Conformation-Family Monte Carlo (CFMC): An Efficient Computational Method for Identifying the Low-Energy States of a Macromolecule

by **Jaroslaw Pillardy**, **Cezary Czaplewski**, **William J. Wedemeyer**, and **Harold A. Scheraga**[1])

Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, N.Y. 14853-1301, U.S.A

Dedicated to *Albert Eschenmoser* on the occasion of his 75th birthday

A highly efficient method, Conformation-Family Monte Carlo (CFMC), has been developed for searching the conformational space of a macromolecule and identifying its low-energy conformations. This method maintains a database of low-energy conformations that are clustered into families. The conformations in this database are improved iteratively by a Metropolis-type Monte Carlo procedure, together with energy minimization, in which the search is biased towards investigating the regions of the lowest-energy families. The CFMC method has the advantages of our earlier potential-smoothing methods (in that it 'coarse-grains' the conformational space and exploits information about nearby low-energy states), but avoids their disadvantages (such as the displacement of the global minimum at large smoothings). The CFMC method is applied to a test protein, domain B of Staphylococcal protein A. Independent CFMC runs yielded the same low-energy families of conformations from random starts, indicating that the thermodynamically relevant conformational space of this protein has been explored thoroughly. The CFMC method is highly efficient, performing as well as or better than competing methods, such as Monte Carlo with minimization, conformational-space annealing, and the self-consistent basin-to-deformed-basin method.

Introduction[2])**.** – Efficient methods for global energy optimization have many applications in computational chemical biology, such as X-ray and NMR structural refinements, the docking of ligands to macromolecules, and the *ab initio* prediction of protein and crystal structures. Three classes of such methods have been studied in our laboratory. The first class is that of Monte Carlo methods with minimization, such as the original MCM method [1][2] and the electrostatically driven Monte Carlo (EDMC) method [3]. The second class is that of deformation-based methods [4] such as the diffusion-equation method (DEM), the distance-scaling method (DSM), and their descendant as the self-consistent basin-to-deformed-basin method (SCBDBM) [5][6]. The third class is that of genetic (recombination) algorithms such as conformational space annealing (CSA) [7][8]. It should be noted that, despite their differences, these successful methods share certain properties; *e.g.*, *1*) structural modification steps are followed immediately by energy minimization, and *2*) previously explored states are used to guide the future exploration of the conformational space. In addition, the

[1]) Phone: (607) 255-4034; fax: (607) 254-4700; e-mail: has5@cornell.edu.

[2]) *Abbreviations:* CFMC: conformation-family Monte Carlo; MCM: Monte Carlo with minimization; EDMC: electrostatically driven Monte Carlo; SCBDBM: self-consistent basin-to-deformed-basin method; CSA: conformational-space annealing; SpA: Staphylococcal protein A; UNRES: united-residue model of proteins; SUMSL: secant unconstrained minimization solver; CTC: Cornell Theory Center; NIH: National Institutes of Health; ECEPP: empirical conformational-energy program for peptides; PDB: Protein Data Bank.

SCBDBM and CSA methods share a third property, namely, *3*) they both enforce a broad sampling of the conformational space, either by smoothing or by requiring that conformations maintain a minimal distance from one another.

A new method, Conformation-Family Monte Carlo (CFMC), is introduced here and tested on a small protein. The CFMC method can be considered as an extension of the original MCM method since, at each iteration of the method, a conformation is perturbed, locally minimized and then subjected to an accept/reject criterion. However, the CFMC method also is conceptually related to deformation-based methods, in that the CFMC conformations are clustered into families, which 'coarse-grain' the conformational space. As in deformation methods, this allows the CFMC method to exploit information about the local structure of the energy landscape to guide the global exploration. More generally, the CFMC method shares the properties *1*–*3* cited above.

This article is organized as follows. In the next section, we describe the CFMC method in detail. The following section describes the application of this method to the well-studied test protein, domain B of Staphylococcal protein A (SpA) [5][8][9]. CFMC performs much better than MCM and SCBDBM, and is roughly equal to CSA in efficiency. A detailed comparison of these methods is also given in this section.

Description of the CFMC Method. – The central element of the CFMC method is the *conformation-family database*, which is an ensemble of conformations clustered into families. To control the computational expense, the number of families and conformations within each family are bounded to N_f families and N_c conformations, respectively. This section describes in detail how this database is created and modified during a CFMC run. For the applications described in this article, we employed the united-residue (UNRES) model and energy function [10], which are described in the *Appendix*.

The conformation-family database for a CFMC run is initialized by successively generating N_f random conformations and minimizing them. (The local minimizations of this study were carried out using the SUMSL algorithm [11].) In the unlikely event that two randomly generated and minimized conformations have an rmsd less than a defined threshold R_c, the lower-energy conformation is kept in the database, and another random conformation is generated. The final N_f conformations and their families constitute the initial conformation-family database.

A newly generated conformation is said to be *connected* to a family of conformations, if there is at least one conformation in the family that differs from it in rmsd by less than a threshold R_f.

The CFMC Protocol (see flowchart in *Fig. 1*). In each iteration of a CFMC run, a new conformation C' is generated from a conformation C already present in the conformation-family database. This conformation C is chosen from a specific family F, denoted the *generative family*. As the CFMC simulation progresses, this generative family can change, as described below. For the first CFMC iteration, the generative family is set to the lowest-energy family in the initial conformation-family database.

An iteration of CFMC consists of four steps, as follows:

