Journal of Computational Physics **151,** 402–421 (1999)
Article ID jcph.1999.6233, available online at http://www.idealibrary.com on IDEAL®

Ab Initio Folding of Peptides by the Optimal-Bias Monte Carlo Minimization Procedure

Ruben A. Abagyan* and Maxim Totrov

Biochemistry Department, Skirball Institute of Biomolecular Medicine, and Courant Institute of Mathematics, New York University, 540 First Avenue, New York, New York 10016

Received August 20, 1998; revised February 15, 1999

Prediction of three-dimensional structures of proteins and peptides by global optimization of the free energy estimate has been attempted without much success for over thirty years. The key problems were the insufficient accuracy of the free energy estimate and the giant size of the conformational space. Global optimization of the free energy function of a peptide in internal coordinate space is a powerful method of structure prediction that outperforms both Molecular Dynamics, bound by the continuity requirement, and Monte Carlo, bound by the Boltzmann ensemble generation requirement. We demonstrate that stochastic global optimization algorithms of the first order, i.e., with local minimization after each iteration (e.g., Monte Carlo-Minimization), have a greater chance of finding the global minimum after a fixed number of function evaluations. Recently, the principle of optimal bias was mathematically justified and the Optimal-Bias Monte Carlo-Minimization algorithm (a.k.a. Biased Probability Monte Carlo-minimization) was successfully applied to theoretical *ab initio* folding of several peptides, resulting in more than a 10-fold increase in efficiency compared to the Monte Carlo-Minimization method. The square-root bias is shown to be comparable in performance with the previously derived linear bias. A 23-residue peptide of beta-beta-alpha structure can be predicted from any random starting conformation.

Key Words: global optimization; protein folding; Monte Carlo; structure prediction.

INTRODUCTION

Ab initio prediction of three-dimensional structures of large macromolecules remains the main theoretical and computational challenge in biology. Most of the globular proteins adopt a unique conformation in aqueous solution. It is believed that the compact and unique conformation of a protein corresponds to the global minimum of its free energy function, at least

* Corresponding author. E-mail: abagyan@earth.med.nyu.edu, Fax: 212-263-8951.

0021-9991/99 $30.00

for small, independently folding protein domains [10, 11]. Therefore, to predict the native folded conformations of a peptide or a small protein one needs to evaluate the free energy in the vicinity of every trial conformation with sufficiently high accuracy [8, 12] to separate the native minimum from billions of low energy alternatives. Furthermore, the native conformation must be *found* in a space of almost 100 dimensions within a reasonable time.

We will argue here that global optimization algorithms, not bound by trajectory continuity or by the canonical ensemble generation requirements of MD[1] or MC, represent the best approach for solving the protein folding problem. In addition, we will argue that first-order stochastic global optimization algorithms (which apply full local minimization after each random move) are superior to zero-order algorithms (which do not). Finally, we will introduce the optimal bias principle, derive the square-root bias, and demonstrate the speed and accuracy of the OBMCM algorithm in predicting the native conformation of large peptides with nontrivial topology.

MOLECULAR DYNAMICS: LARGER STEPS, BETTER SAMPLING

Traditional Newtonian molecular dynamics in Cartesian space remains an important algorithm for sampling the conformational space of peptides [13–15]. This algorithm, optimized and refined over the years, is best applied to simulations with explicit water molecules. Several attempts have been made to predict the native peptide conformation from a "denatured" state through dynamic simulations in water [13–15]; however, these attempts have had only limited success. The longest simulation to date is the 1-μs simulation of the villin headpiece domain which reached a metastable folded state [14].

The requirement for a small time step of integration (about 1 fs) imposes severe limitations on efficiency. Another factor increasing the computational effort is the accuracy of the energy function; this accuracy must be sufficient to distinguish the correct solution from billions of false alternatives, many of which may be very energetically similar to the correct answer.

