Available online at www.sciencedirect.com
SCIENCE DIRECT®
Polymer 45 (2004) 677–686
polymer
www.elsevier.com/locate/polymer

Improved conformational space annealing method to treat β-structure with the UNRES force-field and to enhance scalability of parallel implementation

Cezary Czaplewski[a,b], Adam Liwo[a,b,c], Jarosław Pillardy[a,d], Stanisław Ołdziej[a,b], Harold A. Scheraga[a,*]

[a]*Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA*
[b]*Faculty of Chemistry, University of Gdańsk, ul. Sobieskiego 18, 80-952 Gdańsk, Poland*
[c]*Academic Computer Center in Gdańsk TASK, ul. Narutowicza 11/12, 80-952 Gdańsk, Poland*
[d]*Cornell Theory Center, Cornell University, Ithaca, NY 14853-3801, USA*

Received 25 April 2003; received in revised form 16 May 2003; accepted 16 May 2003

Abstract

Successful application of physics-based protein-structure prediction methods depends on sophisticated computational approaches to global optimization of the conformational energy of a polypeptide chain. One of the most effective procedures for the global optimization of protein structures appears to be the Conformational Space Annealing (CSA) method. CSA is a hybrid method which combines genetic algorithms, essential aspects of the build-up method and a local gradient-based minimization. CSA evolves the population of conformations through genetic operators (mutations, i.e. perturbations of selected geometric parameters, and crossovers, i.e. exchange of selected subsets of geometric parameters between conformations) to a final population optimizing their conformational energy. Implementation of the CSA method with the united-residue force field (UNRES, in which each amino-acid residue is represented by two interaction sites, namely the united peptide group and the united side-chain) was enhanced by introducing new crossover operations consisting of (i) copying β-hairpins, (ii) copying remote strand pairs forming non-local β-sheets, and (iii) copying α-helical segments. A mutation operation, which shifts the position of a β-turn, was also introduced. The new operations promote β-structure, and are essential for searching the conformational space of proteins containing both α- and β-structure; without these operations, excessive preference of α-helical structures is obtained, even though these structures are high in energy. Parallelization of the CSA method has also been enhanced by removing most of the synchronization steps; the improved algorithm scales almost linearly up to 1,000 processors with over 75% average performance.

Keywords: Protein structure prediction; Global optimization; Conformational space annealing

1. Introduction

The theoretical prediction of three-dimensional structures of proteins from a knowledge of only its amino acid sequences is a formidable task [1–3]. All physics-based computational methods for determining protein structure are based on Anfinsen's thermodynamic hypothesis, viz. that the native fold adopted by a polypeptide corresponds to its free energy minimum [4]. Given an amino acid sequence and a fast and accurate approximation of the free energy function, the goal of physics-based protein structure prediction methods is to find the global minimum of this function, which should correspond to the native structure within the applied approximations. Reliable *de novo* protein-structure prediction depends on both an adequate approximation of the free energy function, whose global minimum corresponds to the native fold, and an optimization algorithm that consistently finds that global minimum. This article concentrates on an improvement of the optimization algorithm, and will not discuss in detail the problem of designing an optimal approximation of the free energy function for protein-structure prediction, which is covered elsewhere [5–8].

* Corresponding author. Tel.: +1-607-254-4700; fax: +1-607-255-4034.
E-mail address: has5@cornell.edu (H.A. Scheraga).

0032-3861/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.10.081

Physics-based protein-structure prediction falls into a general and large category of global optimization problems, i.e. problems that involve finding the global minimum or maximum of a target function. All scientific and engineering disciplines have many important problems that fall into this category, and there is a wide range of global optimization algorithms. The difficulty in global optimization of polypeptides arises from the very large conformational space of any reasonable size polypeptide, its high dimensionality, and a ruggedness of this conformational space, i.e. a large number of local minima separated by large barriers. One of the principal limitations impeding successful application of physics-based protein-structure predictions has been time lack of a fast, sophisticated computational approach to global optimization of the conformational energy of a polypeptide chain [1–3]. Most optimization algorithms can be divided roughly into two classes: stochastic and deterministic algorithms. Monte Carlo (MC) based stochastic global optimization methods should be distinguished from the MC methods which are designed to generate a Boltzmann ensemble rather than to identify a single conformation corresponding to the global minimum. Numerous stochastic approaches have been used to attack the global optimization problem in protein structure prediction. MC search [9], simulated annealing [10,11], optimal-bias MC minimization procedure [12,13], MC-stochastic dynamics hybrid [14], torsional flexing [15], catalytic tempering [16], multicanonical jump walking [17], lattice model MC [18], tabu search [19], multiple time step MC [20]. Among the stochastic global optimization algorithms developed in our laboratory for physics-based protein-structure prediction are MC based algorithms (MC with minimization, MCM [21,22]; electrostatically driven MC, EDMC [23–25]; conformation-family MC, CFMC [26]), and genetic algorithms (Conformational Space Annealing, CSA [27–30]). Several deterministic methods have been applied to protein structure prediction: imaginary time Schrödinger equation smoothing [31], packet annealing [32], αBB [33], low-mode search [34], top-down free-energy minimization [35]. Among the deterministic methods, we have studied the build-up method and various deformation-based methods (the diffusion-equation method, DEM [36–38]; the distance-scaling method, DSM [39]). Some methods are hybrids between deterministic and stochastically based methods: the self-consistent basin-to-deformed-basin mapping (SCBDBM) [40] method combines deformation with MC search, and the integrated hybrid method [41] combines αBB and CSA.

At present, one of the most effective procedures for the global optimization of protein structures appears to be the CSA method [27–30]. CSA is a hybrid method, which combines genetic algorithms, essential aspects of the build-up method and a local gradient-based minimization. The method is based on the idea of CSA: in the early stages, it enforces a broad conformational search and then gradually focuses the search into smaller regions with low energy. CSA is designed to search over extremely broad ranges of conformational space, generating numerous local minima before arriving at the global-minimum free energy conformation. Therefore, the CSA searching method allows one to calculate many different groups of low-energy protein structures, one of which is presumably the native structure. CSA was first applied [27,28] to global optimization of peptides containing up to 20 amino acid residues using all-atom models and the ECEPP/3 force field [42]. At present, CSA is also our principal method for optimizing the united-residue (UNRES) energy in our hierarchical procedure for protein-structure prediction [43,44].