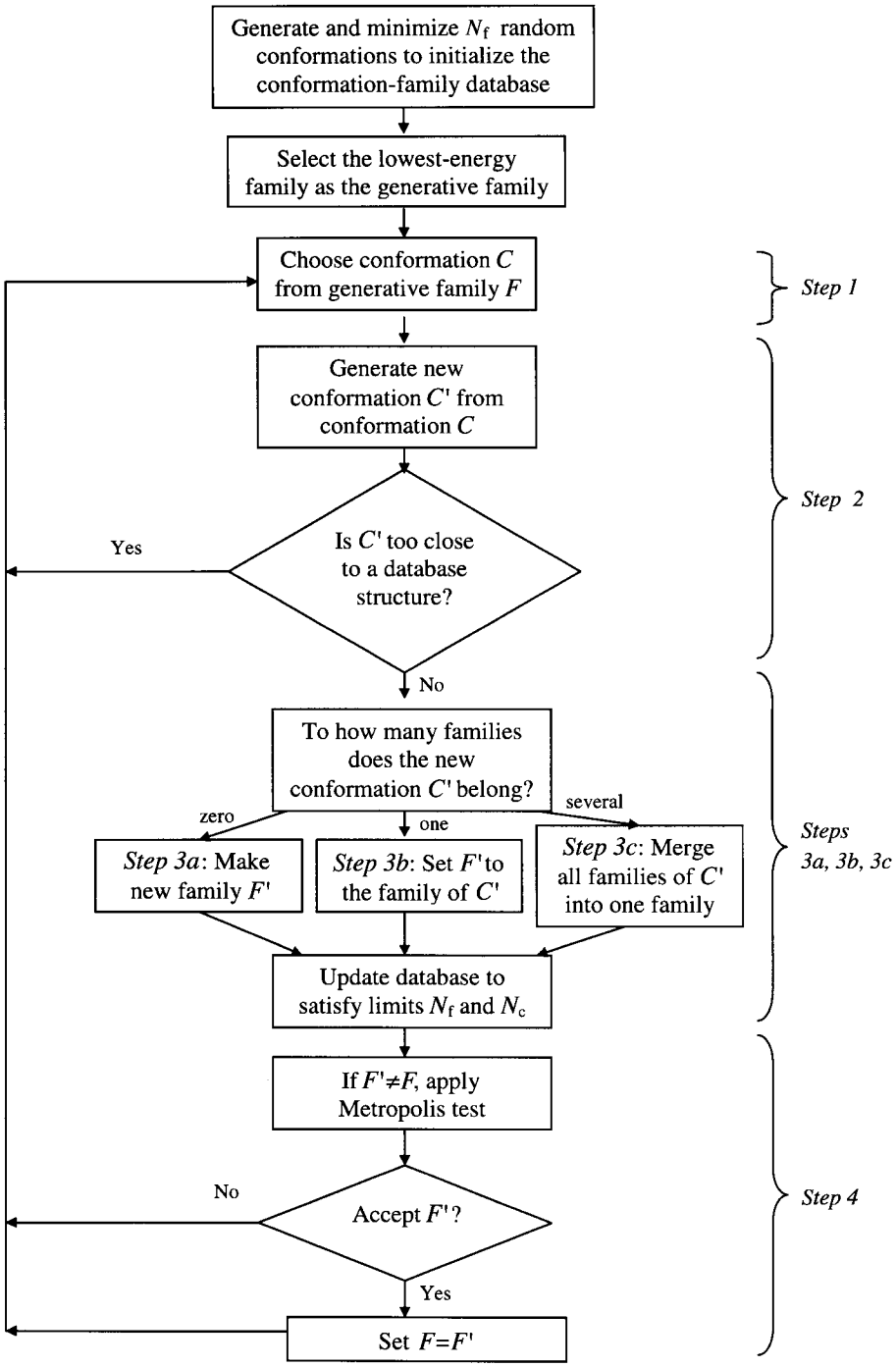

Fig. 1. *Flowchart for the CFMC method*

Step 1. A conformation C is chosen from the generative family F, with a probability proportional to its *Boltzmann* weight.

Step 2. This conformation C is modified to yield a new conformation C'. (The methods for modifying conformations are described below.) If the new conformation C' is closer than R_c in rmsd to another conformation C'' in the database, the lower-energy conformation of the pair is stored in the database and the algorithm returns to *Step 1*.

Step 3a. If the new conformation C' is unconnected to any family in the database, a new family F' is created whose sole member is C'. (Of course, conformations may be added to this new family in subsequent CFMC iterations, as in *Step 3b*.) If the number of families in the database exceeds the limit N_f, the family with the highest energy is eliminated. The algorithm then jumps to *Step 4*.

Step 3b. If the new conformation C' is connected to exactly one family F' in the database, the conformation is added to the family. If the number of conformations in the family exceeds the limit N_c, the conformation with the highest energy is eliminated. This elimination may split the original family into two families; if so, the database is updated accordingly. The algorithm then jumps to *Step 4*.

Step 3c. If the new conformation C' is connected to more than one family in the database, these families are merged into a single family F'. If the number of conformations in the merged family exceeds the limit N_c, the conformations with the highest energy should be eliminated. However, this may also eliminate the conformations that caused the families to be merged in the first place, leading to a kind of oscillation in which families are repeatedly merged and re-divided immediately. Hence, two families are merged if and only if the new conformation C' meets two criteria designed to avoid such oscillations. First, C' must have a lower energy than every conformation of the higher-energy family and, second, C' must have a lower energy than at least one conformation of the lower-energy family. If these two criteria are not met, the conformation C' is rejected and the algorithm returns to *Step 1*; otherwise, the algorithm proceeds to *Step 4*.

Step 4. If the new family F' found in *Step 3* is not identical to the original generative family F of *Step 1*, a Metropolis criterion is applied to determine whether to make F' the generative family. F' becomes the new generative family if it has a lower energy than F, or if its *Boltzmann* factor $exp\ (-\beta\Delta E)$ is greater than a randomly generated number in the interval (0, 1), where $\beta \equiv 1/kT$ as usual, and ΔE is the energy difference between families F and F'. (The energy of a family is defined as that of its lowest-energy conformation.) If this Metropolis criterion is not met, F remains the generative family. At this point, the algorithm returns to *Step 1* and a new CFMC iteration begins.

The Methods for Producing New Conformations. In the second step of a CFMC iteration, the conformation C is modified to yield a new conformation C'. This modification is carried out in several ways, which are described in this subsection.

There are two general classes of moves used in the CFMC method: *internal* (or local) moves, intended to improve the low-energy conformations within a family, and *external* (global) moves, intended to search the conformational space for new families. Within each class, there are two kinds of moves: perturbation and averaging (see *Appendix* for definitions of the variables in the UNRES model).

Perturbations are used for searching the space of the backbone torsional angles γ, as well as that of the side-chain positions. Moves perturbing only one backbone torsional angle (or only one side-chain position) are distinguished from moves perturbing a sequence of consecutive backbone torsional angles (or side chains); for the latter, the number of variables to be perturbed is chosen randomly between 1 and the maximum, pre-defined, number. Each perturbation is carried out in one of two ways: *a*) an angle is changed randomly within a pre-defined range, or *b*) an angle is chosen according to the distribution of angles already present in the database; both ways are used with equal probability. The only difference between internal and external moves of type *a*) is that the pre-defined range for external moves is larger than that for internal moves. The difference between internal and external moves of type *b*) lies in the distribution of angles used. For internal moves, the distribution of the particular variable being perturbed is computed from the *current* conformation-family database; for example, if variable 5 is being perturbed, the distribution of variable 5 in the present conformation-family database is used. For external moves, where the larger variation of a variable is desired, the distribution of the perturbed variable is calculated from the *initial* conformation-family database (obtained by minimizing randomly generated conformations, as noted above). Altogether, eight kinds of perturbation moves are used: two single backbone torsional-angle moves (external and internal), two multiple consecutive backbone-torsional-angle moves (external and internal), two single side-chain moves (external and internal, both consisting of perturbations of the α_{SC} and β_{SC} angles), and two multiple consecutive side-chain moves.

An entirely different kind of move is averaging, for which two different conformations are necessary. For external averaging, these conformations are chosen from different families while, for internal averaging, they are chosen from the same family. From these two conformations, an averaged (or interpolated) conformation is then calculated using a randomly chosen 'mixing ratio' x (in the range from 0 to 1). Thus, every variable of the averaged conformation (*i.e.*, α_{SC_i}, β_{SC_i}, γ_i, and θ_i) is calculated according to *Eqn. 1:*

$$\nu_i^* = \nu_i^{(1)} \cdot x + \nu_i^{(2)} \cdot (1 - x) \tag{1}$$

where x is the 'mixing ratio', and ν is a variable. This type of search focuses mostly on differences between conformations; the averaged conformation would indeed differ very little from the initial two conformations in regions where their secondary structures are similar.

