Journal of Global Optimization **15:** 235–260, 1999.
© 1999 *Kluwer Academic Publishers. Printed in the Netherlands.*

Surmounting the Multiple-Minima Problem in Protein Folding

HAROLD A. SCHERAGA[1], JOOYOUNG LEE[1], JAROSŁAW PILLARDY[1], YUAN-JIE YE[1], ADAM LIWO[2] and DANIEL RIPOLL[3]
[1]*Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA;* [2]*Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland;* [3]*Cornell Theory Center, Cornell University, Ithaca, NY 14853-3801, USA*

(Accepted in original form 5 July 1999)

Abstract. Protein folding is a very difficult global optimization problem. Furthermore it is coupled with the difficult task of designing a reliable force field with which one has to search for the global minimum. A summary of a series of optimization methods developed and applied to various problems involving polypeptide chains is described in this paper. With recent developments, a computational treatment of the folding of globular proteins of up to 140 residues is shown to be tractable.

Key words: Global optimization, Multiple-minima problem, Protein folding, Structure prediction

1. Introduction

The computation of the three-dimensional structures of globular proteins involves two major problems: the acquisition of a reliable potential energy function, and an adequate procedure to search the conformational space to identify the global minimum of the potential energy among the myriad of local energy minima (the multiple-minima problem) and the nearby low-lying energy minima. This paper is concerned only with the latter question, the attempts to circumvent the multiple-minima problem, and will deal only with methods that we have developed for this purpose. A discussion of other global optimization methods has been presented recently by Wales & Scheraga [1].

2. Methodologies

In applications to global optimization of biological macromolecules, we consider those methods that involve only optimization of the potential energy of the system, without making use of ancillary aids such as secondary-structure prediction, homology modeling, use of fragments from a protein-structure database, etc. The methods described below work well with small peptides and fibrous proteins, but only a few of them, at the present stage of development, seem applicable to large systems such as globular proteins. In most of this work, the potential energy was adop-

ted from the Empirical Conformational Energy Program for Peptides (ECEPP [2], ECEPP/2 [3, 4], or ECEPP/3 [5]), augmented recently by UNRES, a united-residue representation [6–10].

2.1. THE BUILD-UP PROCEDURE

Since an exhaustive enumeration of all possible conformations, i.e., a *systematic search*, is not feasible for proteins or even oligopeptides with more than 5 residues, the build-up procedure was developed [11–16]. This method carries out a *truncated* search, relying on the dominance of short-range interactions. Thus, it finds local minima of short fragments by an exhaustive energy-minimization procedure, and then combines these short fragments into longer ones and, again, minimizes the energies of the longer fragments. Then, a selection of the minima is carried out, keeping those that lie within an appropriately chosen upper bound (the cutoff energy) of the lowest-energy fragment. Subsequently, the limited set of minima of one fragment is combined with the set of another fragment to form larger peptides which are also subjected to energy-minimization. As the fragments grow in size, more of the long-range interactions are taken into account. This process is repeated until the whole chain is eventually built up from its constituent parts.

2.1.1. *Summary of the procedure*

1– A single amino acid residue is the smallest fragment used by the build-up procedure to construct a polypeptide conformation. Vásquez et al. [17] reported the ECEPP/2 minimum-energy conformations of terminally-blocked single residues. These conformations were ordered by increasing energy using a cutoff energy of 5 kcal/mol, and were classified according to the code defined by Zimmerman et al. [18]. The ECEPP/3 force field leads to the same energy minima for all blocked amino acids with the exception of the proline and hydroxyproline residues.

2– Given a molecule with n residues, the conformations of $n-1$ dipeptides are generated from the single-residue data. After energy-minimization, the dipeptides are sorted and are subsequently used to construct tripeptides.

3– Generation of larger fragments of the polypeptide chain involves joining two fragments with one or more residues in common, e.g., tetrapeptides can be constructed from two tripeptides having two residues in common. This process is repeated until the whole polypeptide chain is built.

2.1.2. *Difficulties*

The fact that the number of conformations of fragments that must be energy-minimized and stored at each step increases exponentially constitutes a major drawback of the build-up procedure. Aside from using an energy cut-off, a partial solu-

tion to this problem is to retain only those minima whose backbone conformations differ significantly. This approach considerably reduces the number of conformations to be stored at each stage of the procedure; however, it may lead to problems at later stages because the side-chain rotamers that are most favorable energetically in smaller fragments are not necessarily favored in the whole polypeptide chain.

2.1.3. *Applications*

This procedure has been used to treat open-chain [13, 15, 19, 20] and cyclic peptides [21, 22] and fibrous proteins such as collagen [23–25]. Except for fibrous proteins, where advantage is taken of symmetry relations, the method becomes unmanageable for polypeptide chains containing more than about 20 amino acid residues.

2.2. THE SELF CONSISTENT ELECTROSTATIC FIELD METHOD

Based on a large amount of experimental evidence [26–31], Piela & Scheraga [32] postulated that the native conformation of a protein arises when the electrostatic interactions are near optimal, e.g., they proposed that the peptide group dipoles in the native conformation must have approximately optimal orientations in the electric field generated by the whole molecule and its surrounding solvent. Based on this idea, a conformational search method, named the Self-Consistent Electric Field (SCEF) method was developed. The SCEF procedure was implemented as follows:

1– Starting from an arbitrary conformation of the molecule, minimize the total (e.g., ECEPP/3) conformational energy until the nearest local minimum is reached.
2– For this conformation, calculate the *electric field* due to the whole molecule at each CO and NH group of the peptide units, and also in the center of the C′-N peptide bond.
3– Determine the direction of the electric field with respect to the CO and NH bond dipole moments for all peptide groups, and generate *diagnostic rotations*. A *diagnostic rotation* corresponds to the variation that must be applied to a given torsional angle to obtain the best alignment of the peptide-unit dipole with respect to the local electric field. The electrostatic analysis points specifically to the worst oriented *dipole moment* of the peptide groups (e.g., the group between residues i and $i + 1$). The diagnostic rotation then describes a change of the corresponding backbone dihedral angles ψ_i and ϕ_{i+1} required to align the dipole moment of the unit.
4– Carry out the diagnostic rotation.
5– Use the new conformation of the molecule as the starting point in step 1:
   – if a *new* local minimum is reached, then *repeat* the procedure from step 2 for the new local minimum;

– if the *same* local minimum is encountered, then *repeat step 3*, but use the diagnostic rotation for the next worst-oriented dipole.

6– Steps 1–5 are repeated in a self-consistent manner until further application of the procedure does not change the conformation of the molecule.

2.2.1. *Applications*

The procedure was tested on a 19-residue poly(L-alanine) chain [32] with acetyl- and N-methyl amide terminal blocking groups. The starting conformations were a series of partially α-helical conformations representing different degrees of distortion from the canonical right-handed α*-helix*. The right-handed α-helical conformation corresponds to the global energy minimum of the ECEPP/2 (and ECEPP/3) potential function. In the four cases reported, the procedure was able to achieve the conformation corresponding to the global energy minimum in a very short computation time. The SCEF procedure was also used [33] in a restrictive search of the conformational space of the 58-residue protein bovine pancreatic trypsin inhibitor (BPTI). In this application, the algorithm led to a series of conformations with up to 50 kcal/mol lower than the starting conformation.

2.3. THE MONTE CARLO-MINIMIZATION METHOD

The fact that proteins are not static structures but instead undergo fluctuations was the main motivation for the development of the Monte Carlo-Minimization (MCM) method [34, 35]. The MCM method is a *stochastic approach* [36] for *global optimization* of polypeptides and proteins that combines the strength of the Metropolis Monte Carlo method [37] in global *combinatorial optimization* with that of conventional *energy minimization* to find local minima. Even though the *Metropolis Monte Carlo method* can simulate the thermal processes, straightforward applications of the method to polypeptides were proven to be quite inefficient [38–40]. The main reasons for the lack of success are (a) that only small increments of the variables in each step can be used to sample a very complex conformational space, and (b) that large energy barriers tend to confine the sampling within a very restrictive region of the conformational space. To overcome these difficulties, the MCM method includes conventional energy minimization as a second important feature. Thus, the MCM method generates a Markov walk on the hyper-lattice of all discrete energy minima, with Boltzmann transition probabilities.