Genetic algorithms [45], or more general evolutionary algorithms, are based on theories of biological evolution and natural selection. As with any genetic algorithm, CSA evolves the population of possible solutions through genetic operators (mutations and crossovers) to a final population, optimizing a predefined fitness function. To improve the quality of the solution, and to speed up convergence, CSA incorporates local minimization into an evolutionary algorithm. Such a hybrid approach possesses both the global character of genetic algorithms and also the fast convergence of local searches. In other words, a hybrid approach makes a better tradeoff between the computational cost and the extensiveness of the conformational search. In stochastic MC-based methods, the efficiency depends strongly on a good set of moves that produce a relatively high acceptance ratio, while favoring a broad search of conformational space. Similarly, the efficiency of a genetic algorithm can be enhanced by carefully designed recombination operators. In this article, we present new genetic operators designed for treating β-structures more efficiently. Also, changes in parallel implementation of CSA which greatly improve scalability will be discussed.

2. Methods

2.1. The UNRES force field

In the UNRES model [5–7,46–48], a polypeptide chain is represented by a sequence of α-carbon (C^{α}) atoms linked by virtual bonds with attached united side-chains (SC) and united peptide groups (p). Each united peptide group is located in the middle between two consecutive α-carbons, with peptide group p_i being located between C^{α}_i and C^{α}_{i+1}. Only these united peptide groups and the united side-chains serve as interaction sites, the α-carbons serving only to define the chain geometry (see Fig. 1 of reference [46]). All virtual bond lengths (i.e. C^{α}–C^{α} and C^{α}–SC) are fixed; the distance between neighboring C^{α}'s is 3.8 Å corresponding to *trans* peptide groups, while the side-chain angles (α_{SC}

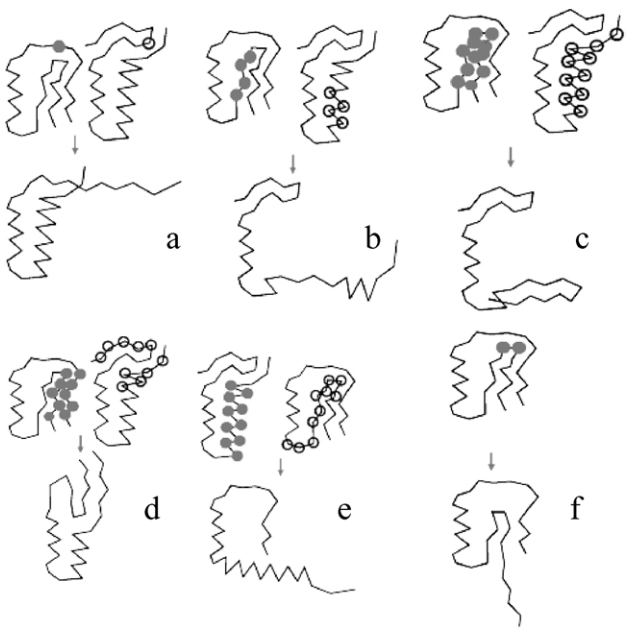


Fig. 1. Illustration of the old and new crossover and mutation operations. Conformations of the bank (which are to be copied) are on the left, while seed conformations (whose conformations are to be replaced) are on the right of the upper parts of each panel. The copied residues of the bank conformations are marked by solid circles, while the residues of the structure to be replaced are marked by open circles. The resulting trial structure is shown on the lower part of each panel. The residues of the resulting trial structure retain all of the original seed conformations except that the replaced residues come from the bank structure. (a) Single residue transfer (operations $\mathcal{O}1$ and $\mathcal{O}2$); (b) transfer of the variables of a number of consecutive residues (operation $\mathcal{O}3$); (c) transfer of a β-hairpin (operation $\mathcal{O}4$); (d) transfer of a non-local pair of β-strands (operation $\mathcal{O}5$); (e) transfer of an α-helical fragment (operation $\mathcal{O}6$); (f) shift of a β-turn (operation $\mathcal{O}7$).

and β_{SC}), and virtual-bond (θ) and dihedral (γ) angles can vary. The energy of the virtual-bond chain is expressed by Eq. (1).

$$U = \sum_{i<j} U_{SC_iSC_j} + w_{SCp} \sum_{i\neq j} U_{SC_ip_j} + w_{el} \sum_{i<j-1} U_{p_ip_j} + w_{tor} \sum_i U_{tor}(\gamma_i) + w_{tord} \sum_i U_{tord}(\gamma_i, \gamma_{i+1}) + w_b \sum_i U_b(\theta_i) + w_{rot} \sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i}) + \sum_{m=2}^{N_{corr}} w^m_{corr} U^m_{corr} \quad (1)$$

The term $U_{SC_iSC_j}$ represents the mean free energy of the hydrophobic (hydrophilic) interactions between the side-chains, which implicitly contains the contributions from the interactions of the side-chain with the solvent. The term $U_{SC_ip_j}$ denotes the excluded-volume potential of the side-chain–peptide-group interactions. The peptide-group interaction potential ($U_{p_ip_j}$) accounts mainly for the electrostatic interactions (i.e. the tendency to form backbone hydrogen bonds) between peptide groups p_i and p_j. U_{tor}, U_{tord}, U_b, and U_{rot} represent the energies of virtual-dihedral angle torsions, double torsions, virtual-bond angle bending, and side-chain rotamers; these terms account for the local propensities of the polypeptide chain. Details of the parameterization of all of these terms are provided in earlier publications [46,47]. Finally, the terms U^m_{corr}, $m = 1, 2, \ldots N_{corr}$ are the correlation or multibody contributions from a cumulant expansion [48] of the restricted free energy (RFE), and the w's are the weights of the energy terms. The multibody terms are indispensable for reproduction of regular α-helical and β-sheet structures. The UNRES force field has been derived as an RFE function of an all-atom polypeptide chain plus the surrounding solvent, where the all-atom energy function is averaged over the degrees of freedom that are lost when passing from the all-atom to the simplified system (i.e. the degrees of freedom of the solvent), the dihedral angles χ for rotation about the bonds in the side-chains, and the torsional angles λ for rotation of the peptide groups about the $C^\alpha \cdots C^\alpha$ virtual bonds. This approach enabled us to derive the multibody terms U^m_{corr}, $m = 1, 2, \ldots, N_{corr}$ by a generalized cumulant expansion of the RFE developed by Kubo [49]. The internal parameters of the individual U's were derived by fitting the analytical expressions to the RFE surfaces of model systems [48] or by fitting the calculated distribution functions to those determined from the PDB [47], while the w's (the weights of the energy terms) were calculated by optimization of the energy gap between the lowest-energy native-like conformation and the lowest-energy non-native conformation (ΔE) and the Z-score (Z, defined as the difference between the mean energy of the native-like structures and the mean energy of the non-native structures divided by the standard deviation of the energy of the non-native structures) of the training proteins [5,6,47].

$$\Delta E = \min_{i\in nat} E_i - \min_{i\in non-nat} E_i \quad (2)$$

$$Z = \frac{(1/N_{nat})\sum_{i=1}^{N_{nat}} E_i - (1/N_{non-nat})\sum_{i=1}^{N_{non-nat}} E_i}{\sqrt{(1/N_{non-nat})\sum_{i=1}^{N_{non-nat}} E_i^2 - \left[(1/N_{non-nat})\sum_{i=1}^{N_{non-nat}} E_i\right]^2}} \quad (3)$$

The force field is able to predict the structures of proteins containing both α-helical and β-sheet structures with a reasonable degree of accuracy, as assessed by tests on model proteins [30,50,51] as well as in the CASP3 [30,52,53], CASP4 [50], and CASP5 blind prediction experiments.