Comparison with Related Optimization Methods. The CFMC method is unusual as a Monte Carlo method in that an ensemble of states is simulated, rather than the single-state characteristic of Metropolis Monte Carlo simulations. In this regard, the CFMC method resembles other ensemble-oriented simulation methods such as the CSA and SCBDBM methods. These three ensemble-oriented methods share other similarities as well. For example, they all maintain a database of conformations, which is initialized to a set of randomly generated (and minimized) conformations; this database is then gradually 'pruned' into shape by successive random moves and accept/reject criteria. In all three methods, the members of this structural database are required to remain well-

separated in order to ensure a broad sampling of the conformational space; as the simulation progresses, this separation requirement is gradually relaxed, allowing the ensemble to converge to the appropriate thermal distribution.

However, the CFMC method also has significant differences from these other methods. In particular, the families of the CFMC method constitute an additional level of organization in the database of conformations; in effect, CFMC moves are made not between conformations, but between families of conformations. Unlike the SCBDBM method, the true *undeformed* energy is evaluated in the CFMC method; the 'smoothing' of the CFMC method occurs by clustering the accepted conformations into families.

The move-regeneration scheme of the CFMC method differs significantly from the CSA and SCBDBM methods. As detailed above, the CFMC relies heavily on perturbations of one or a few adjacent dihedral angles, whereas the CSA method employs a recombination scheme in which pieces of candidate conformations are combined and minimized. Two CFMC moves are distantly related to CSA moves. The first such CFMC move is that in which a single dihedral angle is chosen from a known statistical distribution for that angle; the analogous move in CSA occurs when a single dihedral angle is exchanged. However, this analogy does not extend to moves involving multiple adjacent dihedral angles; in CSA, the values of these angles are exchanged as a block, whereas, in the CFMC method, these angles are chosen independently. The second such CFMC move is that in which two conformations are averaged; this corresponds loosely to the various recombinations that occur in the CSA method. Despite their superficial resemblance, these moves are quite different; the CSA method combines separate pieces of the two conformations, whereas the CFMC method interpolates their conformational variables evenly.

Computational Details. All calculations were carried out on a *PC-Linux* cluster in our laboratory, and on the *PC-Windows NT* cluster in the Cornell Theory Center. Numerical parameters for the CFMC method were chosen as follows. The inter-family cutoff (R_f) varied between 5 Å (initially) and 2 Å (at the end of a simulation); the intra-family cutoff (R_c) varied between 2 Å (initially) and 1 Å (at the end of a simulation); both cutoffs were reduced linearly every 2,500 local minimizations. The maximum number N_f of families was set at 20, and the maximum number N_c of conformations in any family was set at 50. The total number of local minimizations carried out during a full CFMC run was limited to 50,000.

The program has been parallelized at the coarse-grain level, *i.e.*, local minimizations are carried out in parallel on different processors. A single CFMC run is divided into several threads (usually five); all of them use the same database of conformations. The master processor carries out all the database management, as well as generation of trial conformations, and acceptance decisions. Usually 26 processors were used per run, and a single simulation (50,000 local minimizations) of our model system (domain B of Staphylococcal protein A) took about 3.5 wall-clock hours on our *PC-Linux* cluster.

Results and Discussion. – The CFMC method has been applied to a benchmark protein, a truncated version (residues 10–55) of the B domain of Staphylococcal protein A. Several high-resolution NMR structures of the untruncated domain and its

homologues have been published [12], indicating that it folds into a classic up-down, three-helix bundle with nearly parallel helical axes. Global energy minimization of the truncated protein has been carried out in several laboratories, using very different energy functions and minimization protocols [5][8][9]. These computational studies agree that the global energy minimum is a three-helix bundle in which the first two helices are aligned, but the third (C-terminal) helix crosses at a significant angle. These studies also agree in identifying a 'mirror-image' fold as having the next lowest energy after the global minimum.

To assess the effectiveness of the CFMC method in finding the global minimum and in exploring the conformational space of protein A, comparative simulations were carried out using the classical MCM method [1] and the CSA method [7]. An equal amount of computational resources (as measured by the number of local minimizations) was used in all simulations.

A total of 10 simulations using CFMC were carried out starting from the randomly generated conformations, as described in the previous section. In 6 of them, the global minimum, with energy −127.16 kcal/mol was located. In the remaining 4 simulations, another conformation, slightly higher in energy than the previous one (−126.86 kcal/mol), was found. Both conformations belong to the native-fold family, and their rmsd difference in the C^{α} positions is 1.5 Å. As shown in *Fig. 2*, the only structural difference between these conformations occurs in the five C-terminal residues. In the global minimum conformation, this pentapeptide has no regular secondary structure, whereas, in the competing low-energy (−126.86 kcal/mol) conformation, this pentapeptide forms a small helical fragment positioned at a 60° angle to the third helix.

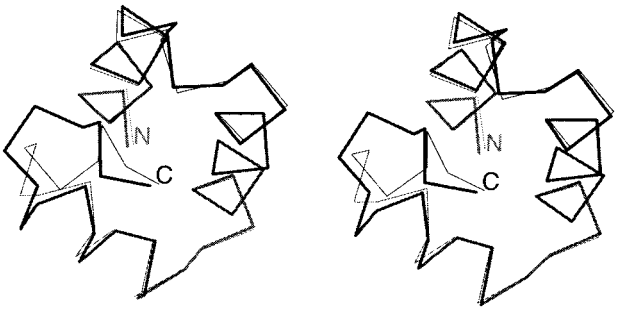

Fig. 2. *Stereoviews of the two conformations found by the CFMC method as the lowest-energy conformations in different runs.* The global minimum is shown in thick lines (−127.16 kcal/mol), and the second minimum is shown in thin lines (−126.86 kcal/mol). Both structures belong to the same family.

The three comparative simulations using the CSA method found conformations very close to the global minimum; one found the global minimum itself, while the other two found conformations differing from the global minimum in the orientations of a few side chains. The results of the ten CFMC simulations (solid lines) and the three CSA simulations (dashed lines) are compared in *Fig. 3*. By contrast, the classical MCM was significantly less successful, as shown in *Fig. 4*. Only two out of the ten MCM simulations located the global minimum, and half of the runs were stuck in high-energy