2.3.1. *Summary of the procedure*

The MCM procedure is implemented as follows:

1– Given an energy-minimized conformation, C^{min}_{curr}, with total energy E^{min}_{curr}, a sampling strategy is used to generate a perturbed conformation C_{pert}. This

Monte Carlo sampling strategy consists of random changes, involving k dihedral angles of the total number N_{dieh} used to describe the molecule. These changes are generated with probabilities 2^{-k} ($k = 1, 2, ..., N_{dieh}$). This selection of probabilities implies that fluctuations involving more degrees of freedom are sampled with successively lower probabilities. Since any local minimum is accessible from any other one after a finite number of random sampling steps, this sampling strategy satisfies ergodicity requirements.

2– The conformation C_{pert} is then subjected to conventional minimization of its potential energy until it reaches the nearest local minimum. The energy-minimization process is carried out with the Secant Unconstrained Minimization Solver (SUMSL) algorithm [41]. The resulting conformation, C_{pert}^{min}, has a total energy E_{pert}^{min} and is, in general, free of atomic overlaps.

3– The Metropolis criterion is used to decide which conformation, C_{pert}^{min} or C_{curr}^{min}, is to be kept. The following criterion is used for acceptance: if the energy difference $\Delta E = E_{pert}^{min} - E_{curr}^{min} < 0$, or (when $\Delta E > 0$) if $e^{-\Delta E/RT}$ is greater than a randomly generated number between 0 and 1, the new conformation, C_{pert}^{min} replaces the current C_{curr}^{min}; otherwise, C_{pert}^{min} is discarded.

2.3.2. *Applications*

The MCM procedure was applied successfully to study the conformational preferences of the pentapeptide Met-enkephalin [34, 35].

2.4. THE ELECTROSTATICALLY DRIVEN MONTE CARLO METHOD

The Electrostatically Driven Monte Carlo (EDMC) method [42–44] is also an iterative procedure for searching the conformational hypersurface of polypeptides consisting of up to 20 amino acid residues. The EDMC method incorporates the best features of the SCEF and MCM methods and combines them with a set of new techniques that leads to a more efficient search of the conformational space.

The search for the global energy minimum of a molecule proceeds as a ‘quasi-random walk’ along a conformational pathway. The pathway followed by the EDMC method is defined by a sequence of energy-minimized conformations encountered over an unbounded number of iterative steps of the algorithm. In practice, however, the number of iterations is finite and is specified by the user at the beginning of the simulation. The underlying assumption behind the EDMC method is that (*a*) the electrostatic interactions and (*b*) thermal fluctuations, both compete in determining the conformation of the polypeptide chain. The electrostatic interactions should lead to conformations representing an improvement of the charge distribution, i.e., the new conformations are expected to have lower electrostatic and total energies; while the thermal fluctuations, on the other hand, introduce disorder within the molecule. These thermal effects could force the molecule to

adopt higher-energy conformations, but may allow the protein to escape from stable local minima of relatively high energy.

The implementation of these ideas is accomplished as follows: Thermal effects are associated with random changes in the molecular conformation, i.e., a small set of randomly-chosen variables is altered randomly. The reordering effect of the electrostatic interactions is viewed, as in the SCEF method, as a tendency of all permanent dipole moments of the polypeptide to attain their best possible alignment in the local electric field produced by the rest of the molecule. In addition, a series of new features [44] has been included in the latest implementation of the EDMC method that accelerates the search and optimizes the process of generation of new conformations.

2.4.1. *The procedure*

An unfolded state of the polypeptide chain, in which the initial values of the variables describing the molecular conformation are assigned randomly, is usually the first accepted conformation on the conformational pathway followed by the EDMC method. The energy of this conformation is minimized to relieve all possible atomic overlaps. The subsequent accepted conformations are obtained through a series of iterations using a variety of techniques described below. An *iteration* of the procedure is defined as a set of manipulations of the currently accepted conformation that leads to its *replacement* by a newly generated conformation.

a) An electrostatic analysis similar to that produced by the SCEF method [32], but extended to consider the permanent dipole moments of polar side-chains, is one of the techniques that the EDMC method uses to generate new conformations. As the first step of an iteration, an electrostatic analysis of the currently accepted conformation is carried out. This analysis is used to determine the alignment of the permanent dipoles with the local electric field produced by the whole molecule. As a result, a series of *diagnostic rotations* that could improve the local dipole alignments with the electric field are produced. The diagnostic rotations are incorporated into a *prediction list* of possible conformational changes and used in subsequent steps within the iteration to generate new conformations.

b) Since none of these predictions may lead to an acceptable conformation, a random and/or biased sampling technique is also used to generate additional conformations. The following procedure is used:
   1. Specification of the mode in which the variable dihedral angles of the *selected residues* are to be altered:
      (i) Select all variables at random;
      (ii) Select the backbone variables randomly within specific regions of the ϕ-ψ map;
      (iii) Select all variables from pre-computed low-energy conformations of the tripeptides included in the sequence;

(iv) Select backbone variables compatible with regular structures (β-sheets or α-helices).

2. Random selection of (i) the number of residues to be affected by the changes, and (ii) their positions in the sequence.

The latest implementation of the EDMC method [44] includes a technique to produce a *cluster analysis* of the conformations. Conformations corresponding to the accepted minima are grouped into clusters using rmsd (root-mean-square deviation) criteria and ranked on the basis of their total energies. In addition, every generated conformation, even if rejected, is associated with an existing cluster or family, but added to it only if its energy is lower than the one corresponding to the best member of that family. The low-energy conformations included in any of the clusters (with the exception of the cluster containing the current accepted minima) can be used within an iteration to generate conformations randomly, using the protocol described in item (b) above.

Conformations generated by any of these two procedures (a or b) are subjected to minimization of the total energy. A newly generated conformation must fulfill two criteria to be accepted:

1. Any generated conformation corresponding to an already accepted minimum that has been encountered more than a pre-defined number of times (usually 5–10) is automatically excluded from further consideration. This analysis of the long–term behavior of the search constitutes one of the criteria to ensure that the search does not become trapped in a set of local minima of the conformational space.
2. When a conformation satisfies the condition stipulated above, its energy E_{new} is compared with the energy, E_{curr}, of the current accepted conformation, and the Metropolis criterion [37], as described for the MCM method, is applied.

If the energy of the new conformation satisfies both tests, the conformation is accepted, replacing the current one, and a new iteration begins.

2.4.2. *Backtrack*

Within an iteration, it may happen that neither the set of electrostatic predictions, nor the set of randomly generated conformations (usually 100 to 200 conformations) produces an acceptable conformation. Under these circumstances, the algorithm assumes that the current local minimum is quite stable and a new procedure named *backtrack* is triggered. The backtrack procedure attempts to displace the search to a different region of the conformational hypersurface by altering the processes of generation and acceptance of conformations in a substantial manner.

The backtrack procedure involves the following:

a) A new set of conformations is generated by changing a large number of variables simultaneously. In particular, the procedure tends to select the variables associated mainly with the backbone of the polypeptide chain; and,

b) the temperature parameter, T, used on the acceptance test is raised either (i) abruptly to a very high value, or (ii) steadily increased by means of a pre-defined heating scheme.

The backtrack procedure proceeds until the acceptance test is satisfied, or until the number of generated conformations reaches a predetermined value. In the first case, the temperature parameter is reset to its original user-specified value, and the generation mechanism is switched back to the standard protocol described above. If the latter situation occurs, the run is terminated based on the assumption that it is practically impossible to escape from the current region of the conformational space.

It should be noted that raising the temperature during backtrack has the effect of increasing the probability of acceptance of conformations with energies much higher than the current local minimum. The backtrack mechanism has been shown to be an effective technique to help the search avoid being trapped in stable, high-energy regions of the conformational space.

2.4.3. *Applications*

The multiple-minima problem has been found to be computationally surmountable by the EDMC method on existing computers for polypeptides sequences consisting of up to 20 amino acid residues.

In applications to Met-enkephalin [43], oxytocin [45], arginine-vasopressin [45], decaglycine [46], a 19-residue chain of poly(L-alanine) [42], and the 20-residue membrane-bound portion of melittin [44], the EDMC algorithm has converged to unique conformations presumed to be the global energy minima for those particular sequences.

In other applications, to a seven-residue peptide epitope [47], and a twelve-residue analogue of mastoparan and mastoparan X [48], the method identified very low-energy conformations, but it is not certain that the global energy minima were attained in these cases.