An important alternative to dynamics in Cartesian space is dynamics in internal coordinate (torsional) space. The first application of torsion dynamics was limited to linear chains [16] and was based on Wittenburg's formalism for connected rigid bodies—which is related, in turn, to equations derived for n-body space satellites [17, 18]. The general equations for internal coordinate molecular dynamics of arbitrary fixed branched biomolecules were first introduced and tested on biomolecules in 1989–1991 [19–21]. Subsequently, two other implementations of torsion dynamics were proposed and applied to x-ray refinement and NMR-structure determination [22, 23] and peptide simulations [24]. This method allows us to easily distinguish between "hard" degrees of freedom, such as bond lengths and bond angles, and "soft" degrees of freedom such as torsion angles. Fixation of bond lengths and bond angles allows an increase of the time step of integration to 2–4 fs, and suppression of fast hydrogen rotations allows another several-fold increase [25] of the minimal time step. A number of interesting ideas have been proposed to increase the sampling power of MD simulations and permit larger time steps of integration [26–28]. In the most recent

[1] Abbreviations used: BMC, biased Monte Carlo [1]; ECEPP/3, empirical force field for polypeptides [2–4]; ICM, internal coordinate mechanics [5]; ICMD, internal coordinate molecular dynamics; MC, Monte Carlo; MCM, Monte Carlo-Minimization [6]; MD, molecular dynamics; OBMCM, BPMC, Optimal-Bias Monte Carlo-Minimization Bias (a.k.a. Biased Probability Monte Carlo-minimization), a quickly converging stochastic global optimization method using random moves derived from the expected local probability distributions [7].

implementation the update time of the "slow forces" was increased to 48 fs or more, which corresponds to a 10-fold increase in efficiency [26].

MONTE CARLO

In most cases, the native protein structure important for biological function can be considered one unique conformation in the mean field of the solvent, rather than a truly dynamic system. In other words, one can introduce a pseudo-free-energy potential that is a function of the peptide conformation. Therefore, a traditional Monte Carlo method—the primary goal of which is the generation of a Boltzmann ensemble—may be replaced by an algorithm aimed at the fastest possible identification of the lowest energy minimum. The latter is the primary goal of global optimization methods. A Monte Carlo method can also be designed to allow better sampling of the phase space [29], but the requirement of canonical ensemble generation defeats the purpose of efficient global optimization. An easy example: for a single harmonic well, a Monte Carlo method will need to calculate at least several dozen values until the convergence criterion is met, while a local minimization method can *find* the potential energy minimum in a single step. In the case of a rugged energy landscape the goals and the strategies are still different, since instead of sampling all the low energy patches of the hyper-surface, the global optimization is only aimed at finding the shortest path to the global minimum.

As we have noted, the native conformation is largely unique with the exception of some flexible surface side chains and surrounding water molecules. Therefore, if the free energy of solvent can be calculated implicitly for every trial conformation based on the electrostatic and surface effects, the free energy of a folded and solvated conformation can be well approximated by a function of a single conformation! An additional assumption is that the vibrational entropy differences between possible folded conformations are not large. If an explicit water model is used, and/or the peptide conformation is essentially dynamic (e.g., no particular backbone conformation dominates the statistical sum), conventional Monte Carlo or molecular dynamics methods become more appropriate.

Which approximation of the solvent in a peptide simulation, explicit or implicit, is more accurate? Clearly the implicit method is faster, but is it less accurate? The answer to this question is still not clear. On one hand, explicit water can potentially take the effects of finite molecular size into account (e.g., form a bridge of a certain length between hydrogen bonding groups). On the other hand, the problems and inaccuracies they introduce can outweigh the benefit. Let us list them: a large increase of the system size (thousands of water molecules); inaccuracies due to insufficient sampling; truncation of the electrostatic interactions or artificial boundary conditions; and limitations of the simple electrostatic and polarization model of a single molecule. The continuous electrostatic models [30, 31], however, are free from the above problems. They automatically consider an infinite shell of solvent and its electrostatic properties to be directly described by the experimentally measured dielectric constant.

STOCHASTIC GLOBAL OPTIMIZATION, OBMCM

A global optimization can reach the minimum much faster than a Monte Carlo procedure simply because the detailed balance condition, compulsory for the Monte Carlo procedure but unimportant for global optimization, is dropped. Therefore, an efficient local energy

optimization, the fastest way to identify the local minimum, can be used. Even the most sophisticated molecular dynamics and Monte Carlo algorithms could still only be applied to sample either the neighborhood of known protein conformations or a complete conformational space of rather small peptides. Typically, when using the above methods one had problems with predicting a new peptide or loop conformation from scratch without knowing the answer. Our goal, however, is to predict the lowest energy conformations of large peptides and protein loops from scratch without any prior structural information.