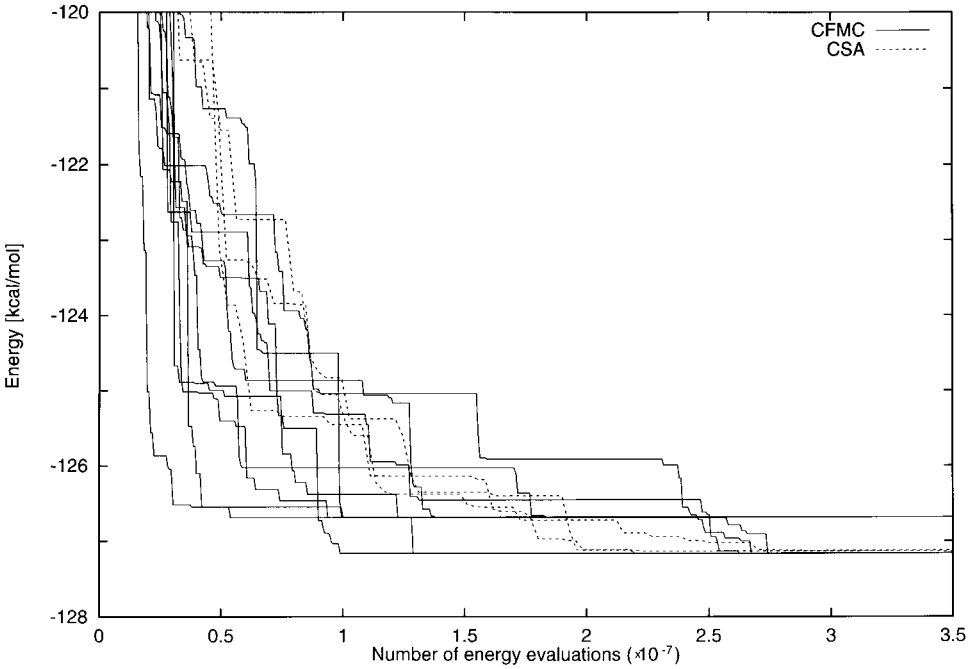

Fig. 3. *Comparison between the CFMC and CSA methods.* The abscissa shows the progress of the simulations (represented by the number of energy evaluations), and the ordinate plots the energy of the lowest-energy conformation obtained for a given number of energy evaluations during the simulations. The ten CFMC simulations are shown with solid lines, and the three CSA simulations are shown with dashed lines. As noted in the main text, the low-energy conformations obtained from the CSA and CFMC methods differed only in two side-chain positions.

conformations. These conformations did not adopt the native fold and their energies were 2–7 kcal/mol higher than the global minimum energy.

A great advantage of the CFMC method is that it not only locates the lowest-energy family of conformations, but also generates a broad distribution of conformations covering a large region of the potential-energy hypersurface. Such a distribution may be analyzed by plotting rmsd from the global minimum conformation *vs.* energy for the whole set of conformations. The plot of a typical distribution of conformations generated by the CFMC method during one complete simulation is shown in *Fig. 5*, *a*, in which each conformation is represented by a single dot. The pathway to the global minimum (global-minimum trajectory) is plotted as a solid line connecting the consecutive, lowest-energy conformations obtained during the run. The content of the database at the end of the simulation is shown in *Fig. 5*, *b*; the number of conformations therein is significantly lower than in *Fig. 5*, *a*, because high-energy conformations (and families) are gradually removed from the database during the simulation.

At the beginning of a simulation, changes in the topology of the lowest-energy conformation are large, as indicated by the large variations in the global minimum trajectory in *Fig. 5*. The lack of significant focusing in particular parts of the

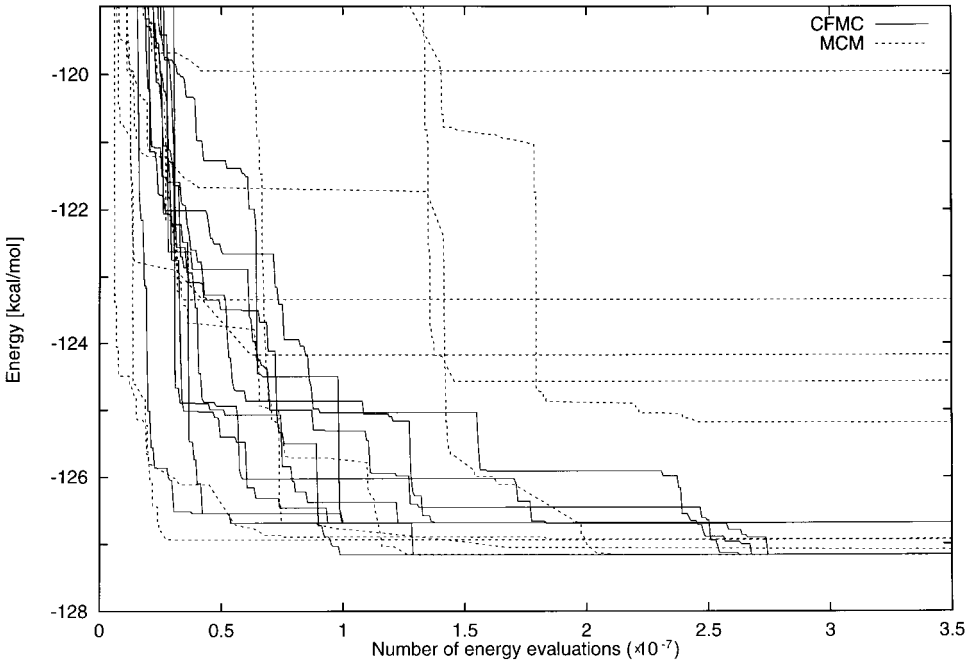

Fig. 4. *Comparison between the CFMC and classical MCM methods.* The abscissa shows the progress of the simulations (represented by the number of energy calculations), and the ordinate plots the energy of the lowest-energy conformation obtained for the given number of energy evaluations during the simulations. The ten CFMC simulations are shown with solid lines, while the ten MCM simulations are shown with dashed lines.

conformational space at the beginning of a simulation is also shown by the uniform distribution of the conformations (no clustering) on the right side of *Fig. 5,a*. As the algorithm progresses, the fluctuations decrease, and the method focuses on the important, low-energy families of conformations. All low-energy families could be classified into two distinct folds: the native fold (lower part of *Fig. 5,a* and *b*) and its mirror image (higher part of *Fig. 5,a* and *b*). Both classes of families are well-sampled during the simulation.

A detailed analysis of the families present in the database at the end of a typical simulation is shown in *Fig. 6,a–c*. There are 20 different families present in the database, 4 of them adopting the mirror-image fold, and the other 16 adopting the native fold. The topological differences between families are shown in three plots: the plot of rmsd from the lowest-energy mirror-image conformation *vs.* rmsd from the global-minimum conformation (*Fig. 6,a*), the plot of energy *vs.* rmsd from the global-minimum conformation (*Fig. 6,b*), and the plot of energy *vs.* rmsd from the lowest-energy mirror-image conformation (*Fig. 6,c*). The conformations belonging to different families are represented by different geometrical symbols (circles, triangles, *etc.*). *Fig. 6,a*, clearly shows that the two folds (native and mirror-image) are geometrically well separated from one another, because their families do not overlap. *Fig. 6,b* and *c*, show that, for each fold, there is a positive correlation between the energy and the rmsd distance from the lowest-energy conformation adopting that fold.