The EDMC method has also been applied to explore the conformational space of larger molecules; however, these searches were restricted to regions of the conformational space close to the native conformations. In an application to the 58-residue protein BPTI [33], the algorithm produced the lowest energy conformation known for BPTI using the ECEPP/2 or ECEPP/3 potential. Recently, the EDMC method has also been used to search the conformational properties of a non-oncogenic p21 protein [49] and a molecular switch designed as a biological logic gate [50].

Recently, the method has also been used quite successfully to study the variation of the conformational properties of a series of oligo- and polypeptides with pH [51–53].

2.5. THE SELF-CONSISTENT MULTITORSIONAL FIELD (SCMTF) METHOD

The SCMTF method [54–56] is based on the fact that the ground-state solution of the Schrödinger equation gives information about the location of the global minimum of a potential function, even for a potential with a very complex structure, i.e., the maximum of the square of the ground-state wave function is very often close to the global minimum. Since it is not possible to solve the many-body Schrödinger equation exactly, the mean field approximation is used. The Schrödinger equation for the motion of the nuclei is given by

$$\hat{H}\Psi = E\Psi, \tag{1}$$

where the Hamiltonian operator $\hat{H}$ is defined by

$$\hat{H} = -\sum_{n=1}^{M} \frac{\hbar^2}{2m_n}\Delta_n + \hat{V}, \tag{2}$$

where Δ_n is the Laplacian operator and $\hat{V}$ is the potential energy operator. Bond lengths and bond angles are kept fixed, so the configurational space is defined by the dihedral angles $\theta = (\theta_1, \ldots, \theta_N)$. In order to solve the Schrödinger equation in dihedral angle space, the Hamiltonian from Equation 1 must be transformed appropriately. Assuming that the Hamiltonian may be approximated by diagonal terms only, the resulting operator in torsional space is given by

$$\hat{H}^{mod} = -\sum_{i=1}^{N} \frac{\hbar^2}{2I_i}\frac{\partial^2}{\partial\theta_i^2} + \hat{V}, \tag{3}$$

where I_i is an averaged moment of inertia. The solution may be approximated by the Hartree-like product of the normalized one-angle wave functions $\phi_i(\theta_i)$ leading to a set of N coupled one-dimensional equations

$$\hat{H}_i\phi_i^{k_i} = \epsilon_i^{k_i}\phi_i^{k_i} \quad i = 1, \ldots, N. \tag{4}$$

The Hamiltonian for the single dihedral angle is given by

$$\hat{H} = -\frac{\hbar^2}{2I_i}\frac{d^2}{d\theta_i^2} + \hat{V}_i^{eff}(\theta_i), \tag{5}$$

where the effective potential $\hat{V}_i^{eff}(\theta_i)$ depends on the mean field created by averaging over the other dihedral angles, according to the probability density distribution $|\phi_i^0|^2$. The above set of equations (Equation 4) is solved iteratively, until the probability density distributions converge.

The SCMTF method was tested initially on terminally-blocked alanine [54], and then successfully applied to oligopeptides: met-enkephalin [54], decaglycine [55], icosalanine [55], and melittin [56].

2.6. METHODS BASED ON THE DEFORMATION OF A POTENTIAL FUNCTION

A promising approach to surmount the multiple-minima problem involves methods based on the deformation of the original rugged energy surface, thereby reducing the number of minima by orders of magnitude, at best even to a single minimum, and simplifying the conformational search greatly. Applying a deformation usually alters locations of all minima; therefore, a procedure for tracking minima between the highly deformed function and the undeformed one (*reversing*, or *reversal procedure*) must be used. An application of a deformation method may be divided into two subproblems: (i) designing an effective deformation of the potential function, and (ii) constructing an appropriate reversing procedure.

2.6.1. *The diffusion equation method*

The basic idea of the method, introduced by Piela et al. [57], is to deform the multivariable function that represents the potential energy in such a manner as to make the shallow wells disappear gradually, while other potential wells grow at their expense. Under the assumption that the shallower wells will disappear more easily than the deep wells, it is possible to envision an iterative procedure that, applied to the potential function, will change its shape, making most of the minima become shallower until they disappear, while leaving a single absorbing minimum related to the lowest minimum of the original function. At this point of the *deformation process*, a simple local minimization algorithm should be able to retrieve the position of the unique minimum from any starting point. However, since the deformation of the potential should likely have altered the location of all minima, the global minimum of the original function is not the same as the minimum of the deformed surface. Its location can, in principle, be attained by slowly reversing the deformation and using standard local minimization procedures. Piela et al. showed that the deformation of the hypersurface can be carried out with the aid of the diffusion equation. In this context, the original shape of the potential function has the meaning of an initial concentration (or temperature) distribution.

Transformation operator. In order to show the basic features of the method, it is worth considering a simple example with a one-dimensional function. Given a function $f(x)$, one may define a transformation by adding its second derivative:

$$f^{[1]}(x) = f(x) + \beta f''(x) \quad \text{for} \beta > 0. \tag{6}$$

The transformation of Equation (6) destabilizes any potential well of *f(x)* by decreasing its depth, i.e., the inflection points of $f(x)$ do not undergo any change, since they correspond to $f'' = 0$, while the regions of the curve where the function is convex or concave will be shifted upward or downward, respectively. For small values of the positive constant β, the net effect of this transformation is that existing extrema of the curve are destabilized. The procedure may be repeated for the new curve $f^{[1]}(x)$ leading to a new transformation. If this procedure is applied

iteratively, we obtain for the N-th iteration

$$f^{[N]}(x) = \left[1 + \frac{t}{N}\frac{d^2}{dx^2}\right]^N f(x); \quad \text{for } \beta > 0, \tag{7}$$

where β has been replaced by t/N, with $t > 0$ being a parameter. The destabilization of the surface is most effective when N tends to infinity. Under this assumption, Equation (7) can be transformed into:

$$\begin{aligned} F(x,t) &= \lim_{N\to\infty}\left(1 + \frac{t}{N}\frac{d^2}{dx^2}\right)^N f(x) \\ &= exp\left(t\frac{d^2}{dx^2}\right) f(x) \\ &= T(t)\, f(x), \end{aligned} \tag{8}$$

where $T(t)$ is defined as:

$$\begin{aligned} T(t) &= \exp\left(t\frac{d^2}{dx^2}\right) \\ &= 1 + t\frac{d^2}{dx^2} + \frac{1}{2!}\left(t\frac{d^2}{dx^2}\right)^2 + \ldots + \frac{1}{k!}\left(t\frac{d^2}{dx^2}\right)^k + \ldots, \end{aligned} \tag{9}$$

where a Taylor series representation of the exponential operator $\exp\left(t\frac{d^2}{dx^2}\right)$ was used in Equation (9). The operator $T(t)$ has some useful properties. First, it is linear and its eigenfunctions are $\sin\omega x$ and $\cos\omega x$, ω being a real constant:

$$T(t)\,\sin\omega x = a(\omega, t)\,\sin\omega x \tag{10}$$

$$T(t)\,\cos\omega x = a(\omega, t)\cos\omega x \tag{11}$$

where the eigenvalues $a(\omega, t)$ are expressed as

$$a(\omega, t) = \exp\left(-\omega^2 t\right). \tag{12}$$

Because of the factor $a(\omega, t)$, the operator $T(t)$ has the property of flattening sines and cosines. The flattening effect is more pronounced for high-frequencies, i.e., functions with large ω's. Thus, high-frequency components of $f(x)$ will vanish first when the operator $T(t)$ is applied, and the new function $T(t)\,f(x)$ is usually left with many fewer minima.

Diffusion equation. When the Taylor series given in Equation (9) converges, its sum is a solution of the diffusion or heat conduction equation:

$$\frac{\partial^2 F}{\partial x^2} = \frac{\partial F}{\partial t} \tag{13}$$

where the variable t represents time. Additionally, Equation (13) is solved with the initial condition $F(x, 0) = f(x)$. The function F usually represents a concentration or a temperature distribution. If the function $f(x)$ is bounded, a solution of Equation (13) exists for any positive value of t. The procedure described above represents a spontaneous mass transport (or flow of heat) in a medium for an initial distribution of concentration (or temperature) given by the function $f(x)$ (which in our case represents the conformational energy). Governed by the diffusion equation and independent of the initial conditions, the concentration (or temperature), will evolve with time in such a manner that it will become constant for $t = \infty$. However, it is expected that the concentration (or temperature) will exhibit a single minimum for certain (very large) finite values of t. This single minimum should represent the last trace of the potential well corresponding to the global minimum of the original hypersurface $f(x)$.