A number of stochastic global optimization methods have been developed [32]. They can be divided into methods with (first-order) and without (zero-order) local minimization after each step. These methods may be further subdivided according to the way in which random moves are made, their temperature scheduling, and their use of simulation history. The sampling bias, aimed at faster convergence, can be done according to the experimentally observed preferences [1, 7, 33] defined either as a continuous function [7] or as a discrete grid function [1, 33]. Another productive idea aimed at improved sampling is to make local backbone moves [34, 35], or restrict the random moves according to the kinetic escape time estimate (the so-called diffusion process-controlled Monte Carlo method [36]). Several excellent reviews of other zero-order Monte Carlo methods, which are not specifically aimed at the global optimization of a pseudo-free-energy potential, were published recently [29, 37].

Generally all the stochastic methods with minimization outperform the methods without local minimization (similarly the efficiency of a local minimization method critically depends on the use of energy derivatives). Indeed, the introduction of the Monte Carlo-Minimization (MCM) method [6] was a major step forward. With MCM it became possible to identify the global minimum of Met-enkephalin in fewer than 100,000 energy evaluations. However, larger peptides—especially in a non-alpha-helical conformation—still required a prohibitively large number of energy evaluations.

The next radical step forward was extension of the MCM method with the principle of optimal bias (OBMCM, a.k.a. BPMC). The goal of the optimal bias is to use *a priori* information about local probability distributions in the best possible way in order to minimize the number of random steps to the global minimum. For macromolecules the algorithm relies on the assumption that the local energy landscape due to local sequential interactions in a peptide is perturbed in a quasi-random way by non-local interactions that are comparable in magnitude. This assumption leads us to the formulation of a simple stochastic model and the square-root bias rule [38] (see the next section).

BENCHMARKS

To optimize and test new energy functions and global optimization methods, we rely on experimentally characterized peptides with compact unique topologies supported without disulfide bonds and metals. Until recently there were not many suitable candidates. Most small peptides remain unfolded in aqueous solution; among those peptides which do assume a specific conformation, the majority adopt a simple α-helical conformation [39]. Fortunately, several beta-forming peptides and the smallest compact mixed-topology peptide [40] (BBA1) have been discovered. The 23-residue BBA1 peptide contains the synthetic residue 3-(1,10-phenanthrol-2-yl)-L-alanine (Fen), and was later replaced by another 23-residue peptide (BBA5) which contains no synthetic amino acids and forms a stable beta-beta-alpha structure [9]. BBA5 became the smallest protein-like folded peptide and provides the ideal benchmark for the evaluation of theoretical folding algorithms.

The geometrical model used in all calculations is the extended internal coordinate model (ICM) [5]. This model is convenient for tree-like branched polymers and may be applied to any number of arbitrarily interacting molecules. A typical ICM system will have all covalent bonds lengths and bond angles fixed; torsion angles that are not involved in rigid rings are free. Analytical equations of motion have been derived for an arbitrarily constrained ICM tree, together with equations which permit the efficient calculation of pair-wise energy terms with respect to free internal coordinates [5, 19].

The required accuracy of the energy function was estimated as less than 1 kcal/mol per residue [8]. To achieve this accuracy we have previously used full-atom models, the ECEPP/3 vacuum force field [2–4], improved electrostatic solvation [7], and solvent-accessible surface-dependent estimates of side chain entropies [7]. This function worked well in simulations of 12 and 16 residue helices and with the BBA1 peptide, but we noticed in a number of BBA5 and beta hairpin simulations that the function is biased toward alpha helices.

To find the global energy minimum of the target peptides, we used a Monte Carlo minimization-based global optimization algorithm [6] in conjunction with a special set of biased random moves. In previous studies that employed Monte Carlo-based algorithms, investigators have focused on simulation temperature schedules [41] and the acceptance criteria (e.g., [42]). Several years ago we first suggested that the rational design of random moves is the key to a radical increase of the sampling efficiency [7]. A strategy was outlined for dividing the internal coordinates into groups of strongly coupled variables (zones), and it was proposed that the random moves could be biased according to a pre-calculated continuous local probability distribution. In addition, we derived the statistically optimal algorithm to bias the random move within the zone.