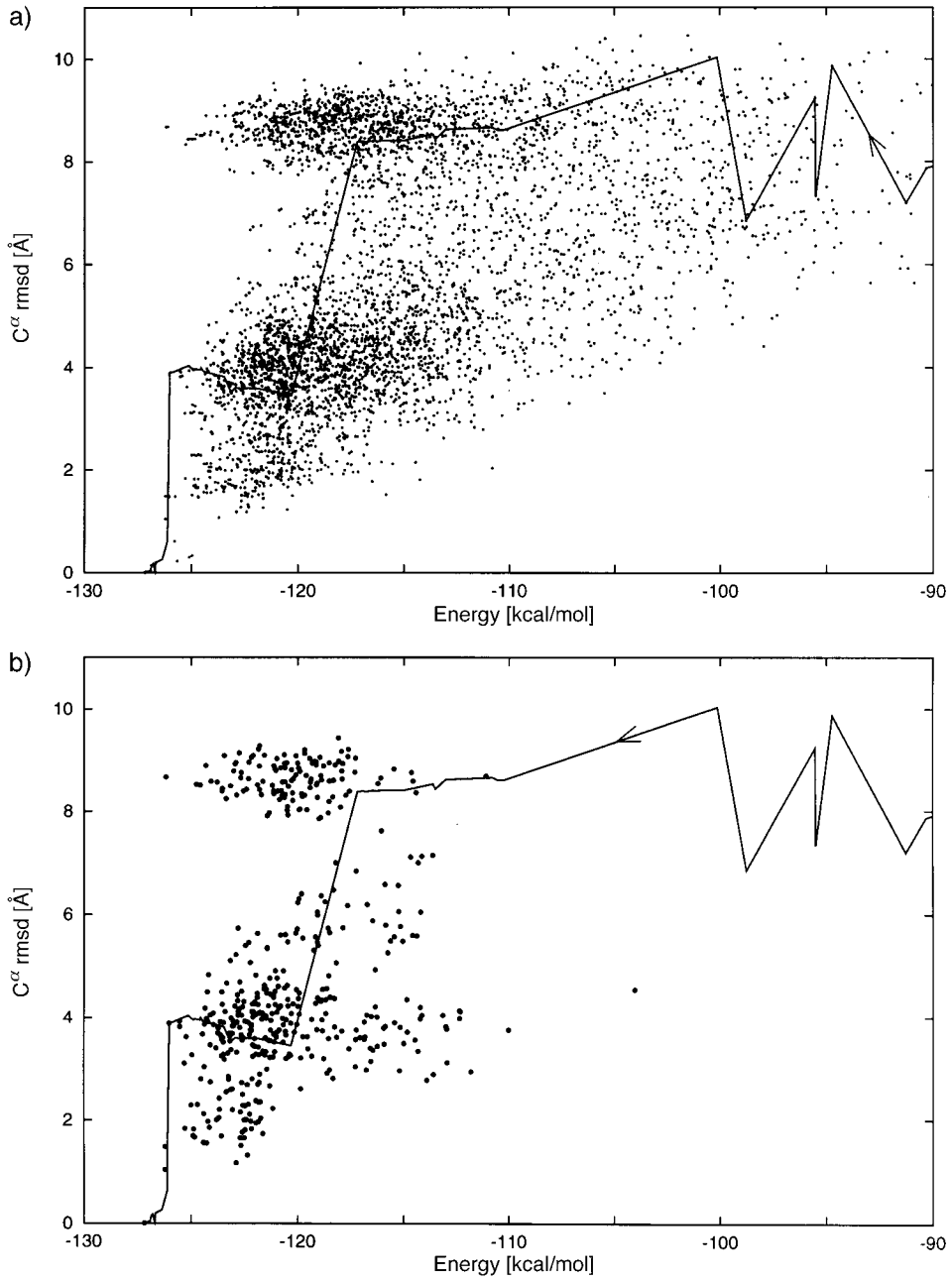

Fig. 5. a) *Plot of C^α rmsd from the lowest-energy conformation* vs. *energy for all conformations obtained during the simulation;* b) *plot of C^α rmsd from the lowest-energy conformation* vs. *energy for all conformations present in the database at the end of the simulation*. The pathway to the global minimum (global-minimum trajectory) is plotted as a solid line connecting the consecutive, lowest-energy structures obtained during the run in chronological order.

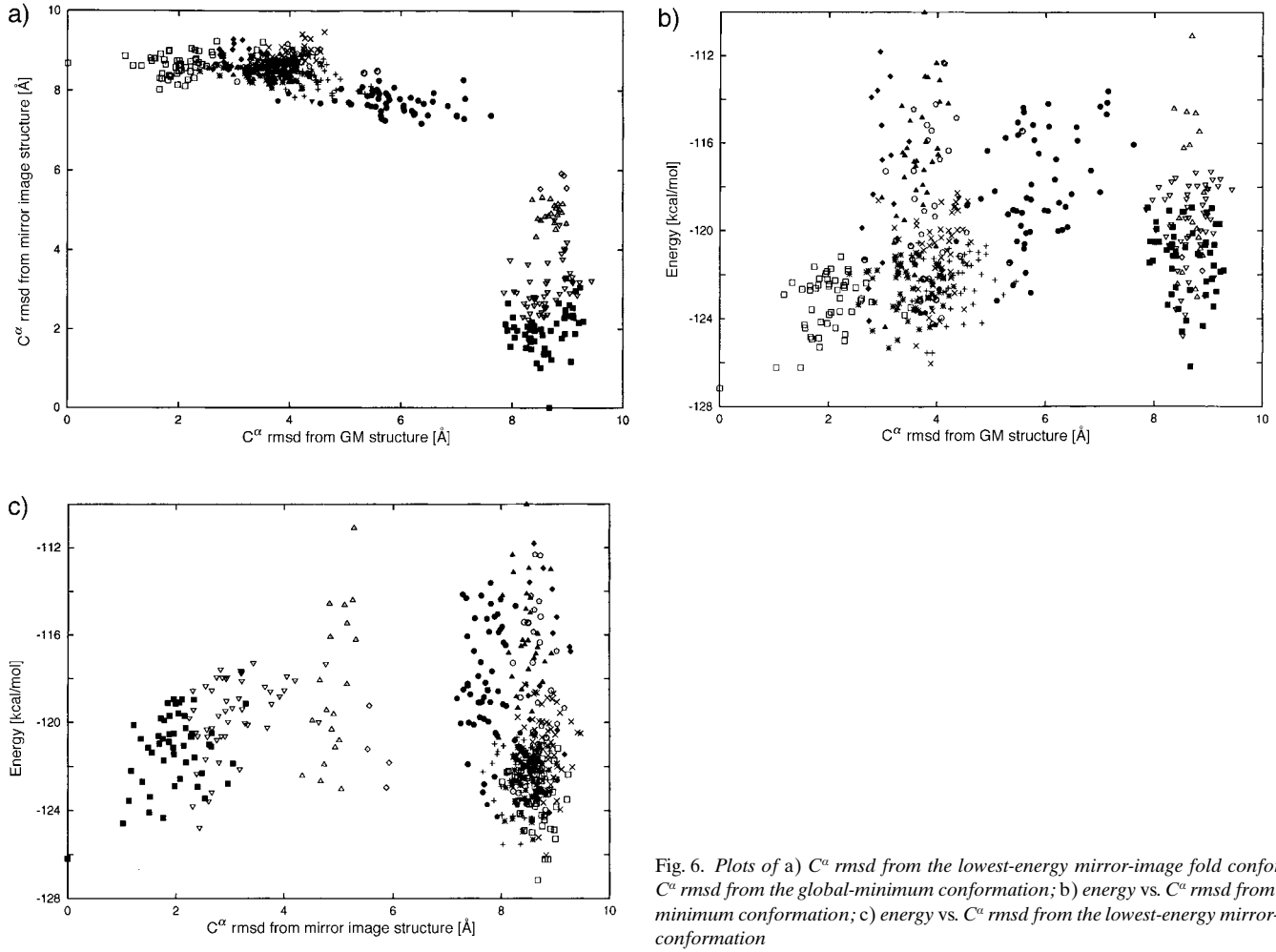

Fig. 6. *Plots of* a) *C^α rmsd from the lowest-energy mirror-image fold conformation* vs. *C^α rmsd from the global-minimum conformation;* b) *energy* vs. *C^α rmsd from the global-minimum conformation;* c) *energy* vs. *C^α rmsd from the lowest-energy mirror-image fold conformation*

The lowest-energy conformations of the four lowest-energy families adopting the native-like fold are shown in *Fig. 7*, while the lowest-energy conformations from the two lowest-energy families of the mirror-image fold are shown in *Fig. 8*. The main difference between low-energy, native-like families is the structure of the C-terminus. In the global minimum conformation (*Fig. 7,a*), the C-terminal residues are packed parallel to the end of the N-terminal helix. By contrast, the other three families have C-termini rotated by 90° relative to the global-minimum conformation. In the second family, the C-terminus points towards the middle of the N-terminal helix (*Fig. 7,b*); in the third family, it is packed antiparallel to the N-terminal helix (*Fig. 7,c*) while, in the fourth family, the C-terminal helix extends continuously to include the C-terminal residues (*Fig. 7,d*). The two lowest-energy families of the mirror-image fold exhibit small differences in the loop regions and at the C-terminus (*Fig. 8,a* and *b*).