2.6.1.1. *Extension of the procedure to higher dimensions.* This deformation procedure can be extended to higher dimensions. By analogy with Equation (6), in the case of an n dimensional space, the function $f(\mathbf{x})$ (with $\mathbf{x} = [x_1, x_2, \ldots, x_n]$) is destabilized by adding the trace of the Hessian. For the n-dimensional case, the operator T(t), appearing in Equation (8), is replaced by

$$T(t) = T_1(t)\, T_2(t) \ldots T_n(t) \tag{14}$$

with

$$T_i(t) = \exp\left(t \frac{\partial^2}{\partial x_i^2}\right). \tag{15}$$

The operator $T_i(t)$ has the property

$$T_i(t) f(x_j) = f(x_j) \quad \text{for } j \neq i\ . \tag{16}$$

The function $F(\mathbf{x}, t)$ appearing in Equation (8) also satisfies the diffusion equation for the multidimensional case

$$\Delta F = \frac{\partial F}{\partial t} \tag{17}$$

where the operator $\partial^2/\partial x^2$ in Equation (13) has been replaced by the Laplacian:

$$\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}\ . \tag{18}$$

2.6.2. *The Distance Scaling Method (DSM)*

The DSM [58, 59] is another procedure to deform the potential energy hypersurface. Instead of solving the diffusion equation, the deformation is carried out by scaling the distance variables in the potential energy function. In the DSM method, the site-site distance r_{ij} is transformed as follows:

$$\tilde{r}_{ij}(t) = \frac{r_{ij} + t r_{\circ,ij}}{1 + bt} \tag{19}$$

The parameter $r_{\circ,ij}$ in Equation (19) has the meaning of the position of the minimum in the pairwise-interaction term under consideration. On increasing t, the original function is flattened, but the position of its minimum and the function value at the minimum remain the same, if a value of the parameter b is taken as 1 (as in the original formulation of the DSM). The parameter b controls the position of the minimum and remains constant during the calculations. If the parameter b is greater than 1, this means that the position of the minimum of the deformed site-site function will shift to larger values while, for $b < 1$, it will shift towards zero, and a two-body potential will become purely attractive for $t = 1/(1 - b)$. For the parameter $b = 0$, the deformation becomes especially simple; the original two-body function is shifted toward the origin of the coordinate system (Shift Method (SM) [58]). It is relatively easy to choose $r_{\circ,ij}$, if there is a minimum. However, there are two-body functions (e.g., for electrostatic interactions) that are monotonic. In this case, it is reasonable to choose $r_{\circ,ij}$ so large that the function value at this point is close to zero; thus, this energy contribution will effectively be eliminated for large deformation.

The potential energy function $F(\mathbf{x})$ is deformed by applying the transformed site-site distance to all its two-body components:

$$F(\mathbf{x}, t) = \sum_{i,j} u_{ij}[\tilde{r}_{ij}(t)] \tag{20}$$

where u_{ij} is a two-body potential function. The two-body components of the potential function are flattened and, therefore, barriers between different basins of local minima of the potential function $F(\mathbf{x}, t)$ gradually disappear (while t increases), resulting in merging minima and a decrease in their number.

2.6.3. *Reversing procedure*

As mentioned earlier, the positions of minima of the deformed function $F(\mathbf{x}, t_0)$ are, in general, different from those of the original function f. Furthermore, the position of a single minimum of $F(\mathbf{x}, t_0)$, cannot generally be used as a starting point in a minimization of $f(\mathbf{x})$, since it will probably lead to a local minimum not related to the starting one. Consequently, a reversing procedure must be used to retrieve the relations between minima of the deformed and undeformed functions. The simplest version of this procedure consists of a series of local minimizations

carried out on gradually less deformed surfaces, where the starting point on the less deformed surface $F(\mathbf{x}, t - \Delta t)$ (where Δt is a small deformation interval) is a result of local minimization on $F(\mathbf{x}, t)$.

It was hoped that tracking the lowest-energy minimum obtained with the maximally deformed energy surface back to the original energy surface would locate the global minimum of the original energy function [57] (*single trajectory approach*). The main advantage in this case is that the method is completely deterministic. However, this approach works only for relatively simple systems; in more complex cases, the global minimum in the original energy surface is represented as a higher-energy minimum in the deformed energy surface and *vice versa*. Moreover, during the reversal of the deformation, a single trajectory often branches, forcing the algorithm to track only one possibility (tracking all of them is effectively impossible because of the exponential growth of the number of trajectories as the reversal progresses). This problem also remains when a *multiple trajectory* approach is used, in which all minima in the deformed energy surface are traced back to the original energy surface. An attempt to alleviate this problem is the *multiple trajectory perturbation approach (MTPA)*. In this approach, each structure encountered at a particular reversal step of the deformation is perturbed and then energy-minimized, and a pre-defined number of lowest-energy structures is taken to the next step of the reversal; the addition of this perturbation step alleviates the problem of splitting the trajectories as the deformation decreases. This approach is by far more successful than the single or multiple-trajectory approaches and was applied in the theoretical prediction of crystal structures [60, 61]. However, it does not work for highly demanding applications, such as very large Lennard-Jones clusters or large polypeptide chains.

The most recent approach to global optimization, using the idea of potential function deformation, is the utilization of a local search with self-consistent mapping of the deformed and undeformed minima (Self-Consisted Basin-to-Deformed-Basin Mapping, SCBDBM) [62, 63]. The underlying principle is the location of large regions of conformational space containing low-energy minima by coupling them to some of the greatly reduced number of minima on the highly deformed surface. The whole procedure consists of macro-iterations, in which the parameter t controls the deformation changes between two extreme values, t_{max} and t_{min} ($t = 0$ corresponds to the original energy surface). The first macroiteration is initialized with randomly-generated conformations, while the next macroiterations are fed with the results of the previous ones. Each macroiteration consists of the following steps: (i) reversal of the deformation from t_{max} to t_{min}; the reversal is accompanied by carrying out a limited search in the neighborhood of the minima at each stage of the reversal; (ii) collection of the new low-energy minima in the t_{min}-deformed energy surface; (iii) tracking the images of these minima while increasing the deformation, up to t_{max}. Steps (i)–(iii) are iterated, until no new minima are found or a pre-defined number of iterations is exceeded. In the initial macroiteration, t_{min} is greater than 0 and t_{max} is chosen so that the deformed energy surface has only a

few minima. In each next macroiteration, the new t_{max} is set at t_{min} of the previous macroiteration and t_{min} is decreased, to reach 0 in the last macroiteration.

2.6.4. *Applications*

The DEM with a single trajectory reversing procedure has been applied to:

- Clusters of Lennard-Jones atoms with 8–19, 33 and 55 atoms in a cluster; the global minimum was found for all of them (except 8, 9 and 12) [64].
- Water clusters [65].
- A single terminally blocked alanine [66].
- The pentapeptide Met-enkephalin [66] for which the method led to practically the same global-minimum backbone structure obtained by other methods. The test, however, was carried out under more restrictive conditions since only the backbone dihedral angles ϕ and ψ were considered as variables.

The simplified version of the DSM (SM) has been applied successfully with a single trajectory reversing procedure to small Lennard-Jones atomic clusters [58] and water clusters [67]. The DSM coupled with molecular dynamics as a searching tool and a multi-trajectory reversing procedure has been applied to Lennard-Jones clusters containing up to 66 atoms, and was able to locate the non-icosahedral global minimum for LJ_{38} [59].

The DEM and the DSM have also been applied to predict crystal structures without making use of ancillary information such as the space group. The multi-trajectory reversing procedure with perturbations was used in this case. The DEM and DSM were used to predict the crystal structure of the rigid, nonpolar molecule, hexasulfur, with a Lennard-Jones interaction potential, and that of the rigid, polar molecule, benzene [60, 61]. Fixing only the molecular geometry and the interaction potential, the unit cell dimensions, space groups and the number of molecules in the unit cell were all computed, and the experimental crystal structures were located successfully. For benzene, the calculation succeeded even when the number of molecules in the unit cell was allowed to be twice the experimental value, which made the global optimization problem considerably harder.