2.2. The conformational space annealing method

The CSA algorithm [27–30] is summarized below because the details of its implementation are necessary for

the discussion of new genetic operators and the redesign of its parallel implementation. The CSA method begins with a randomly-generated population of conformations which are energy minimized to generate the *first bank* of conformations. The first bank is meant to represent a sparse sampling of the conformational space that captures short-range interactions. From the initial population, a number of conformations (called seeds) are selected as parents for the trial population. These 'seed' conformations are altered in a non-random fashion to create new trial conformations. As in any genetic algorithm, the trial population is generated by the use of genetic operators: mutations and crossovers. Unlike traditional genetic algorithms, the mutation operator applied in CSA does not change the value of the selected variable randomly; instead, it uses values of the corresponding variables in the initial population (the first bank) or in the current population of conformations as a pool of random numbers. A copy of the first bank is used as a source of 'random' variables, which are not uniformly distributed, but their distribution is determined by intramolecular interactions at this stage determined mainly by steric overlap. The crossover operators copy a set of variables, representing a continuous segment of the polypeptide chain of various size taken from a randomly selected conformation in the current population, to a selected parent conformation (seed). This is described in detail in Section 2.3. Attention is paid to assure that all trial conformations are significantly different from each other and from parent conformations. After generation, all trial conformations are energy minimized.

The next step of the CSA algorithm is the update of the current population (the bank) without increasing its size. Each trial conformation is compared to each existing conformation of the bank. If the trial conformation is similar to an existing conformation of the bank, only the lower-energy conformation out of these two is preserved. If the trial conformation is not similar to any existing conformation in the bank, it represents a new distinct region of conformational space. Then it replaces the highest-energy conformation in the bank, if its energy is lower than the highest energy in the bank, otherwise it is discarded. The distance between conformations i and j is defined as the difference of their virtual-bond angles and virtual-bond dihedral angles [Eq. (9) of Ref. 52]. If the distance, D_{ij}, is less than or equal to some predefined cutoff value, D_{cut}, conformations i and j are considered similar, otherwise they are considered different. CSA achieves its efficiency by beginning with a large value of D_{cut} to essentially search all possible structures, and then gradually reduces ('anneals') D_{cut} by reducing the minimum distance between the conformations of the bank and focusing the search in low-energy regions of conformational space. After updating the current population, the seed conformations are selected from the set of conformations not selected as seeds previously; in addition, attention is paid to cover the conformational space as broadly as possible by selecting conformations not similar to each other as seed conformations.

2.3. Introducing new crossover operators

As mentioned in Section 2.2, the crossover operations copy some variables from randomly selected conformations of the bank to the corresponding variables of the seed conformations. In the version of the CSA method used in the CASP3 experiment [30,43], the following operators were applied:

$\mathcal{O}1$: *Exchange of backbone or side-chain variables of a single residue*. Residue i is selected at random and either the backbone (γ_i, θ_i) or the side chain (α_i, and β_i) variables are copied from the selected bank conformation to the selected seed conformation.

$\mathcal{O}2$: *Exchange of all variables of a single residue*. This operation differs from that of $\mathcal{O}1$ by exchanging all variables (γ_i, θ_i, α_i, and β_i).

$\mathcal{O}3$: *Exchange of the variables of n consecutive residues*. The variables of residues from i_1 to i_2, where $2 < i_2 - i_1 < n/3$, n being the number of residues in the molecule, are copied from the selected bank conformation to the selected seed conformation.

A schematic representation of the operations described in $\mathcal{O}1$–$\mathcal{O}3$ is shown on Fig. 1(a) and (b). These operations are sufficient to search the conformational space of proteins with simple topology. However, it can easily be demonstrated that operation $\mathcal{O}3$ introduces a strong bias towards structures with efficient short-range interactions, i.e. the α-helices (see Section 3.1 for a numerical demonstration of this fact). Suppose we have a 20-residue α-helical segment in the bank conformation which is to be copied. Then, if we copy a 10-residue fragment, the number of ways in which at least a five residue α-helical fragment is included is 11 (the number of 10-residue helical segments of the 20-residue helix) +5 (the number of helical segments at least five residues long including the amino terminus of the 20-residue helix and the residues preceding it) +5 (the number of helical segments at least five residues long including the carboxy terminus of the 20-residue helix and the residues following it) = 21. If we have a 20-residue β-hairpin in the structure which is to be copied, the number of 10-residue fragments with at least one β-hairpin peptide–peptide contact (corresponding to the middle β-turn) is 7, i.e. three times smaller than the number of fragments with a helical segment of at least five residues. Therefore, this recombination scheme includes a strong bias towards α-helical structure even for simple combinatorial reasons. Moreover, elements of β-structure are stabilized by long-range interactions and, thereby, formation of the sufficiently regular structures with optimal interactions is more difficult compared to the stabilization of α-helices with short-range interactions.

To overcome this undesirable bias, we introduced two new crossover operations which copy portions of the β structure:

𝒪4: *Transfer of a single β-hairpin*. The bank conformations are analyzed for the presence of β-hairpins (see Section 2.4). If a β-hairpin is detected in a bank conformation, the corresponding variables are transferred to a selected seed conformation.

