *Toy model for protein folding*, Phys. Rev. E 48 (1993), pp. 1469–1477.

[308] F. H. STILLINGER AND T. A. WEBER, *Nonlinear optimization simplified by hypersurface deformation*, J. Stat. Phys. 52 (1988), pp. 1429–1445.

[309] P. STOLORZ, A. LAPEDES AND Y. XIA, *Predicting protein secondary structure using neural nets and statistical methods*, J. Mol. Biol. 225 (1992), pp. 363–377.

[310] J. E. STRAUB, *Optimization techniques with applications to proteins*, in New Developments in Theoretical Studies of Proteins (R. Elber, ed.), World Scientific, in press.

[311] DE WITT L. SUMNERS, ed., *New Scientific Applications of Geometry and Topology*, Proc. Symp. Appl. Math. 45, Amer. Math. Soc., Providence, RI, 1992.

[312] S. SUN, *Reduced representation model of protein structure prediction: statistical potential and genetic algorithms*, Protein Sci. 2 (1993), pp. 762–785.

[313] S. SUN, *Reduced representation approach to protein tertiary structure prediction: statistical potential and simulated annealing*, J. Theor. Biol. 172 (1995), pp. 13–32.

[314] R. SUSNOV, R. B. NACHBAR, C. SCHUTT AND H. RABITZ, *Sensitivity of molecular structure to intramolecular potentials*, J. Phys. Chem. 95 (1991), pp. 8585–8597.

[315] O. TAPIA, *Solvent effects on biomolecules and reactive systems. An overview of the theory and applications*, in Computational Chemistry, S. Fraga, ed., Elsevier, New York 1992, pp. 694–721.

[316] W. R. TAYLOR ed., *Patterns in Protein Sequence and Structure*, Springer, New York 1992.

[317] O. TELEMEN AND B. JÖNSSON, *Vectorizing a general purpose molecular dynamics simulation program*, J. Comp. Chem. 7 (1986), pp. 58–66.

[318] T. S. THACHER, A. T. HAGLER AND H. RABITZ, *Application of sensitivity analysis to the establishment of intermolecular potential functions*, J. Amer. Chem. Soc. 113 (1991),

pp. 2020–2033.

[319] D. J. Tobias, J. E. Mertz and C. L. Brooks, *Nanosecond time scale folding dynamics of a pentapeptide in water*, Biochemistry 30 (1991), pp. 6054–6058.

[320] A. E. Torda and W. F. van Gunsteren, *Molecular modeling using nuclear magnetic resonance data*, in Reviews in Computational Chemistry, Vol. III, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1992, pp. 143–172.

[321] J. M. Troyer and F. E. Cohen, *Simplified models for understanding and predicting protein structure*, in Reviews in Computational Chemistry, Vol. II, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1991, pp. 57–80.

[322] M. E. Tuckerman and B. J. Berne, *Molecular dynamics in systems with multiple time scales*, J. Comp. Chem. 95 (1992), pp. 8362–8364.

[323] M. E. Tuckerman, B. J. Berne and A. Rossi, *Molecular dynamics algorithms for multiple time scales: systems with disparate masses*, J. Comp. Chem. 94 (1991), pp. 1465–1469.

[324] P. Tufféry, C. Etchebest, S. Hazout and R. Lavery, *A critical comparison of search algorithms applied to the optimization of protein side-chain conformations*, J. Comp. Chem. 14 (1993), pp. 790–798.

[325] P. Ulrich, W. Scott, W. F. van Gunsteren and A. Torda, *Protein structure prediction force fields: parametrization with quasi Newtonian dynamics*, submitted to Proteins.

[326] H. C. Urey and C. A. Bradley, *The vibrations of pentatonic tetrahedral molecules*, Phys. Rev. 38 (1931),1969–1978.

[327] S. Vajda and C. Delisi, *Determining minimum energy conformations of polypeptides by dynamic programming*, Biopolymers 29 (1990), pp. 1755–1772.

[328] B. van der Graaf and J. A. Baas, *The implementation of constraints in molecular mechanics to explore potential energy surfaces*, Rev. Trav. Chim. Pays-Bas 99 (1980), pp. 327.

[329] A. van der Hoek, *Parallel global optimization of proteins*, Masters thesis, Erasmus Universiteit Rotterdam, 1994.

[330] W. F. van Gunsteren, *Constrained dynamics of flexible molecules*, Mol. Phys. 40 (1980), pp. 1015–1019.

[331] W. F. van Gunsteren, *Computer simulation by molecular dynamics as a tool for modelling of molecular systems*, Mol. Simul. 3 (1989), pp. 187–200.

[332] W. F. van Gunsteren and H. C. Berendsen, *Algorithms for macromolecular dynamics and constrained dynamics*, Mol. Phys. 34 (1977), pp. 1311–1327.

[333] W. F. van Gunsteren and H. C. Berendsen, *Algorithms for Brownian dynamics*, Mol. Phys. 45 (1982), pp. 637–647.