The SCBDBM method has been applied to united-residue polyalanine chains with a length of up to 100 residues and to locate the currently known lowest-energy conformation of staphylococcal protein A [62]. So far, it has successfully located very low energy structures of polyalanine chains, predicting that the most stable structure is a straight α-helix up to 70 residues. For 70–80 residues the most stable form is bent in the middle of the α-helix and, from 80 residues upward, the most stable structure is a three-helix bundle. For Protein A, a minimum very close to the experimental structure has been located. In another application, the SCBDBM has also been able to locate global minima for Lennard-Jones clusters of sizes up to 100 atoms, except those consisting of 75–78 atoms [63].

2.7. THE CONFORMATIONAL SPACE ANNEALING (CSA) METHOD

The CSA method [68–70] combines essential aspects of the build-up procedure and a genetic algorithm. The CSA method searches the whole conformational space in its early stages and then narrows the search to smaller regions with low energy as the distance cut-off, D_{cut}, which defines the similarity of two conformations, is reduced. The distance between conformations i and j, D_{ij}, is defined in terms of the differences between all variable dihedral angles that define the geometry of the polypeptides [68–70]. A flow chart of the CSA algorithm is presented in Figure 1. As in genetic algorithms [71], CSA starts with a pre-assigned number (usually 50) of randomly generated and subsequently energy-minimized conformations. This pool of conformations is called the *bank*. At the beginning, the bank is a sparse representation of the entire conformational space. A number of dissimilar conformations (usually 20) are then selected from the bank, excluding those that have already been used; they are called *seeds*. Each seed conformation is modified by changing from one to one-third of the total number of variables pertaining to a contiguous portion of the chain; the new variables are selected from one of the remaining bank conformations, rather than being picked at random. Each conformation is energy minimized to give a trial conformation. Thirty trial conformations are generated for each seed (a total of 600 conformations). This is the most time-consuming part of the computation, but it is highly suitable for parallel computing [70]. For each trial conformation, α, the closest conformation A from the bank (in terms of distance $D_{\alpha A}$) is determined. If $D_{\alpha A} < D_{\text{cut}}$ (D_{cut} being the current cut-off criterion), α is considered similar to A; in this case α replaces A in the bank, if it is also lower in energy. If α is not similar to A, but its energy is lower than that of the highest-energy conformation in the bank, B, α replaces B. If neither of the above conditions holds, α is rejected. The narrowing of the search is accomplished by setting D_{cut} to a large value initially (usually one-half of the average pair distance in the bank) and gradually reducing it as the search progresses. Special attention is paid to selecting seeds that are far from each other in conformational space. One round of the procedure is completed when there is no seed to select (i.e., all conformations from the bank have already been used). The round is repeated a predetermined number of times. The greatest advantage of the CSA method is that it always finds distinct families of low-energy conformations.

With the ECEPP/3 all-atom force field, the CSA method has been successful in obtaining the global minimum of peptides containing up to 20 amino acid residues with 113 variable dihedral angles [68–70]. The average wall clock time to find the global minimum of a polypeptide with 20 amino acid residues, based on twenty-four independent runs, was about 4.5 hours with 32 processors of an IBM SP2 supercomputer. The corresponding wall clock time for a pentapeptide was 36 seconds with 16 processors [70]. However, an extensive search of the conformational space of globular proteins ($\sim$ 100 amino acid residues), represented

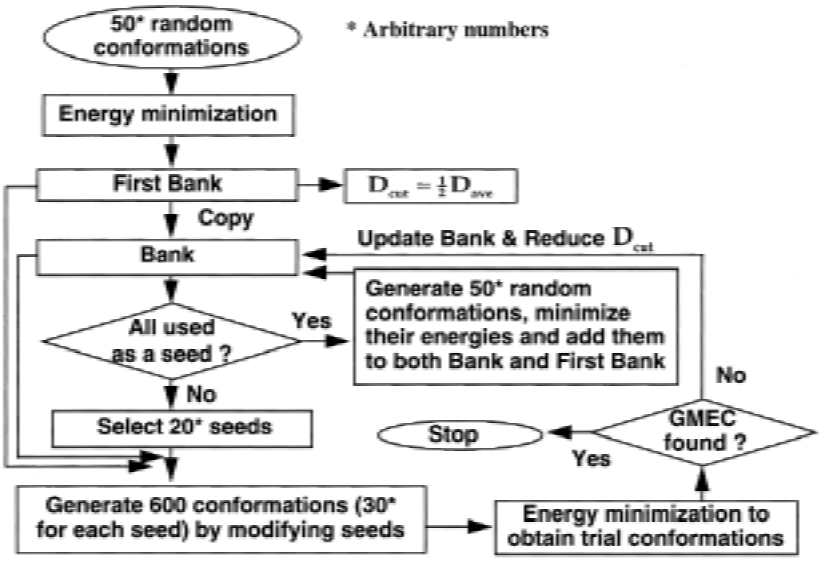

Figure 1. Flow chart of the CSA algorithm.

by an all-atom force field such as ECEPP/3, is out of the question because of the astronomical computational cost.

Recently, we have developed a united-residue force field, UNRES [6–10], with which the search of the conformational space of globular proteins ($\sim$ 100 amino acid residues) is treatable by efficient optimization methods such as CSA. With the UNRES force field, the CSA method successfully located the native-like conformations of two helical proteins (the 10–55 residue fragment of protein A and the 75 residues of apo calbindin D9K) among the low-energy ones [72]. Alternative structures (which were mirror images of the native folds) were also found. The global minimum of a 46-residue fragment of protein A was also identified. The average wall clock time to find the global minimum based on seven independent runs was about 14 hours with 32 processors of an IBM SP2 supercomputer. Details of the implementation of the CSA procedure with the UNRES force field are provided elsewhere [72]. It should be noted that the structures of protein A or apo calbindin D9K or the class of fold that they represent were not used in the optimization of the force field.

The ensemble of conformations of protein A generated by the CSA method have also been used to investigate the kinetics of the folding transitions to see how these conformations evolve toward the global minimum. The kinetics of protein folding were described by a master equation that was described by a Laplace transformation [73]. The calculations show that native proteins fold fast cooperatively, and that the global minimum can be reached after a sufficiently long folding time regardless of the initial state or of the existence of local energy minima. The conformation of a protein molecule can transform from non-native states to the native state even if

it originates in different conformational families, as in the case of the mirror-image family of protein A. Furthermore, it was found that a protein molecule can fold to its global minimum through different paths from different starting conformations. A protein molecule adopts a set of conformations, when the folding reaches the equilibrium distribution, in which the global minimum has the largest probability. This is the basis for simulating the folding of a native protein by searching for the global minimum on its potential energy hyper-surface.

2.8. THE HIERARCHICAL APPROACH TO GLOBAL OPTIMIZATION

Whereas the foregoing methods work well with small systems, a new hierarchical approach [75] has the potential of treating larger systems, e.g., globular proteins. The hierarchical approach is based on two recent developments, a united-residue force field, UNRES [6–10] and the CSA method [68–70, 72]. An extensive conformational search is carried out with the CSA method using the UNRES force field.

Once a set of families of low-energy united-residue conformations has been identified by the CSA method, they are subsequently converted to all-atom chains in the following steps [7]: (a) positioning of the peptide groups between consecutive C^{α}'s so as to achieve optimal alignment of the peptide group dipoles (the dipole-path method) [6]; (b) further optimization of the *backbone* conformations using the Electrostatically-Driven Monte Carlo (EDMC) method [44]; (c) adding the side chains with partial optimization of their degrees of freedom; (d) final refinement of the all-atom chains using the EDMC method and exploration of the flexible loop regions with the use of the Gō-Scheraga algorithm [76, 77]. In the all-atom calculations, the ECEPP/3 [5] force field with the SRFOPT solvation free energy contribution [78] is used. The SRFOPT set of parameters seems to work better [78] than other solvation parameter sets when used in combination with the ECEPP/3 force field.

The critical step of the algorithm is the global conformational search at the united-residue level. Because the structures are selected for further stages based on the energy relations calculated using the united-residue force field, the quality of this force field is of central importance. We provide here a short description of our UNRES force field; for details, the reader is referred to the original papers [6–10].