𝒪5: *Transfer of a pair of remote interacting β-strands*. The bank conformations are analyzed for the presence of remote strands forming β-sheets. If such a pair is found, a pair of remote interacting strands is selected and the contacts between these strands are transferred to the seed conformations. This is accomplished as follows:

1. The variables describing the geometry of the two strands are copied from the bank structure to the seed structure.
2. A local minimization of a simplified potential-energy function containing a harmonic penalty for the preservation of the contacts between the copied strands, defined by Eq. (4), is carried out in the variables that do not correspond to well-defined secondary structure of the original seed structures or to the copied strands.

$$V = \sum_{ij} \tilde{U}_{\mathrm{SC}_i\mathrm{SC}_j} + w_{\mathrm{el}} \sum_{ij} \tilde{U}_{\mathrm{p}_i\mathrm{p}_j} + w_{\mathrm{SCp}} \sum_{ij} \tilde{U}_{\mathrm{SC}_i\mathrm{p}_j} + w_{\mathrm{tor}} \sum_i U_{\mathrm{tor}}(\gamma_i) + w_{\mathrm{tord}} \sum_i U_{\mathrm{tord}}(\gamma_i, \gamma_{i+1}) + w_{\mathrm{b}} \sum_i U_{\mathrm{b}}(\theta_i) + w_{\mathrm{rot}} \sum_i U_{\mathrm{rot}}(\alpha_i, \beta_i) + w_{\mathrm{dis}} \sum_{i\in s1, j\in s2} (d_{\mathrm{C}_i^\alpha \mathrm{C}_j^\alpha} - d_{\mathrm{C}_i^\alpha \mathrm{C}_j^\alpha}{}^\circ)^2 \quad (4)$$

where $\tilde{U}_{\mathrm{SC}_i\mathrm{SC}_j}$, $\tilde{U}_{\mathrm{p}_i\mathrm{p}_j}$, and $\tilde{U}_{\mathrm{SC}_i\mathrm{p}_j}$ are 'soft-sphere' [54] versions of the $U_{\mathrm{SC}_i\mathrm{SC}_j}$, $U_{\mathrm{p}_i\mathrm{p}_j}$, and $U_{\mathrm{SC}_i\mathrm{p}_j}$ potentials, in which the Lennard–Jones-like potential is replaced with the function defined by Eq. (5), and $s1$ and $s2$ denote the first and the second copied strand, respectively, $d_{\mathrm{C}_i^\alpha \mathrm{C}_j^\alpha}$ is the distance between C_i^α and C_j^α in the trial conformation and $d_{\mathrm{C}_i^\alpha \mathrm{C}_j^\alpha}{}^\circ$ is the corresponding distance in the structure from the bank from which the fragment is copied.

$$U = \begin{cases} \frac{1}{4}(r_{ij}^2 - r_{ij}{}^{\circ 2}) & \text{for } r_{ij} < r_{ij}{}^\circ \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

with r_{ij} being the distance between the interacting sites and $r_{ij}{}^\circ$ being the collision distance.

3. Local minimization of the UNRES energy function of Eq. (1) supplemented with the distance constraints corresponding to the contacts between the all C^α-atoms of the copied strands is carried out with the same conditions as in point 2.
4. Local minimization of the UNRES energy function of Eq. (1) supplemented with the distance constraints corresponding to the contacts between the C^α-carbon atoms of the copied strands is carried out with variation of all geometric parameters of the chain.
5. Unrestricted optimization of the UNRES energy function of Eq. (1) is carried out.

An example in which a non-local fragment of β-structure is copied is shown in Fig. 2.

The reason for using the complex procedure described in points 2–5 is that unrestricted local minimization of the UNRES energy function supplemented with interstrand distance constraints typically destroys the secondary structure already present in the seed structure, because there are many clashes between the interacting sites after copying a strand pair. Freezing the variables corresponding to fragments with well-defined secondary structure, with initial use of a 'soft-sphere' long-range interaction potential, facilitates the removal of the worst overlaps in step 2. The overlaps are further released in step 3 when the full UNRES energy function is applied; however, the variables corresponding to segments with well-defined secondary structure are still frozen. In step 4 all geometric parameters are varied; however, the distance constraints to maintain the introduced β-sheet are still imposed; these constraints are released in step 5, which produces a relaxed hybrid structure. It should be stressed that the restraints are imposed only temporarily to enable copying non-local elements of geometry, but the whole process is analogous to copying a contiguous part of the chain in operation 𝒪3.

Apart from the operations described above, that copy portions of β-structure, we introduced the following

𝒪6: Copying an α-helical structure from a bank to a seed conformation. This is accomplished as in operation 3: the bank conformations are scanned for the presence of α-helices and a randomly selected α-helix is copied in the corresponding place of the sequence of the seed conformation.

𝒪7: A mutation operator, which shifts the turn of a β-hairpin

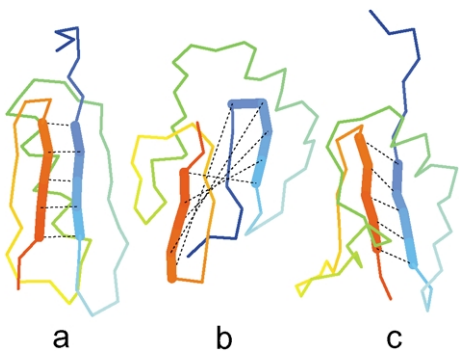


Fig. 2. Example of a hybrid structure (c) generated by transfer of a non-local pair of interacting strands from a bank conformation (a) to a seed conformation (b). The copied strands are marked by thick lines and selected target contacts between the strands are marked by dashed lines.

of a selected seed conformation by ± 2 residues. This operation facilities the correcting of misplaced β-turns, which otherwise would be difficult to accomplish.

A schematic representation of the new crossover operations is shown in Fig. 1(c)–(f).

2.4. Recognizing secondary structure

To recognize secondary structure, we adopted the procedure used in our dipole-path method to reconstruct an all-atom backbone from the C^{α} trace [55,56] using the following algorithm

1. The electrostatic-contact map between peptide groups is constructed. Two peptide groups are in contact, if their average electrostatic-interaction energy computed from Eq. (5) of Ref. [57] is less than the cutoff value ΔE_{cut}; we use $E_{\mathrm{cut}} = -0.3$ kcal/mol for peptide groups separated by more than three $C^{\alpha}\cdots C^{\alpha}$ virtual bonds and $E_{\mathrm{cut}} = -0.5$ kcal/mol for peptide groups separated by three virtual bonds. A contact between peptide groups p_i and p_j means that either C_i^{α} is close to C_j^{α} and C_{i+1}^{α} is close to C_{j+1}^{α}, if the segments of the chain consisting of peptide groups p_i and p_j are parallel, or C_i^{α} is close to C_{j-1}^{α} and C_{i+1}^{α} is close to C_j^{α}, if they are antiparallel.
2. An α-helix is defined in segment C_i^{α}–C_j^{α}, if the following two conditions hold
   (a) Every peptide group in the segment is in electrostatic contact with its third neighbor.
   (b) $10^{\circ} < \gamma_k < 80^{\circ}$ for all $i < k < j$. This condition eliminates left-handed helices.
3. A β-hairpin is defined in segment C_i^{α}–C_{i+2k-1}^{α} if, for every $i \le j < i + k$ peptide group, p_j is in electrostatic contact with peptide group $p_2(i + k - 1) - j$.
4. Two strands from C_i^{α} to C_{i+k}^{α} and from C_j^{α} to C_{j+k}^{α} form a parallel β-sheet, if peptide group p_l, where $i \le l < i + k$, is in electrostatic contact with peptide group p_{j+l-i}.
5. Two strands from C_i^{α} to C_{i+k}^{α} and from C_j^{α} to C_{j-k}^{α} form an anti-parallel β-sheet if peptide group p_l, where $i \le l < i + k$, is in electrostatic contact with peptide group p_{j-l+i}.