Here we discuss different aspects of an efficient stochastic global optimization technique for large molecular systems and present theoretical *ab initio* folding of the detailed atomic model of the BBA5 peptide with the BPMC procedure.

INTERNAL COORDINATES

When we predict the conformation of a large flexible molecule, we must do it in the right space. The choice is the following: Cartesian coordinates, torsion angles, a full set of internal coordinates, and the inter-atomic distances. We will focus on the third choice, which is ideally suited for imposing the covalent structure constraints and generating natural conformational rearrangements for multi-molecular arbitrarily fixed systems.

The general scheme of the ICM tree is shown in Fig. 1. In this model atoms are constructed sequentially from the origin as a directed graph, while the preceding bond length, a bond angle, and a dihedral angle define the geometrical position of each node. This dihedral angle can be of two types at a branching point. The dihedral angle defining the "main branch" will be referred to as a torsion angle and is usually free, while the difference between the above torsion angle and the dihedral angle defining the secondary branch will be referred to as a phase angle and is usually constrained. This choice of independent internal coordinates makes setting of chemical constraints trivial.

Accordingly, the BBA5 molecule (chemical structure Acet-Tyr-Arg-Val-DPro-Ser-Tyr-Asp-Phe-Ser-Arg-Ser-Asp-Glu-Leu-Ala-Lys-Leu-Leu-Arg-Gln-His-Ala-Gly-COOH) is represented by a directed graph in which all bond lengths, bond angles, and phase angles are constrained. The number of free torsion angles is 129. This number includes the peptide plane ω angles that are restrained by the torsion potential to 180° but are free to change.

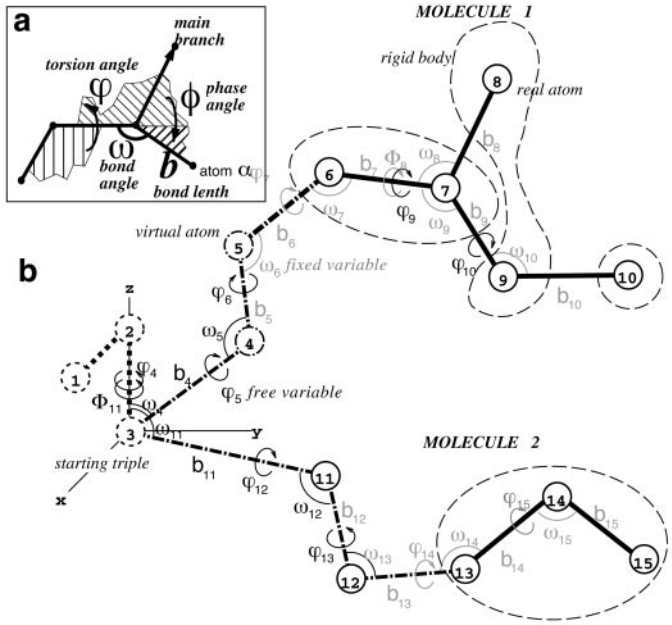

FIG. 1. An internal coordinate representation of one or several molecules in which any subset of bond lengths, as well as bond, phase, and torsion angles which determine the directed tree-like graph, can be constrained. The graph can contain virtual atoms and bonds. The first six internal coordinates determine the rotation and translation of the whole molecule. Analytical derivatives of a pair-wise energy function with respect to the four types of variables for such an object are given in [5, 19, 20] and the equations of motion are given in [19–21, 25].

The number of essential backbone torsion angles is 45. The model contains 385 atoms and includes hydrogen atoms.

Evenly distributed random values are assigned to all the free torsion angles, except the ω angles, to generate the initial conformation for each simulation.

ADDITIONS TO THE ECEPP/3 POTENTIALS AND ELIMINATION OF THE α-HELICAL BIAS

The BBA5 23-residue peptide has about 70 essential degrees of freedom; yet it folds into a unique β-β-α arrangement. Previously, we argued [8, 12] that a highly accurate free energy evaluation ($<$1 kcal/mol) for each trial conformation is necessary to recognize the native state among zillions of alternatives. To meet this requirement we introduced corrections in the all-atom vacuum force field ECEPP3 [2–35], and appended to it the solvation free energy [7] and the entropic contribution [7],

$$E = E_{\text{vw}} + E_{\text{hbonds}} + E_{\text{torsions}} + E_{\text{electr}} + E_{\text{solv}} + E_{\text{entropy}}.$$