In the UNRES force field, a polypeptide chain is represented by a sequence of C^{α} atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p) located in the middle between the consecutive C^{α} atoms (Figure 2). All the virtual bond lengths (i.e. C^{α}–C^{α} and C^{α}–SC) are fixed, while the backbone as well the virtual-bond angles can vary. The free energy of the virtual-

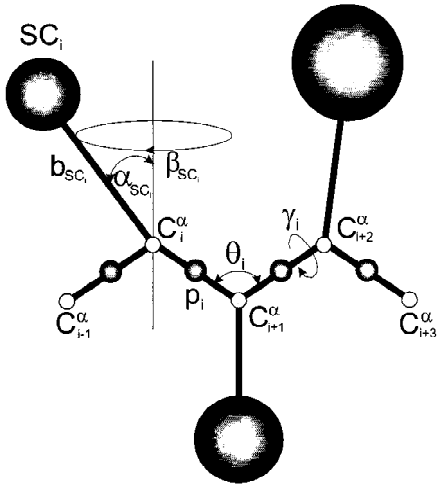

Figure 2. United-residue representation of a polypeptide chain. The interaction sites are side-chain centroids of different sizes (SC) and the peptide-bond centers (p) are indicated by dashed circles, while the α-carbon atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual C^{α}–C^{α} bonds have a fixed length of 3.8 Å, corresponding to a trans peptide group; the virtual-bond (θ) and dihedral (γ) angles are variable. Each side chain is attached to the corresponding α-carbon with a fixed 'bond length', b_{SC_i}, variable 'bond angle', α_{SC_i}, formed by SC_i and the bisector of the angle defined by C^{α}_{i-1}, C^{α}_i, and C^{α}_{i+1}, and with a variable 'dihedral angle' β_{SC_i} of counterclockwise rotation about the bisector, starting from the right side of the C^{α}_{i-1}, C^{α}_i, C^{α}_{i+1} frame.

bond chain is expressed by Equation (21).

$$U = \sum_{i<j} U_{SC_i SC_j} + \sum_{i \neq j} U_{SC_i p_j} + w_{el} \sum_{i<j-1} U_{p_i p_j} + w_{tor} \sum_i U_{tor}(\gamma_i) + w_{loc} \sum_i [U_b(\theta_i) + U_{rot}(\alpha_{SC_i}, \beta_{SC_i})] + w_{corr} U_{corr} \tag{21}$$

The term $U_{SC_i SC_j}$ consists of the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains. It, therefore, implicitly contains the contributions arising from the interactions with the solvent. The terms $U_{SC_i p_j}$ denote the excluded-volume potential of the side-chain – peptide-group interactions. The peptide-group interaction potential ($U_{p_i p_j}$) accounts mainly for the electrostatic interactions between them or, in other words, for their tendency to form backbone hydrogen bonds. U_{tor}, U_b, and U_{rot} denote the energies of virtual-dihedral angle torsions, virtual-angle bending, and side-chain rotamers; these terms reflect the local propensities of the polypeptide chain. Finally, the multibody (or cooperative) term U_{corr} arises from the fact that details of the all-atom chain are lost when converting it into the simplified chain. Their functional forms can be derived taking advantage of the fact that the free energy function of the simplified chain can be obtained by integrating the Boltzmann factor of the energy of the all-atom chain over

Table 1. Summary of the results for the best-predicted contiguous fragments of the seven targets submitted to CASP3

Model[a]	No. of aa[b]	Fragment[c]	rmsd[d]	Percentage[e]
T0056_1	114/114	59 (71–129)	5.8	52
T0061_1	89/76	61 (25–85)	4.2	80
T0073_1	48/48	30 (1–30)	1.0	63
T0074_2	98/95	53 (160–212)	5.8	56
T0076_3	140/139	62 (76–137)	6.8	45
T0079_1	129/116	61 (9–68)	5.9	52
T0084_1	37/30	13 (21–33)	0.9	42

[a] The last digit indicates the model number of the target. Up to five models were submitted for each target. The total number of models submitted is 22.
[b] The numbers correspond to: length of the (entire sequence)/ (experimentally observed).
[c] The first number indicates the size of the fragment analyzed. In parentheses, the first and last residues of the fragment are given.
[d] The calculation of the rmsd (in Å) is carried out by using the positions of the C^{α} carbon atoms of the two fragments at optimal superposition requiring one to one match of corresponding residues.
[e] Ratio (in percentage) of residues predicted over those observed in the experimental structures.

'less important' degrees of freedom, given the configuration of the simplified chain [10]. The w's denote relative weights of the respective energy terms. The force field was parameterized based on distribution and correlation functions calculated from protein structures from the PDB or by averaging the all-atom energy functions, as well as by Z-score optimization (maximizing the ratio of the gap, between the energy of the native structure and the lowest-energy non-native structure, to the average energy of the non-native structures).

With this approach, we attempted blind predictions on seven target proteins provided for the Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP3) [79, 80]. The amino acid sequences of these targets had been volunteered by experimental structural biologists who were in the process of determining their three-dimensional structures by NMR spectroscopy or X-ray crystallography. The results of the conformational search carried out with CSA were very promising. The method identified distinct families of low-energy conformations some of which share a common core (see Table 1). For target T0061, an 89 amino acid residue-protein (HDEA, *E. coli* [81], PDB entry: 1BG8), the structure of the core represents 80% of the experimentally-observed structure. In this particular case, the rmsd between the crystal and our prediction was 4.2 Å for the C^{α} atoms [75, 80].

In addition, large portions ($\sim$ 60 amino acid residues) of the three-dimensional structures of the target proteins with sizes ranging from 89 to 140 amino acid residues were predicted correctly. The rmsd's for the C^{α} atoms from the experi-

mental structures ranged from 4.2 to 6.8 Å , demonstrating the robustness of our approach.

It must be pointed out that our simulations were carried out on *isolated* polypeptide chains (with inclusion of solvent effects). Some of the targets that we predicted exist as complexes or in multimeric forms in the crystals examined by X-ray diffraction. Our treatment does not include all these possible arrangements. However, in these cases, the conformational search method has identified families of conformations that shared similar features with the experimental structures. For example, for proteins with two domains, such as targets T0076 and T0079, our predictions also led to two domains. For T0079, the difference between the experimental and calculated structures arises mainly from the manner in which the two domains associate with each other. The experimental structure of T0079 was determined by X-ray crystallography as a complex with DNA. The DNA molecule was not included in our simulations. The models that we submitted contain two domains, as in the experimental structure. However, these domains are packed tightly together in our models while, in the experimental structure, they bind to DNA. The computational cost for the largest target (140 amino acid residues) was less than 150 hours with 64 processors of an IBM SP2 supercomputer.

3. Conclusions and outlook

In this paper, we discussed global optimization methods that were developed and used in our laboratory to study the protein folding problem. Most of them were applied to study relatively small polypeptides represented by all-atom models. We have demonstrated the usefulness of these methods in studying the protein folding and other hard optimization problems.

The protein folding problem is quite different from other hard optimization problems such as the traveling salesman problem, where the cost function to be optimized is well-defined. In protein folding, the energy function is not provided in an exact form, and an approximate energy function is used. Therefore, finding the global minimum of a given energy function does not necessarily solve the protein folding problem unless the function is accurate enough. However, without efficient search methods, one can never be sure if the solution for a given potential function is really the global minimum. This means that the acquisition of a reliable optimization method is the only way to approach the protein folding problem. A powerful optimization method is a tool that will allow us to judge and improve the performance of a given force field.

In order to treat larger systems in reasonable time, we found it necessary to reduce the complexity of the protein folding problem by using a united-residue force field, UNRES. Our results on globular proteins (containing up to 140 amino acid residues) demonstrate that it is possible to predict a significant portion of protein structure by using only a potential function and a powerful method of con-

formational search, without the aid of knowledge-based information provided by multiple-sequence alignment, secondary-structure prediction, or fold recognition.

There previously were two major obstacles preventing the practical implementation of the energy-based methods: lack of an efficient global optimization method and insufficient quality of the force fields. As we have shown, the first obstacle has been largely overcome by the hierarchical approach, whose decisive stage is the global optimization of the energy of a highly simplified polypeptide chain with the CSA method, as well as the SCBDBM and other efficient global optimization methods developed in our laboratory. Secondly, the force fields (e.g., UNRES, ECEPP) are now being improved systematically by fine-tuning their parameters with the help of global optimization methods. In conclusion, global optimization methods are useful tools for surmounting the multiple-minima problem in protein folding.