We also discovered a conformation that may reflect an intermediate fold between the global-minimum and the mirror-image folds (*Fig. 9*). The global-minimum conformation can be described by its inter-helical angles $\Omega_{12} = 175°$, $\Omega_{23} = 133°$, and $\Omega_{13} = 50°$. For comparison, the mirror-image fold has inter-helical angles $\Omega_{12} = 108°$, $\Omega_{23} = 124°$, and $\Omega_{13} = 17°$. Thus, in the global-minimum fold, the N-terminal and the middle helices are aligned and the C-terminal helix crosses over them whereas, in the 'mirror-image' fold, the N-terminal and C-terminal helices are aligned, and the middle helix crosses over them. By contrast, in the intermediate conformation, the three helices are nearly perpendicular to each other, with inter-helical angles $\Omega_{12} = 110°$, $\Omega_{23} = 99°$, and $\Omega_{13} = 89°$. Thus, there seems to be a low-energy pathway connecting the global-minimum and 'mirror-image' folds, in which the N-terminal helix moves between the middle and C-terminal helices (*Fig. 9*).

The results of the simulations for protein A clearly show the advantage of the family-based approach to MCM over the classical implementation of the MCM algorithm. The CFMC method not only correctly located low-energy structures in all the runs carried out, but also generated a broad distribution of structures in the conformational space. The effectiveness of the method is comparable to that of the CSA method; in all CFMC simulations, the family containing the global-minimum structure was located correctly.

The families of low-energy conformations generated by the CFMC method may be applicable in simulating the folding of proteins. If such families can be identified with the kinetic species of the folding protein, and if the transition rates between such families can be determined, then a relatively simple master equation can be solved to reproduce the kinetics of folding [13].

This research was supported by grants from the *National Science Foundation* (MCB95-13167), and from the *National Institutes of Health* (GM-14312). Some of the computations were carried out at the *Cornell Theory Center* (CTC), which is funded in part by the *NIH National Center for Research Resources*, and the *CTC*'s *Advanced Cluster Computing Consortium*.

Appendix: The UNRES Protein Representation and Potential. – Recently, a novel united-residue model of proteins (UNRES) has been proposed [10]. This model has been shown to be computationally efficient and remarkably accurate in predicting native protein conformations [8][14]. This section reviews the properties of this model.

The Representation of Proteins in UNRES. The polypeptide backbone of a protein is represented in UNRES by its C^{α} trace. By assumption, all virtual bond lengths are 3.8 Å, *i.e.*, *cis*-peptide bonds are not

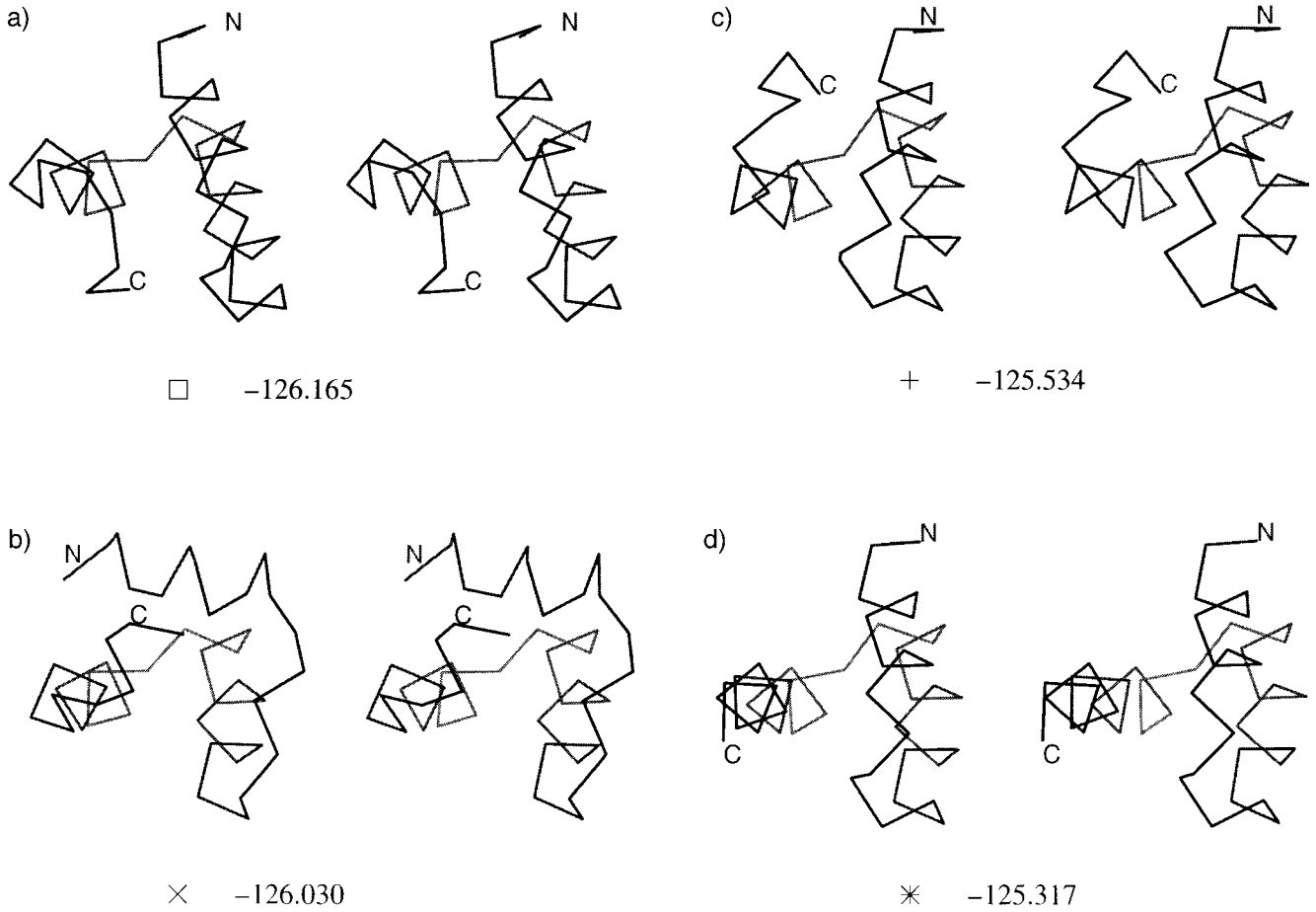

Fig. 7. *Stereoviews of the lowest-energy conformations of the four lowest-energy families adopting the native fold*. The symbols below each stereofigure correspond to those used in *Fig. 6*. For example, the square symbol under *Fig. 7,a*, corresponds to the lowest-energy (global minimum) family of *Fig. 6*.