2.5. Improving scalability of the CSA algorithm

Any global optimization method, including CSA, applied to protein-structure prediction, typically requires a huge computational effort. Even the fastest processors available are not fast enough to carry out these kinds of computations in real time. To solve this problem, the CSA method takes advantage of massively parallel computing. The CSA algorithm can be divided into two parts: the algorithmic part described above and local minimization of trial conformations, the latter being the most computationally intensive. Parallelization by a master/worker approach, in which the master executes only the algorithmic part of CSA while minimization of trial conformations is distributed to all workers, has been described in detail in our earlier work [44]. In brief, the master generates a number of trial conformations and sends them one by one to the worker processors for local minimization; a current conformation is sent to the first node, which is not busy with local minimization at the moment. After all trial conformations are sent, all processes are synchronized, until all energy-minimized conformations are returned to the master processor. Because local minimizations take very different CPU time depending on the conformation, synchronization violates load balancing. This violation depends strongly on the ratio of the number of trial conformations to the number of processors; about 80% efficiency is achieved when this ratio is 10:1 which, however, impairs massive parallelization of the algorithm (e.g. with typically 600 trial conformations, good scalability can be expected only up to 60 processors) [44]. In practice, the scalability of the algorithm has been far from perfect, reaching 43% on 100 processors and 200 trial conformations [44].

To improve the load balancing, we eliminated the synchronization step. The new algorithm consists of the following steps.

1. At the beginning of the procedure, the master generates a number of random conformations, which are sent to the workers for local minimization. This step is completed when all conformations have been energy-minimized and therefore synchronization does take place here.
2. In a given iteration, the master generates a number of trial conformations (it has to be greater than the number of worker processors). The conformations are sent to the worker processors for local minimization; however, the master concludes the iteration as soon as all trial conformations have been sent. The conformations that are still not returned will be collected in the next iteration.
3. The collected energy-minimized conformations (both from the current iterations and from the previous iterations returned in the current iteration) are used to update the bank. If a pre-defined total number of conformations has already been generated and energy minimized, the procedure stops and the remaining conformations are collected in the final synchronization step; otherwise, the procedure returns to step 2.

3. Results and discussion

3.1. Significance of β-structure copying operations in searching the conformational space

Our test case was the 61-residue IgG α + β protein (PDB code: 1IGD [58]). This protein consists of N- and the C-terminal hairpins packed together to form a parallel β-sheet

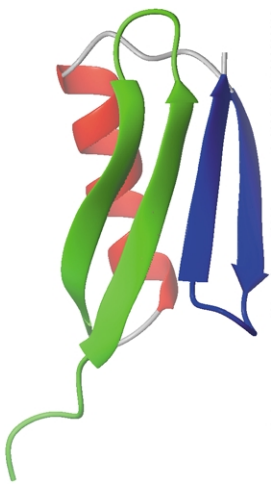


Fig. 3. Native structure of the 1IGD protein. The N-terminal hairpin, α-helix, and the C-terminal hairpin are colored green, red, and blue, respectively.

and the middle α-helix packed against the β-structure (Fig. 3). This protein was one of the toughest tests of the UNRES force field, because a very small change of force-field parameters results in a dramatic change in the structure of the lowest-energy conformation. Also, the conformational search with the UNRES force field proved particularly hard for this protein.

Using our hierarchical method of force-field optimization [7], we recently managed to obtain a 'caldera-like' force field by optimizing the UNRES potential using 1IGD as the benchmark protein [59]. The lowest-energy structure of 1IGD in this force field has all three native secondary structure elements with proper packing of β-hairpins into a parallel β-sheet (Fig. 4(c)). The 'caldera-like' property means that the search goes quickly to the global minimum, because there is a strong negative energy gradient with increasing degree of native-likeness.

Fig. 5 compares the time course of two CSA runs with different settings, and representative structures of

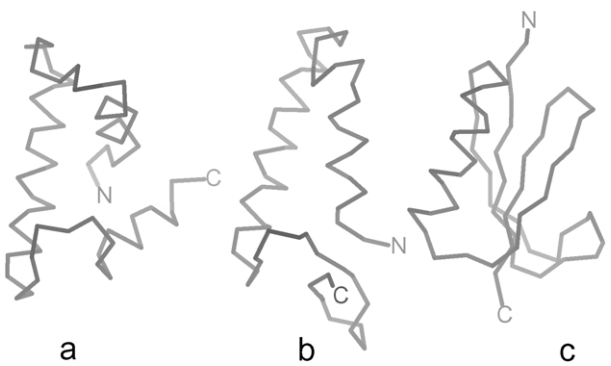
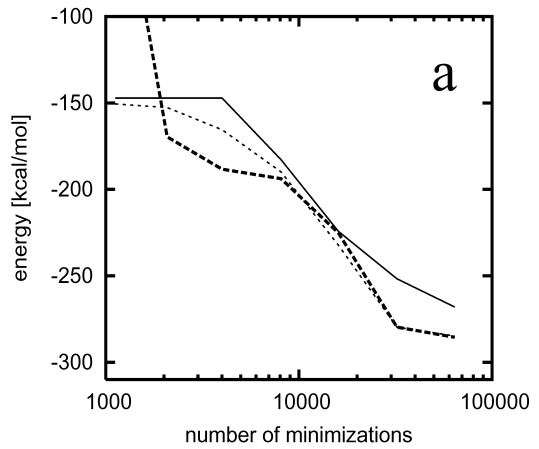


Fig. 4. (a) The lowest-energy structure obtained in the CSA run without including β-sheet transfer operations ($E = -285.5$ kcal/mol); (b) the second-lowest energy structure in that run ($E = -284.8$ kcal/mol); (c) the lowest-energy structure (native-like) obtained in the run with inclusion of β-sheet transfer operations ($E = -308.7$ kcal/mol).