References

1. Wales, D.J. and Scheraga, H.A. (1999), Global optimization of clusters, crystals and biomolecules, *Science*, 285: 1368–1372.
2. Momany, F.A., McGuire, R.F., Burgess, A.W. and Scheraga, H.A. (1975), Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions and intrinsic torsional potential for the naturally occurring amino-acids, *J. Phys. Chem.* 79: 2361–2381.
3. Némethy, G., Pottle, M.S. and Scheraga, H.A. (1983), Energy parameters in polypeptides. IX. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids, *J. Phys. Chem.* 87: 1883–1887.
4. Sippl, M.J., Némethy, G. and Scheraga, H.A. (1984), Intermolecular potentials from crystal data. VI. Determination of empirical potentials for O-H$\cdots$O-C hydrogen bonds from packing configurations, *J. Phys. Chem.* 88: 6231–6233.
5. Némethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S. and Scheraga, H.A. (1992), Energy parameters in polypeptides. X. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides, *J. Phys. Chem.* 96: 6472–6484.
6. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S. and Scheraga, H.A. (1993), Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment, *Protein Science* 2: 1697–1714.
7. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S. and Scheraga, H.A. (1993), Prediction of protein conformation on the basis of a search for compact structures; test on avian pancreatic polypeptide, *Protein Science* 2: 1715–1731.
8. Liwo, A., Oldziej, S., Pincus, M.R., Wawak, R.J., Rackovsky, S. and Scheraga, H.A. (1997), A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data, *J. Comput. Chem.* 18: 849–873.
9. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S., Oldziej, S. and Scheraga, H.A. (1997), A united-residue force field for off-lattice protein-structure simulations. II: Parameterization of short-range interactions and determination of the weights of energy terms by Z-score optimization, *J. Comput. Chem.* 18: 874–887.
10. Liwo, A., Kazmierkiewicz, R., Czaplewski, C., Groth, M., Oldziej, S., Wawak, R.J., Rackovsky, S., Pincus, M.R. and Scheraga, H.A. (1998), A united-residue force field for off-lattice protein-

structure simulations. III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials, *J. Comput. Chem.* 19: 259–276.
11. Scheraga, H.A. (1974), Prediction of protein conformation, in C.B. Anfinsen and A.N. Schechter (eds.), *Current Topics in Biochemistry* (pp. 1–42), Academic Press, New York.
12. Simon, I., Némethy, G. and Scheraga, H.A. (1978), Conformational energy calculations of the effects of sequence variations on the conformations of two tetrapeptides, *Macromolecules* 11: 797–804.
13. Pincus, M.R., Klausner, R.D. and Scheraga, H.A. (1982), Calculation of the three-dimensional structure of the membrane-bound portion of melittin from its amino acid sequence, *Proc. Natl. Acad. Sci., USA* 79: 5107–5110.
14. Scheraga, H.A. (1983), Recent progress in the theoretical treatment of protein folding, *Biopolymers* 22: 1–14.
15. Vásquez, M. and Scheraga, H.A. (1985), Use of buildup and energy-minimization procedures to compute low-energy structures of the backbone of enkephalin, *Biopolymers* 24: 1437–1447.
16. Gibson, K.D. and Scheraga, H.A. (1987), Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization, *J. Comput. Chem.* 8: 826–834.
17. Vásquez, M., Némethy, G. and Scheraga, H.A. (1983), Computed conformational states of the 20 naturally occurring amino acid residues and of the prototype residue α-aminobutyric acid, *Macromolecules* 16: 1043–1049.
18. Zimmerman, S.S., Pottle, M.S., Némethy, G. and Scheraga, H.A. (1977), Conformational analysis of the twenty naturally occurring amino acid residues using ECEPP, *Macromolecules* 10: 1–9.
19. Vásquez, M. and Scheraga, H.A. (1988), Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data, *J. Biomol. Struct. & Dyn.* 5: 705–755.
20. Vásquez, M. and Scheraga, H.A. (1988), Variable-target-function and build-up procedures for the calculation of protein conformation. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data, *J. Biomol. Struct. & Dyn.* 5: 757–784.
21. Dygert, M., Gō, N. and Scheraga, H.A. (1975), Use of a symmetry condition to compute the conformation of gramicidin S, *Macromolecules* 8: 750–761.
22. Némethy, G. and Scheraga, H.A. (1984), Hydrogen bonding involving the ornithine side chain of gramicidin S, *Biochem. Biophys. Res. Commun.* 118: 643–647.
23. Miller, M.H. and Scheraga, H.A. (1976), Calculation of the structures of collagen models. Role of interchain interactions in determining the triple-helical coiled-coil conformation. 1. Poly(glycyl-prolyl-prolyl), *J. Polymer Sci.: Polymer Symposia No. 54*, pp. 171–200.
24. Miller, M.H., Némethy, G. and Scheraga, H.A. (1980), Calculation of the structures of collagen models. Role of interchain interactions in determining the triple-helical coiled-coil conformation. 2. Poly(glycyl-prolyl-hydroxyprolyl), *Macromolecules* 13: 470–478.
25. Miller, M.H., Némethy, G. and Scheraga, H.A. (1980), Calculation of the structures of collagen models. Role of interchain interactions in determining the triple-helical coiled-coil conformation. 3. Poly(glycyl-prolyl-alanyl), *Macromolecules* 13: 910–913.
26. Levitt, M. and Chothia, C. (1976), Structural patterns in globular proteins, *Nature* 261: 552–558.
27. Wada, A. (1976), The α-helix as an electric macro-dipole, *Adv. Biophys.* 9: 1–63.
28. Perutz, M.F. (1978), Electrostatic effects in proteins, *Science* 201: 1187–1191.
29. Hol, W.G.J., Halie, L.M. and Sander, C. (1981), Dipoles of the α-helix and β-sheet: their role in protein folding, *Nature* 294: 532–536.
30. Chou, K.-C., Némethy, G. and Scheraga, H.A. (1983), Energetic approach to the packing of α-helices. 1. Equivalent helices, *J. Phys. Chem.* 87: 2869–2881.
31. Hol, W.G.J. (1985), The role of the α-helix dipole in protein function and structure, *Prog. Biophys. molec. Biol.* 45: 149–195.

32. Piela, L. and Scheraga, H.A. (1987), On the multiple-minima problem in the conformational analysis of polypeptides. I. Backbone degrees of freedom for a perturbed α-helix, *Biopolymers* 26: S33–S58.
33. Ripoll, D.R., Piela, L., Vásquez, M. and Scheraga, H.A. (1991), On the multiple-minima problem in the conformational analysis of polypeptides. V. Application of the self-consistent electrostatic field and the electrostatically driven Monte Carlo methods to bovine pancreatic trypsin inhibitor, *Proteins: Struc., Func., and Gen.* 10: 188–198.
34. Li, Z. and Scheraga, H.A. (1987), Monte Carlo-minimization approach to the multiple-minima problem in protein folding, *Proc. Natl. Acad. Sci., USA* 84: 6611–6615.
35. Li, Z. and Scheraga, H.A. (1988), Structure and free energy of complex thermodynamic systems, *J. Molec. Str. (Theochem)* 179: 333–352.
36. Bharucha-Reid, A.T. (1960), *Elements of the theory of Markov processes and their applications*, McGraw-Hill, New York.
37. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), Equation of state calculations by fast computing machines, *J. Chemical Physics* 21: 1087–1092.
38. Hagler, A.T., Stern, P.S., Sharon, R., Becker, J.M. and Naider, F. (1979), Computer simulation of the conformational properties of oligopeptides. Comparison of theoretical methods and analysis of experimental results, *J. Am. Chem. Soc.* 101: 6842–6852.
39. Rapaport, D.C. and Scheraga, H.A. (1981), Evolution and stability of polypeptide chain conformation: a simulation study, *Macromolecules* 14: 1238–1246.
40. Paine, G.H. and Scheraga, H.A. (1985), Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. I. Backbone structure of enkephalin, *Biopolymers* 24: 1391–1436.
41. Gay, D.M. (1983), Algorithm 611. Subroutines for unconstrained minimization using a model/trust-region approach, *ACM Trans. Math. Software* 9: 503–524.
42. Ripoll, D.R. and Scheraga, H.A. (1988), On the multiple-minima problem in the conformational analysis of polypeptides. II. An electrostatically driven Monte Carlo method-tests on poly(L-alanine), *Biopolymers* 27: 1283–1303.
43. Ripoll, D.R. and Scheraga, H.A. (1989), The multiple-minima problem in the conformational analysis of polypeptides. III. An electrostatically driven Monte Carlo method; tests on enkephalin, *J. Protein Chem.* 8: 263–287.
44. Ripoll, D.R., Liwo, A. and Scheraga, H.A. (1998), New developments of the electrostatically driven Monte Carlo method – Test on the membrane bound portion of melittin, *Biopolymers* 46: 117–126.
45. Liwo, A., Tempczyk, A., Ołdziej, S., Shenderovich, M.D., Hruby, V.J., Talluri, S., Ciarkowski, J., Kasprzykowski, F., Łankiewicz, L. and Grzonka, Z. (1996), Exploration of the conformational space of oxytocin and arginine-vasopressin using the electrostatically-driven Monte Carlo and molecular dynamics methods, *Biopolymers* 38: 157–175.
46. Ripoll, D.R., Vásquez, M.J. and Scheraga, H.A. (1991), The electrostatically driven Monte Carlo method: Application to conformational analysis of decaglycine, *Biopolymers* 31: 319–330.
47. Ripoll, D.R. (1992), Conformational study of a peptide epitope shows large preferences for β-turn conformations, *Int. J. Pepide Protein Res.* 40: 575–581.
48. Faerman, C.H. and Ripoll, D.R. (1992), Conformational analysis of a twelve-residue analogue of mastoparan and mastoparan X, *Proteins* 12: 111–116.
49. Liwo, A., Gibson, K.D., Scheraga, H.A., Brandt-Rauf, P.W., Monaco, R. and Pincus, M.R. (1994), Comparison of the low energy conformations of an oncogenic and a non-oncogenic p21 protein, neither of which binds GTP or GDP, *J. Protein Chem.* 13: 237–251.
50. Ashkenazi, G., Ripoll, D.R., Lotan, N. and Scheraga, H.A. (1997), A molecular switch for biological logic gates: conformational studies, *Biosensors & Bioelectronics* 12: 85–95.