The individual terms are calculated as follows.

$$E_{\text{vw}} = \sum_{i,j} \left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{CB_{ij}}{d_{ij}^{6}} \right),$$

$$E_{\text{hbonds}} = \sum_{i,j} \left(\frac{A'_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{10}} \right),$$

where i, j are a pair of atoms separated by less than 7.5 Å and more than two chemical bonds; d_{ij} is an inter-atomic distance; A, B, A', and D are parameters for each two atom types; and C is 0.5 for a pair separated by three bonds and 1 otherwise.

$$E_{\text{torsion}} = \sum_{\text{torsions}} U_m \cos(n\phi - \phi_0),$$

$$E_{\text{electr}} = \sum_{i,j} \frac{332 q_i q_j}{\varepsilon d_{ij}},$$

where q_i are electric charges, distance-dependent dielectric constant $\varepsilon = 4 \cdot d_{ij}$, and U_m are individual torsion barriers. The solvation contribution in this simulation was calculated as a sum of products of solvent-accessible surface areas (a_i) by the solvation energy densities (σ_i) derived from the water–vacuum transfer energies [8],

$$E_{\text{solv}} = \sum_i \sigma_i a_i.$$

Finally, the entropic contribution from the protein side chains is calculated from the maximal burial entropies for each residue type and their relative accessibilities,

$$E_{\text{entropy}} = -RT \sum_{\text{residue } \rho} \Delta S_\rho^{\max} a_\rho / a_\rho^{\max}.$$

A number of simulations with peptides of different topologies convinced us that the current form of the ECEPP3 force field has a bias toward alpha helices. To compensate for this bias we imposed a soft torsion potential on the backbone ψ angle: $0.5(1 + \cos(90 + \psi))$. A stronger torsion potential of 1 kcal/mol amplitude was applied to Val, Ile, and Thr residues. The corrected potential was used for simulations with all peptides, including the α-helical ones. Without the correction the lowest energy conformation of most of the peptides that we simulated is dominated by one α-helical element.

DERIVATION OF THE OPTIMAL BIAS

Several years ago we introduced the idea of the optimal bias in a Monte Carlo-based stochastic global optimization procedure. This procedure increased the efficiency of the peptide structure predictions by at least an order of magnitude [7].

Let us pose a mathematical problem of the optimal bias in random sampling specifically directed at finding a single right answer, and give a solution in three guessing games. The underlying model we have in mind is a chain molecule described as a set of torsion angles x_i, where each x_i is either a scalar representing an individual angle or a vector representing several correlated angles (as ψ-ϕ backbone angles or a group of side chain angles). One can see, however, that the consideration below is general and is applicable to a wide class of global optimization problems.

Game 1 (always bet on the best). Suppose that we are guessing values of only one vector x which is distributed according to $S(x)$. How should we guess to maximize the probability of the correct guess? In a continuous case this is equivalent to finding a function $f(x)$ maximizing the integral $\langle P \rangle$,

$$\langle P \rangle = \int S(x) f(x)\, dx,$$

under the normalizing condition, $\int f(x)\, dx = 1$.

The answer is trivial and uninteresting. Actually we should always guess the same most likely value $x^{S_{\max}}$ ($f(x) = \delta(x - x^{S_{\max}})$). This sampling strategy is obviously not very productive, since it does not really correspond to what we expect from a global optimization procedure, which is to predict all the variables correctly. In this game we will always be guessing the same value and even though the probability of the correct guess in each case is maximal, we will never be successful with any value other than the most probable.

Game 2 (linear bias). In the second guessing game we are trying to find the best guessing probability function $f(x)$ to achieve *simultaneous* prediction of n x-values with certain accuracy $h(x, x^0)$, given the fact that x_i^0-values are distributed according to a known distribution $S(x)$.

It has been shown [6] that the guessing (or sampling) probability function maximizing the product of n integrals,

$$P = \prod_{i=1}^{n} \int h(x, x_i^0)\, f(x)\, dx,$$

is actually identical to the expected probability distribution, $f(x) = S(x)$.

Game 3 (square root bias). Here we are actually going to guess many times until we get the correct answer and we assume that it can be done independently for each x. It can be proven that the function $f(x)$ minimizing the average number of guesses required to find the correct answer is a square root of the distribution function $S(x)$. $f(x) = C\sqrt{S(x)}$ ($f(x)$ normalized to 1 after the root is taken).