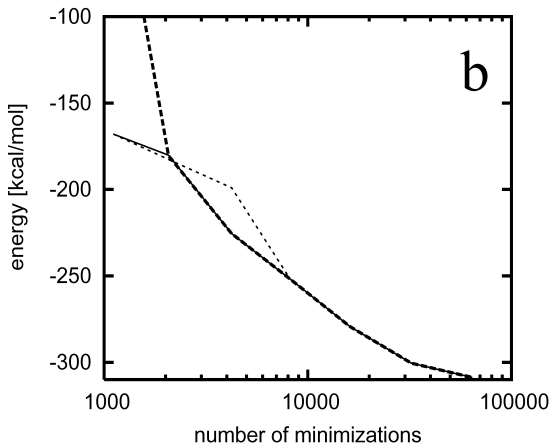


Fig. 5. Plots of the lowest energy of structures with native secondary-structure elements obtained during the course of the CSA simulations with a 'caldera-like' UJNRES force field without (a) and with (b) inclusion of β-sheet transfer operations vs. the number of minimizations. Solid line: energies of structures with N-terminal hairpin; dashed line: energies of structures with middle α-helix; dotted line: energies of structures with a C-terminal hairpin. When β-sheet transfer operations are included (part b) the lowest-energy structure has all three elements after less than 10,000 energy minimizations.

these runs are shown in Fig. 4. In the first run, no β-sheet-promoting operations were included while, in the second one, these operations were added (operations 3 and 4 of Section 2.3). In the first run, the resulting lowest-energy structure is a three-helix bundle (Fig. 4(a)), regardless of the fact that the native-like α + β structure is about 25 kcal/mol lower in energy than this structure. The second lowest-energy structure contains the C-terminal β-hairpin (Fig. 4(b)), but the N-terminal part is folded into an α-helix. When the β-hairpin moves are added, the lowest-energy structure is a native-like structure (Figs. 4(c) and 5(b)).

As shown in Fig. 5(a), structures with the N- arid the C-terminal β-hairpins already appear early in the run that does not include the new β-structure-promoting operations. However, initially, they have substantially higher energy

than structures with α-helices formed in place of β-hairpins, because it is easier for an α-helical fragment to attain an optimal geometry, compared to a β-hairpin that is stabilized by long-range contact, which has a lower entropy of formation. Moreover, as mentioned in Section 2.3, the chance that a β-sheet fragment will be copied from a bank conformation containing β-structure is significantly lower than the probability of copying an α-helical fragment from a bank conformation containing α-helical structure. Thus, the original version of CSA method contains a significant bias towards forming α-helical structure.

The tendency of the original CSA method to falsely promote α-helical structure is even better illustrated in our second example. In this example, we used a version of the UNRES force field from an early stage of optimization on 1IGD; this force field locates the native-like structure in CSA runs, but the lowest-energy structure is a full β-sheet structure. The time course of the CSA runs is shown in Fig. 6. The three-helix bundle is the lowest-energy structure if no β-sheet-promoting operations are included (Fig. 7(a)). When these operations are included, the lowest-energy structure is a full β-sheet (Fig. 7(b)) with almost the same energy as the three-helix bundle. This β-sheet structure has quite short range contacts, as all strands that are consecutive in sequence are packed together with each other. Introduction of more β-sheet-promoting operations leads to a β-sandwich-like structure (Fig. 7(c)) with more than 20 kcal/mol lower energy. It is interesting to note that the lowest-energy structure with long-range contacts between the N- and C-terminal β-strands (Fig. 7(d)) is attained only when a copy of a non-local β-sheet is included (operation ℴ4 of Section 2.3).

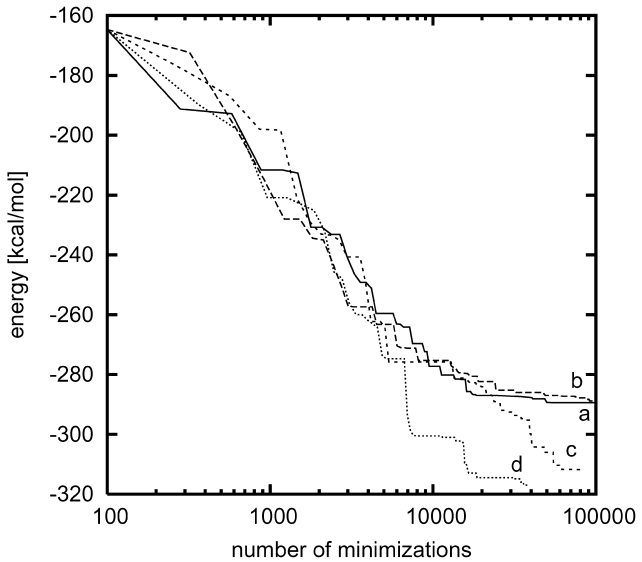


Fig. 6. Plots of the lowest energy vs. the number of minimizations during the course of the CSA simulations with an UNRES force field that gives an incorrect all-β structure as the global minimum of 1IGD: (a) no β-structure-transfer moves included; (b) some β-hairpin-transfer moves included; (c) more β-hairpin-transfer moves included; (d) β-hairpin and non-local β-structure-transfer moves included.

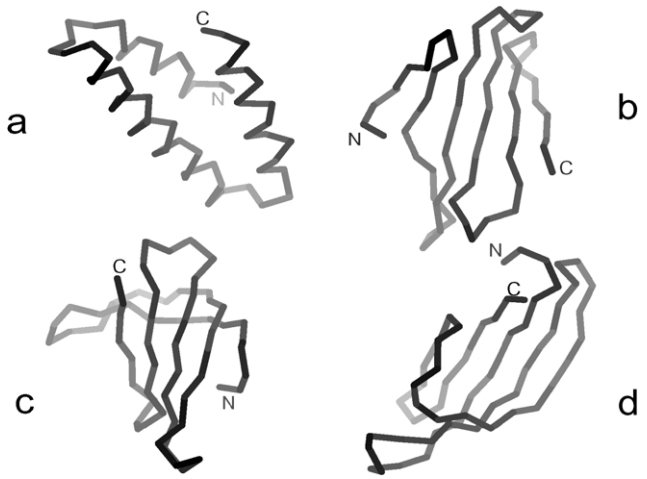


Fig. 7. (a)–(d) The lowest-energy structures corresponding to runs (a)–(d) of Fig. 6. The energies are −289.4 kcal/mol, −290.1 kcal/mol, −311.8 kcal/mol, and −317.3 kcal/mol, respectively.