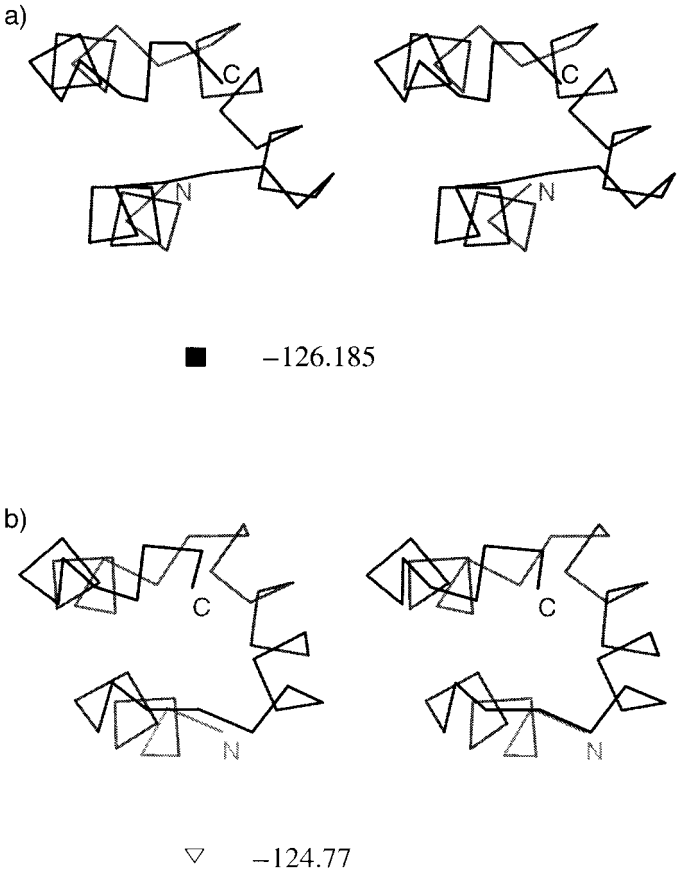

Fig. 8. *Stereoviews of the lowest-energy conformations of the two lowest-energy families adopting the mirror-image fold*. As in *Fig. 7*, the symbols below each stereofigure correspond to those plotted in *Fig. 6*.

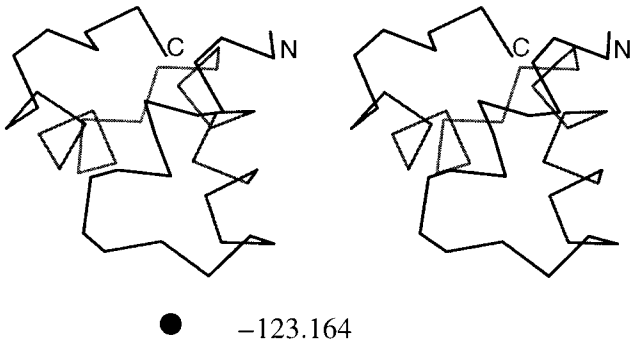

Fig. 9. *Stereoview of a low-energy conformation that adopts an intermediate fold between the global-minimum and mirror-image folds.* The filled circle below the stereofigure corresponds to that used for the same family of conformations in *Fig. 6*.

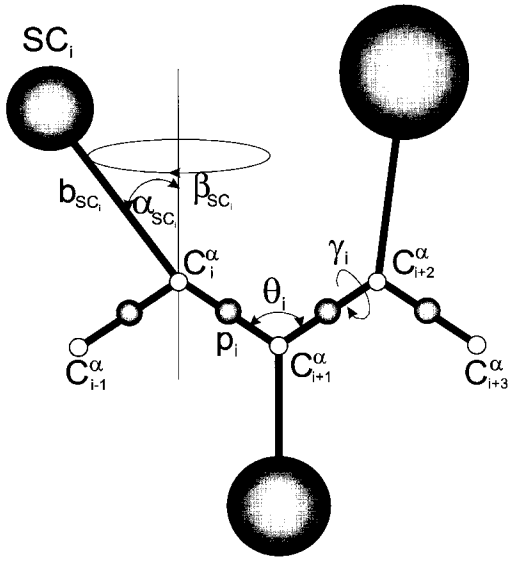

Fig. 10. *United-residue representation of a polypeptide chain*. The interaction sites are side-chain centroids of different sizes (SC) and peptide-bond centers (p) indicated by solid circles, while the C^α-atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual $C^\alpha - C^\alpha$ 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 C^α-atom with a fixed 'bond length', b_{SC}, that varies between residue types; likewise, the radius of side-chain sphere is fixed but varies between residue types. The variable side-chain 'bond angle', α_{SC}, is defined by the vector SC_i and the bisector of the angle defined by C^α_{i-1}, C^α_i, and C^α_{i+1}. Similarly, the variable side-chain 'dihedral angle', β_{SC}, is defined as a counterclockwise rotation about the bisector, starting from the C^α_{i+1} side of the C^α_{i-1}, C^α_i, C^α_{i+1} frame.

considered. Therefore, the only conformational variables of the backbone are θ (the angle between successive $C^\alpha - C^\alpha$ virtual bonds) and γ (the dihedral angle between four successive C^α-atoms), as indicated in *Fig. 10*. The peptide bond itself is modeled as a point dipole positioned in the middle of each $C^\alpha - C^\alpha$ virtual bond.

The protein side chains are represented in UNRES by simple spheres of different sizes for each type of residue. The center of this side-chain sphere is separated from its C^α-atom by a constant distance b_{SC} that varies between residue types. Consequently, the only conformational variables for the side chain are α_{SC} (the conical angle formed by the side chain) and β_{SC} (its dihedral angle), as indicated in *Fig. 10*.

The UNRES Energy Function. Each residue in the UNRES model has only two centers of interaction: the peptide-bond dipoles and the side-chain spheres. In particular, the C^α-atoms serve to define the chain geometry, but do not contribute to the energy evaluations. This drastic reduction in the number of interaction centers (when compared to all-atom models of proteins) renders UNRES much more efficient computationally. This efficiency is essential for a thorough exploration of the thermodynamically relevant space.

The UNRES energy function is given by *Eqn. 2*.

$$U = \sum_{i<j} U_{SC_iSC_j} + \sum_{i\neq j} U_{SC_ip_j} + w_{el} \sum_{i<j} U_{p_ip_j} + w_{tor} \sum_i U_{tor}(\gamma_i) + w_{loc} \sum_i [U_b(\theta_i) + U_{rot}(\alpha_i,\beta_i)] + w_{corr} U_{corr} \qquad (2)$$

$U_{SC_iSC_j}$ represents the interaction energy between the side chains, which contains their solvent interactions implicitly. $U_{SC_ip_j}$ represents the excluded-volume interactions that prevent the overlap of the peptide bonds with the side chains. $U_{p_ip_j}$ represents the energy of electrostatic interactions between peptide groups, including their tendency to form backbone H-bonds. U_{tor} represents the intrinsic energy of rotation about the virtual $C^\alpha - C^\alpha$ bonds, which biases the backbone dihedral angles in order to simulate the intrinsic secondary structure propensities of the protein. U_b represents the bending energy of the virtual-bond valence angles (which also contributes to secondary structure formation), while U_{rot} represents the energy of different rotameric states of

the side chains. Finally, U_{corr} is a multibody correlation energy, which arises from the fact that some degrees of freedom are eliminated when moving to the reduced UNRES representation from the all-atom representation. The w coefficients are the weights of the respective energy terms.