[334] W. F. van Gunsteren, H. C. Berendsen and J. C. Rullmann, *Stochastic dynamics for molecules with constraints. Brownian dynamics of n-alcanes*, Mol. Phys. 44 (1981), pp. 69–95.

[335] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, North-Holland, Amsterdam 1981.

[336] D. L. Veenstra, D. M. Ferguson and P. A. Kollman, *How transferable are hydrogen parameters in molecular mechanics calculations?*, J. Comp. Chem. 13 (1992), pp. 971–978.

[337] G. A. Voth and E. V. O'Gorman, *An effective barrier model for describing quantum mechanical activated rate processes in condensed phases*, J. Chem. Phys. 94 (1991), pp. 7342–7352.

[338] L. L. Walsh, *Navigating the Brookhaven protein data bank*, Cabos Communication 10 (1994), pp. 551–557.

[339] A. Warshel, *Computer Modeling of Chemical Reactions in Encymes and Solutions*, Wiley-Interscience, New York 1991.

[340] M. Watanabe and M. Karplus, *Dynamics of molecules with internal degrees of freedom by multiple time-step methods*, J. Chem. Phys. 99 (1993), pp. 8063–8074.

[341] S. J. Weiner, P. A. Kollmann, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta and P. Weiner, *A new force field for molecular mechanical simulation of nuclear acids and proteins*, J. Amer. Chem. Soc. 106 (1984),765–784.

[342] S. J. Weiner, P. A. Kollmann, D. T. Nguyen and D. A. Case, *An all atom force field for simulations of proteins and nuclear acids*, J. Comp. Chem. 7 (1986), pp. 230–252.

[343] L. Wesson and D. Eisenberg, *Atomic solvation parameters applied to molecular dynamics of proteins in solution*, Protein Sci. 1 (1992), pp. 227–235.

[344] L. T. Wille and J. Vennik, *Computational complexity of the ground-state determination of atomic clusters*, J. Phys. A 18 (1985), pp. L419–422.

[345] D. E. Williams, *Net atomic charge and multipole models for the ab initio molecular electric potential*, in Reviews in Computational Chemistry, Vol. II, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publ., New York 1991, pp. 219–271.

[346] S. J. Wodak and M. J. Rooman, *Generating and testing protein folds*, Curr. Opinion Struct. Biol. 3 (1993), pp. 247–259.

[347] P. G. Wolynes, *Spin glass ideas and the protein folding problem*, in Spin Glasses and Biology, D. L. Stein, ed., World Scientific, New York 1992, pp. 225–259 .

[348] P. G. Wolynes, J. N. Onuchic and D. Thirumalai, *Navigating the folding routes*, Science 267 (1995), pp. 1619–1620.

[349] Z. Wu, *The effective energy transformation scheme as a general continuation approach to global optimization with application to molecular conformation*, SIAM J. Optimization, to appear.

[350] G. Xue, *Molecular conformation on the* CM-5 *by parallel two-level simulated annealing*, J. Global Optim. 4 (1994), pp. 187–208.

[351] G. Xue, *Improvements on the Northby algorithm for molecular conformation: better solutions*, J. Global Optim. 4 (1994), pp. 425–440.

[352] G. L. Xue, A. J. Zall and P. M. Pardalos, *Rapid evaluation of potential energy functions in molecular and protein conformations*, in Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos et al., eds., Amer. Math. Soc., Providence, RI, 1996, pp. 237–249.

[353] D. M. York, T. A. Darden and L. G. Pedersen, *The effect of long-range electrostatic interactions in simulations of macromolecular crystals: a comparison of the Ewald and truncated list methods*, J. Chem. Phys. 99 (1993), pp. 8345–8348.

[354] D. M. York, A. Wlodawer, L. G. Pedersen and T. A. Darden, *Atomic-level accuracy in simulations of large protein crystals*, Proc. Natl. Acad. Sci. USA 91 (1994), pp. 8715–8718.

[355] T. Y. Young and K. Fu, *Handbook of Pattern Recognition and Image Processing*, Academic Press, New York 1986.

[356] K. Yue and K. A. Dill, *Inverse protein folding problem: designing polymer sequences*, Proc. Natl. Acad. Sci. USA 89 (1992), pp. 4136–4167.

[357] G. Zanotti, *Protein crystallography*, in Fundamentals of Crystallography, C. Giacovazzo, ed., Oxford Univ. Press, Oxford 1992, pp. 535–597.

[358] G. Zhang and T. Schlick, LIN: *A new algorithm to simulate the dynamics of biomolecules by combining implicit-integration and normal mode techniques*, J. Comp. Chem. 14 (1993), pp. 1212–1233.

[359] G. Zhang and T. Schlick, *The Langevin/implicit-Euler/normal-mode scheme* (LIN) *for molecular dynamics at large time steps*, J. Chem. Phys. 101 (1994), pp. 4995–5012.

[360] F. Zu-Kang and M. J. Sippl, *Optimal superimposition of protein structures: ambiguities and implications*, Folding and Design 1 (1996), pp. 123–132.