51. Ripoll, D.R., Vorobjev, Y.N., Liwo, A., Vila, J.A. and Scheraga, H.A. (1996), Coupling between folding and ionization equilibria. Effects of pH on the conformational preferences of polypeptides, *J. Mol. Biol.* 264: 770–783.
52. Vila, J.A., Ripoll, D.R., Villegas, M.E., Vorobjev, Y.N. and Scheraga, H.A. (1998), Role of hydrophobicity and solvent-mediated charge-charge interactions in stabilizing α-helices, *Biophys. J.* 75: 2637–2646.
53. Vila, J.A., Ripoll, D.R., Vorobjev, Y.N. and Scheraga, H.A. (1998), Computation of the structure-dependent pk_a shifts in a polypentapeptide of the Poly[f_v(IPGVG), f_e(IPGEG)] family, *J. Phys. Chem. B* 102: 3065–3067.
54. Olszewski, K.A., Piela, L. and Scheraga, H.A. (1992), Mean-field theory as a tool for intramolecular conformational optimization. 1. Tests on terminally-blocked alanine and Met-enkephalin, *J. Phys. Chem.* 96: 4672–4676.
55. Olszewski, K.A., Piela, L. and Scheraga, H.A. (1993), Mean field theory as a tool for intramolecular conformational optimization. 2. Tests on the homopolypeptides decaglycine and icosalanine, *J. Phys. Chem.* 97: 260–266.
56. Olszewski, K.A., Piela, L. and Scheraga, H.A. (1993), Mean field theory as a tool for intramolecular conformational optimization. 3. Test on melittin, *J. Phys. Chem.* 97: 267–270.
57. Piela, L., Kostrowicki, J. and Scheraga, H.A. (1989), The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method, *J. Phys. Chem.* 93: 3339–3346.
58. Pillardy, J., Olszewski, K.A. and Piela, L. (1992), Performance of the shift method of global minimization in searches for optimum structures of clusters of Lennard-Jones atoms, *J. Phys. Chem.* 96: 4337–4341.
59. Pillardy, J. and Piela, L. (1995), Molecular dynamics on deformed potential energy hypersurfaces, *J. Phys. Chem.* 99: 11805–11812.
60. Wawak, R.J., Gibson, K.D., Liwo, A. and Scheraga, H.A. (1996), Theoretical prediction of a crystal structure, *Proc. Natl. Acad. Sci., USA* 93: 1743–1746.
61. Wawak, R.J., Pillardy, J., Liwo, A., Gibson, K.D. and Scheraga, H.A. (1998), Diffusion equation and distance scaling methods of global optimization; Applications to crystal structure prediction, *J. Phys. Chem.* 102: 2904–2918.
62. Pillardy, J., Liwo, A., Groth, M. and Scheraga, H.A., An efficient deformation-based global optimization method for off-lattice polymer chains; self-consistent basin-to-deformed-basin mapping (SCBDBM). Application to united-residue polypeptide chains, *J. Phys. Chem.* B103: 7353–7366.
63. Pillardy, J., Liwo, A. and Scheraga, H.A. An efficient deformation-based global optimization method [self-consistent basin-to-deformed-basin mapping (SCBDBM)]; Application to Lennard-Jones atomic clusters, *J. Phys. Chem.*, in press.
64. Kostrowicki, J., Piela, L., Cherayil, B.J. and Scheraga, H.A. (1991), Performance of the diffusion equation method in searches for optimum structures of clusters of Lennard-Jones atoms, *J. Phys. Chem.* 95: 4113–4119.
65. Wawak, R.J., Wimmer, M.M. and Scheraga, H.A. (1992), Application of the diffusion equation method of global optimization to water clusters, *J. Phys. Chem.* 96: 5138–5145.
66. Kostrowicki, J. and Scheraga, H.A. (1992), Application of the diffusion equation method for global optimization to oligopeptides, *J. Phys. Chem.* 96: 7442–7449.
67. Pillardy, J., Olszewski, K.A. and Piela, L. (1992), Theoretically predicted lowest-energy structures of water clusters, *J. Mol. Struct.* 270: 277–285.
68. Lee, J., Scheraga, H.A. and Rackovsky, S. (1997), New optimization method for conformational energy calculations on polypeptides: Conformational space annealing, *J. Comput. Chem.* 18: 1222–1232.

69. Lee, J., Scheraga, H.A. and Rackovsky, S. (1998), Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing, *Biopolymers* 46: 103–115.
70. Lee, J. and Scheraga, H.A. (1999), Conformational space annealing by parallel computations: extensive conformational search of Met-enkephalin and of the 20-residue membrane-bound portion of melittin, *Int. J. Quant. Chem.* 75: 255–265.
71. Goldberg, D.E. (1989), *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley, Reading, MA.
72. Lee, J., Liwo, A. and Scheraga, H.A. (1999), Energy-based *de novo* protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K., *Proc. Natl. Acad. Sci. USA* 96: 2025–2030.
73. Ye, Y.-J. and Scheraga, H.A. (1999), Kinetics of protein folding, in "Slow Dynamics in Complex Systems: Eighth Tohwa University International Symposium", Eds. M. Tokuyama and I. Oppenheim. AIP Conference Proceedings 469, pp. 452–475, Amer. Inst. Phys.
74. Ye, Y.-J., Ripoll, D.R. and Scheraga, H.A. (1999), Kinetics of cooperative protein folding involving two separate conformational families, Computational and Theoretical Polymer Science, 9: 359–370.
75. Liwo, A., Lee, J., Ripoll, D.R., Pillardy, J. and Scheraga, H.A. (1999), Protein structure can be predicted by global optimization of a potential energy function, Proc. Natl. Acad. Sci., USA, 96: 5482–5485.
76. Gō, N. and Scheraga, H.A. (1970), Ring closure and local conformational deformations of chain molecules, *Macromolecules* 3: 178–187.
77. Palmer, K.A. and Scheraga, H.A. (1991), Standard-geometry chains fitted to X-ray derived structures; Validation of the rigid-geometry approximation. I. Chain closure through a limited search of 'loop' conformations, *J. Comput. Chem.* 12: 505–526.
78. Vila, J., Williams, R.L., Vásquez, M. and Scheraga, H.A. (1991), Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor, *Proteins: Struc., Func., and Gen.* 10: 199–218.
79. Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction;
http://predictioncenter.llnl.gov/casp3/Casp3.html, (1998), Asilomar Center, Pacific Grove, CA.
80. Lee, J., Liwo, A., Ripoll, D.R., Pillardy, J. and Scheraga, H.A. (1999), Calculation of protein conformation by global optimization of a potential energy function, *Proteins: Struc., Func., and Gen.*, suppl. 3: 204–208.
81. Yang, F., Gustafson, K.R., Boyd, M.R. and Wlodawer, A. (1998), Crystal structure of Escherichia coli HdeA, *Nature Struct. Biol.* 5: 763–764.