The terms $U_{SC_iSC_j}$, U_{tor}, U_b, and U_{rot} were parameterized originally from a set of 195 high-resolution non-homologous structures from the PDB [10]. In this parameterization, the distribution $\rho(\mathbf{X})$ of the conformational variables $\mathbf{X}$ obtained from the PDB is, by assumption, related to the corresponding restricted free energy $u(\mathbf{X})$ by the equation:

$$\rho(\mathbf{X}) = \rho_0(\mathbf{X}) \exp[-\beta u(\mathbf{X})] \tag{3}$$

where $\rho_0(\mathbf{X})$ is the reference distribution function in the absence of any interactions.

The parameters of the excluded-volume potential $U_{SC_iP_j}$ were chosen to reproduce the typical backbone geometries of the polypeptide chain in short model helices and sheets, *e.g.*, to give the screw sense of the helix characteristic of α-helices and not of 3_{10}- or π-helices. The expression for $U_{P_iP_j}$ was derived by averaging the energy of electrostatic interactions of the peptide groups and parametrized by fitting the average restricted free-energy surface corresponding to the interaction of two peptide groups, calculated by using the ECEPP/2 force field [15]. The current UNRES force field also includes multibody terms, U_{corr}, that arise from correlated interactions of two neighboring pairs of peptide groups. The weights of the energy terms were determined by optimization of the Z-score function of the phosphocarrier protein from *Streptococcus faecalis* (an 89-residue α/β protein).

The U_{tor} term has since been re-parametrized using the ECEPP/3 potential energy [16]. This approach is preferable, since the parametrization relies on a well-tested potential function based on semi-empirical and model-compound studies, rather than on a statistical survey of the PDB. This newly re-parametrized U_{tor} potential was used in these investigations with its weight w_{tor} set equal to 1.

REFERENCES

[1] Z. Li, H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611; Z. Li, H. A. Scheraga, *J. Mol. Struct. (THEOCHEM)* **1988**, *179*, 333.

[2] D. J. Wales, J. P. K. Doye, *J. Phys. Chem. A* **1997**, *101*, 5111; D. J. Wales, H. A. Scheraga, *Science* **1999**, *285*, 1368.

[3] D. R. Ripoll, H. A. Scheraga, *Biopolymers* **1988**, *27*, 1283; D. R. Ripoll, H. A. Scheraga, *J. Protein Chem.* **1989**, *8*, 263; D. R. Ripoll, A. Liwo, H. A. Scheraga, *Biopolymers* **1998**, *46*, 117.

[4] J. Kostrowicki, L. Piela, H. A. Scheraga, *J. Phys. Chem.* **1989**, *93*, 3339; J. Kostrowicki, L. Piela, B. J. Cherayil, H. A. Scheraga, *J. Phys. Chem.* **1991**, *95*, 4113; J. Pillardy, K. A. Olszewski, L. Piela, *J. Phys. Chem.* **1992**, *96*, 4337; P. Amara, D. Hsu, J. E. Straub, *J. Phys. Chem.* **1993**, *97*, 6715; J. P. Ma, D. Hsu, J. E. Straub, *J. Chem. Phys.* **1993**, *99*, 4024; J. Pillardy, L. Piela, *J. Phys. Chem.* **1995**, *99*, 11805; R. V. Pappu, R. K. Hart, J. W. Ponder, *J. Phys. Chem. B* **1998**, *102*, 9725; I. Andricioaei, J. E. Straub, *J. Comput. Chem.* **1998**, *19*, 1445.

[5] J. Pillardy, A. Liwo, M. Groth, H. A. Scheraga, *J. Phys. Chem. B* **1999**, *103*, 7353.

[6] J. Pillardy, A. Liwo, H. A. Scheraga, *J. Phys. Chem. A* **1999**, *103*, 9370; J. Pillardy, R. J. Wawak, Y. A. Arnautova, C. Czaplewski, H. A. Scheraga, *J. Am. Chem. Soc.* **2000**, *122*, 907.

[7] J. Lee, H. A. Scheraga, S. Rackovsky, *J. Comput. Chem.* **1997**, *18*, 1222.

[8] J. Lee, A. Liwo, H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025.

[9] A. Kolinski, J. Skolnick, *Proteins: Struct. Funct. Genet.* **1994**, *18*, 353; Z. Guo, C. L. Brooks III, E. Boczko, *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10161; D. O. V. Alonso, V. Dagett, *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 133.

[10] A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, H. A. Scheraga, *J. Comput. Chem.* **1997**, *18*, 849; A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, H. A. Scheraga, *J. Comput. Chem.* **1997**, *18*, 874; A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, H. A. Scheraga, *J. Comput. Chem.* **1998**, *19*, 259; A. Liwo, J. Pillardy, C. Czaplewski, J. Lee, D. R. Ripoll, M. Groth, S. Rodziewicz-Motowidlo, R. Kazmierkiewicz, R. J. Wawak, S. Oldziej, H. A. Scheraga, 'Proceedings of Fourth Annual International Conference of Computational Molecular Biology (RECOMB 2000)', Tokyo, Japan, 8–11 April 2000.

[11] D. M. Gay, *ACM Trans. Math. Software* **1983**, *9*, 503.

[12] H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, I. Schimada, *Biochemistry* **1992**, *40*, 9665; M. A. Starovasnik, N. J. Skelton, M. P. O'Connell, R. F. Kelley, D. Reilly, W. J. Fairbrother, *Biochemistry* **1996**, *35*,

15558; M. T. Tashiro, R. Tejero, D. E. Zimmerman, B. Celda, B. Nilsson, G. T. Montelione, *J. Mol. Biol.* **1997**, *272*, 573.

[13] R. A. Sendak, D. M. Rothwarf, W. J. Wedemeyer, W. A. Houry, H. A. Scheraga, *Biochemistry* **1996**, *35*, 12978; Y.-J. Ye, D. R. Ripoll, H. A. Scheraga, *Comput. Theor. Polymer Sci.* **1999**, *9*, 359.

[14] A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A*. **1999**, *96*, 5482; J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, H. A. Scheraga, *Proteins: Struct. Funct. Genet.* **1999**, *Suppl. 3*, 204; J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, J. A. Saunders, K. D. Gibson, H. A. Scheraga, *Int. J. Quant. Chem.* **2000**, *77*, 90.

[15] F. A. Momany, R. F. McGuire, A. W. Burgess, H. A. Scheraga, *J. Phys. Chem.* **1975**, *79*, 2361; G. Némethy, M. S. Pottle, H. A. Scheraga, *J. Phys. Chem.* **1983**, *87*, 1883.

[16] G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, H. A. Scheraga, *J. Phys. Chem.* **1992**, *96*, 6472.

Received May 22, 2000