3.2. Improving scalability of the CSA algorithm

Fig. 8 compares the speedup of the original CSA algorithm with synchronization, after generating and minimizing new conformations with the parallel implementation introduced in this work. Frequent synchronization causes substantial deterioration of the performance with an increase in the number of processors. The speedup of the original parallel CSA algorithm depends strongly on the ratio of the number of trial conformations to the number of processors. The scalability curves for CSA runs with 50, 100, 200, and 400 trial conformations generated per iteration (Fig. 8(a)) show almost linear scaling amounting to 80% of the maximum speedup for a number of processors equal to 5, 10, 20 and 40, respectively; however, the speedup goes down for more processors. Removing synchronization in the new algorithm results in linear scaling up to 80 processors with only 200 trial conformations generated per iteration. This allows massively parallel computations, as CSA scales with 75% average efficiency until up to 1,000 processors (Fig. 8(b)) with only 1,100 trial conformations generated per iteration.

4. Conclusions

In this work we have proposed an improvement of the CSA method, applied to the UNRES force field to treat proteins with β- as well α-structure. We introduced two new crossover operations, which copy β-hairpins or fragments of β-structure composed of remote strands. These new operations are essential in a conformational search of proteins containing both α and β structure; without including them, the system is likely to end up in an α-helical conformation despite the fact, that this conformation is high in energy. This is caused by the fact that, in the original implementation of CSA, contiguous segments of

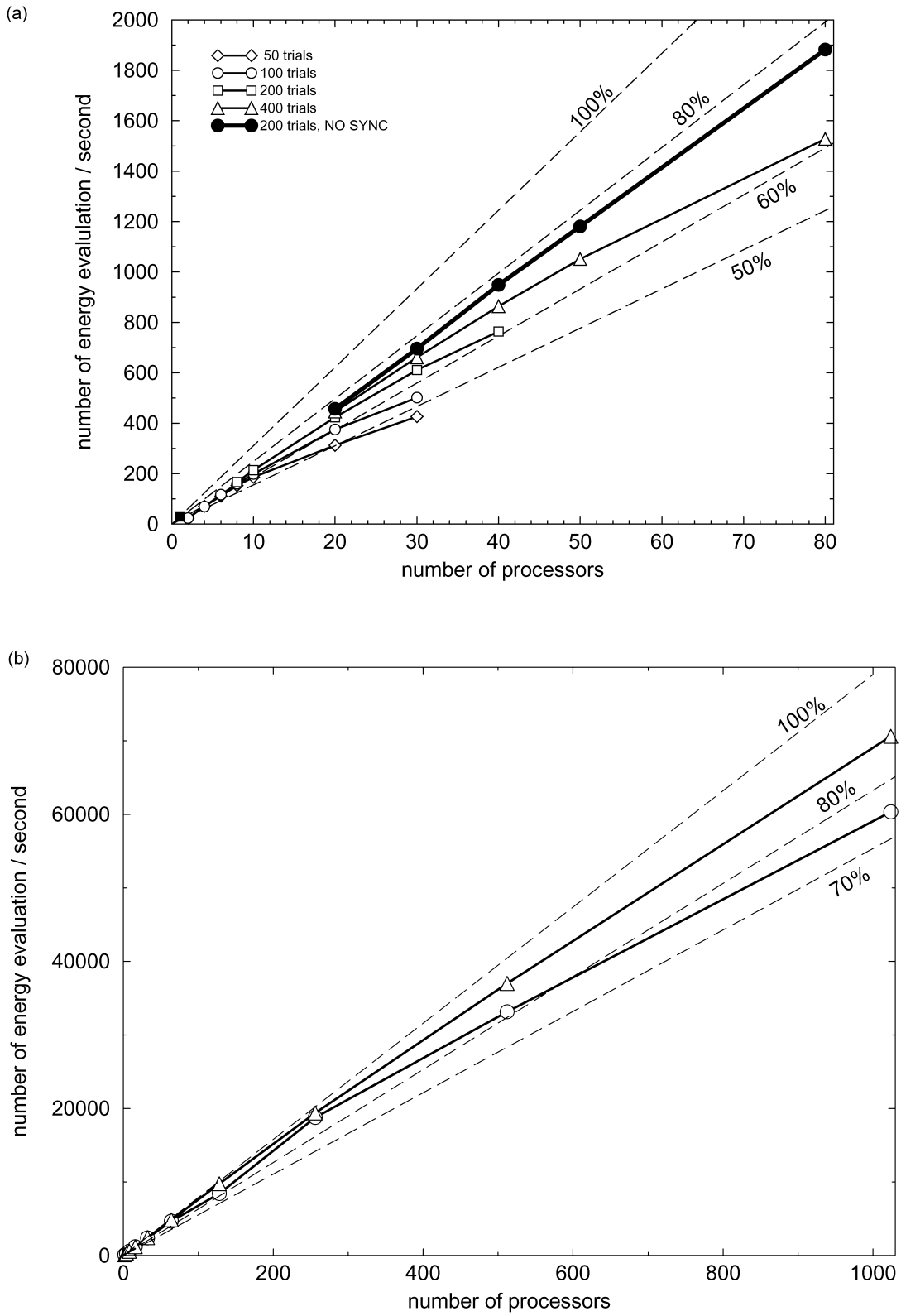


Fig. 8. (a) Comparison of the speedup of the old parallel implementation of the CSA algorithm, with synchronization after each CSA iteration with 50 (◇), 100 (○), 200 (□), and 400 (△) trial conformations generated per iteration, with the speedup of the new implementation (● and heavy line). The dashed lines correspond to 50, 60, 80, and 100% of the theoretical efficiency, respectively. (b) Scalability plot of the new CSA algorithm in massively parallel computations. (○): average efficiency (averaged over all iterations); (△): peak efficiency.

the structure are copied; this impairs the transfer of segments of β-sheet structures, while it does not impair the transfer of α-helical structure.

The new operations can be generalized to copying more than two-strand portions of β-structures and portions of supersecondary and tertiary structure, such as the helix–turn–helix motifs, zinc finger α/β motifs, etc. as well as to transferring established patterns of side-chain contacts. Work on this is in progress in our laboratory. We have also removed excessive synchronization from the parallel implementation of the CSA algorithm. Synchronization is now carried out only after the first bank of conformations is generated and upon termination of the procedure. The new algorithm scales almost linearly up to 1,000 processors with 75% average efficiency.

Acknowledgements

This work was supported by grants from the National Institutes of Health (GM-14312), the National Science Foundation (MCB00-03722), the Fogarty Foundation (TW1064), and grant BW 8000-5-0234-3 from the Polish State Committee for Scientific Research (KBN). Support was also received from the National Foundation for Cancer Research. This research was conducted by using the resources of (a) The National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (b) our 392-processor Beowulf cluster at Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (c) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk, (d) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and (e) The Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

References

[1] Park BH, Levitt M. J Mol Biol 1995;249:493–507.
[2] Kihara D, Lu H, Kolinski A, Skolnick J. Proc Natl Acad Sci USA 2001;98:10125–30.
[3] Scheraga HA, Pillardy J, Liwo A, Lee J, Czaplewski C, Ripoll DR, Wedemeyer WJ, Arnautova YA. J Comput Chem 2002;23:28–34.
[4] Anfinsen CB. Science 1973;181:223–30.
[5] Lee J, Ripoll DR, Czaplewski C, Pillardy J, Wedemeyer WJ, Scheraga HA. J Phys Chem B 2001;105:7291–8.
[6] Pillardy J, Czaplewski C, Liwo A, Wedemeyer WJ, Lee J, Ripoll DR, Arlukowicz P, Ołdziej S, Arnautova YA, Scheraga HA. J Phys Chem B 2001;105:7299–311.
[7] Liwo A, Arlukowicz P, Czaplewski C, Ołdziej S, Pillardy J, Scheraga HA. Proc Natl Acad Sci USA 2002;99:1937–42.
[8] Arnautova YA, Jagielska A, Pillardy J, Scheraga HA. J Phys Chem B, 2003;107:7143–7154.
[9] Derreumaux P. J Chem Phys 1999;111:2301–10.
[10] Kirkpatrick S, Gelatt CD, Vecchi MP. Science 1983;220:671–80.
[11] Andricioaei I, Straub J. Phys Rev E 1996;53:R3055–8.
[12] Abagyan RA, Totrov M. J Mol Biol 1994;235:983–1002.
[13] Abagyan RA, Totrov M. J Comput Phys 1999;151:402–21.
[14] Guarnieri F, Still W. J Comput Chem 1994;15:1302–10.
[15] Kolossvary I, Guida W. J Comput Chem 1993;14:691–8.
[16] Stolovitzky G, Berne B. Proc Natl Acad Sci USA 2000;97:11164–9.
[17] Xu H, Berne B. J Chem Phys 1999;110:10299–306.
[18] Skolnick J, Kolinski A, Yaris R. Biopolymers 1989;27:1059.
[19] Morales L, Garduño-Juàrez R, Aguilar-Alvarado JM, Riveros-Castro F. J Comput Chem 2000;21:147–56.
[20] Hetenyi B, Bernacki K, Berne B. J Chem Phys 2003;117:8203–7.
[21] Li Z, Scheraga HA. Proc Natl Acad Sci USA 1987;84:6611–5.
[22] Li Z, Scheraga HA. J Mol Struct (Theochem) 1988;179:333–52.
[23] Ripoll DR, Scheraga HA. Biopolymers 1988;27:1283–303.
[24] Ripoll DR, Scheraga HA. J Prot Chem 1989;8:263–87.
[25] Ripoll DR, Liwo A, Scheraga HA. Biopolymers 1998;46:117–26.
[26] Pillardy J, Czaplewski C, Wedemeyer WJ, Scheraga HA. Helv Chim Acta 2000;83:2214–30.
[27] Lee J, Scheraga HA, Rackovsky S. J Comput Chem 1997;18:1222–32.
[28] Lee J, Scheraga HA, Rackovsky S. Biopolymers 1998;46:103–15.
[29] Lee J, Scheraga HA. Int J Quant Chem 1999;75:255–65.
[30] Lee J, Liwo A, Scheraga HA. Proc Natl Acad Sci USA 1999;96:2025–30.
[31] Amara P, Hsu D, Straub JE. J Phys Chem 1993;97:6715–21.
[32] Oresić M, Shalloway D. J Chem Phys 1994;101:9844–57.
[33] Androulakis IP, Maranas CD, Floudas CA. J Glob Opt 1995;7:337–63.
[34] Kolossvary I, Guida W. J Am Chem Soc 1996;118:5011–9.
[35] Church B, Shalloway D. Proc Natl Acad Sci USA 2001;98:6098–103.
[36] Piela L, Kostrowicki J, Scheraga HA. J Phys Chem 1989;93:3339–46.
[37] Kostrowicki J, Piela L, Cherayil BJ, Scheraga HA. J Phys Chem 1991;95:4113–9.
[38] Kostrowicki J, Scheraga HA. J Phys Chem 1992;96:7442–9.
[39] Pillardy J, Piela L. J Comput Chem 1997;18:2040–9.
[40] Pillardy J, Liwo A, Groth M, Scheraga HA. J Phys Chem B 1999;103:7353–66.
[41] Klepeis J, Pieja M, Floudas C. Comp Phys Commun 2003;151:121–40.
[42] Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini C, Zagari A, Rumsey S, Scheraga HA. J Phys Chem 1992;96:6472–84.
[43] Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Proc Natl Acad Sci USA 1999;96:5482–5.
[44] Lee J, Pillardy J, Czaplewski C, Arnautova Y, Ripoll DR, Liwo A, Gibson KD, Wawak RJ, Scheraga HA. Comp Phys Commun 2000;128:399–411.
[45] Goldberg DE, Genetic algorithms in search, optimization and machine learning, Reading, MA: Addison-Wesley; 1989.
[46] Liwo A, Ołdziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. J Comput Chem 1997;18:849–73.
[47] Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Ołdziej S, Scheraga HA. J Comput Chem 1997;18:874–87.
[48] Liwo A, Czaplewski C, Pillardy J, Scheraga HA. J Chem Phys 2001;115:2323–47.
[49] Kubo R. J Phys Soc Jpn 1962;17:1100–20.
[50] Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kaźmierkiewicz R, Ołdziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye YJ, Scheraga HA. Proc Natl Acad Sci USA 2001;98:2329–33.
[51] Liwo A, Pillardy J, Kaźmierkiewicz R, Wawak RJ, Groth M, Czaplewski C, Ołdziej S, Scheraga HA. Theor Chem Acc 1999;101:16–20.
[52] Lee J, Liwo A, Ripoll DR, Pillardy J, Saunders JA, Gibson KD, Scheraga HA. Int J Quant Chem 2000;71:90–117.
[53] Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Prot Struct Funct Genet Suppl 1999;3:149–70.
[54] Vásquez M, Scheraga HA. J Biomol Struct Dyn 1988;5:705–55.
[55] Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. Prot Sci 1993;2:1697–714.
[56] Kaźmierkiewicz R, Liwo A, Scheraga HA. J Comput Chem 2002;23:715–23.
[57] Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. Prot Sci 1993;2:1715–31.
[58] Derrick JP, Wigley DB. J Mol Biol 1994;243:906–18.
[59] Liwo A, Czaplewski C, Ołdziej S, Pillardy J, Arlukowicz P, Scheraga HA. J Phys Chem B, in preparation.