Let us derive the optimal sampling strategy for the simplest case of two states (see [31] for a complete description). To derive the rule, we must formulate a different criterion of optimality (the objective function).

The guessing problem formulated in this way can be understood even in the case of two fixed choices. Let us imagine the following game: each time you are given one of two choices and you guess until you are right, but every time you *forget* your previous choice. (Independence of the previous selection is actually well justified in the real case in which the environment of each residue changes after each step/guess.) After the problem is solved and the number of your guesses is recorded you are given another problem, and so on, and the two types of choices are offered with frequencies s_1 and s_2, respectively ($s_1 + s_2 = 1$).

THEOREM. *The optimal random guessing strategy minimizing the average number of guesses is to guess with relative frequencies f_1 and f_2 so that*

$$\frac{f_1}{f_2} = \sqrt{\frac{s_1}{s_2}}.$$

Proof. Let us calculate the average number of guesses N_{guesses} until the correct answer is given, provided that previous guesses are forgotten and thus each guess is independent. If in our random guessing strategy the probability of a correct guess in a single trial is f, and the successful result can be achieved through the first correct guess ($p_1 = f$), the first incorrect guess and the second correct guess ($p_2 = (1 - f) f$), the first two incorrect guesses and the correct guess ($p_3 = (1 - f)^2 f$), etc., the average number of guesses reads

$$\begin{aligned} N_{\text{guesses}} &= p_1 + 2p_2 + 3p_3 + \cdots \\ &= f + 2(1 - f) f + 3(1 - f)^2 f + \cdots + n(1 - f)^{n-1} + \cdots = 1/f, \end{aligned}$$

since each member of this series divided by f is a derivative of x^n, where $x = (1 - f)$, and the series $1 + x + x^2 + x^3 + \cdots + x^n$ converges to $1/(1 - x)$ as n tends to infinity. After taking the derivative and multiplying it by f we arrive at $N_{\text{guesses}} = 1/f$.

Now, in the case of two choices, we will guess the first choice with unknown probability f_1 and the second choice with $f_2 = 1 - f_1$. The average number of guesses reads

$$\langle N_{\text{guesses}} \rangle = S_1/f_1 + S_2/f_2.$$

To derive the optimal f_1 and f_2, let us set the derivative of $\langle N_{\text{guesses}} \rangle$ with respect to f_1 to 0:

$$S_1/f_1^2 - S_2/f_2^2 = 0,$$

or

$$S_1/f_1^2 = S_2/f_2^2,$$

or

$$f_1/f_2 = (S_1/S_2)^{1/2}.$$

This sampling strategy is also valid for multiple discrete states as well as for a set of n continuous variables [31].

DESIGN OF RANDOM MOVES

In the previous section we concluded that by biasing the random steps according to the expected local probability distribution we improve the efficiency of the stochastic global optimization procedure. Since we are optimizing tree-like branched polymers described geometrically by a set of internal coordinates, the next question is how to divide all the internal variables into groups (further referred to as zones). We will assign an expected probability distribution to each zone according to the structural preferences in the database of several thousand known protein structures or according to a short calculation on the fragments of a given peptide. The simplest choice is to change one angle at a time and use a square root of an average distribution of this type of angle, as found in numerous known protein structures or in a separate calculation, to determine its sampling probability. However, by choosing zones of several correlated angles (for example, the backbone ψ and ϕ angles) we can further improve the accuracy of the expected probability distribution and therefore improve the sampling efficiency [7].

Previously we implemented three types of random moves that allowed us to consider three types of optimization problems such as:

- *ab initio* peptide folding (zone move: global change of an individual group of ψ-ϕ angles or the side chain χ-angles),
- loop prediction in homology modeling and design (loop move: local loop rearrangements of two types; Fig. 2),
- molecular association (pseudo-Brownian move: incremental rotation and translation of the whole molecule; Fig. 2).

The zone and loop moves are intramolecular, while the pseudo-Brownian move is necessary for sampling of intermolecular arrangements. Let us introduce the essential properties of random moves. If the move is generated regardless of the current geometry, we will call it *static*, while the move taking into account the changing geometry will be called *dynamic*.

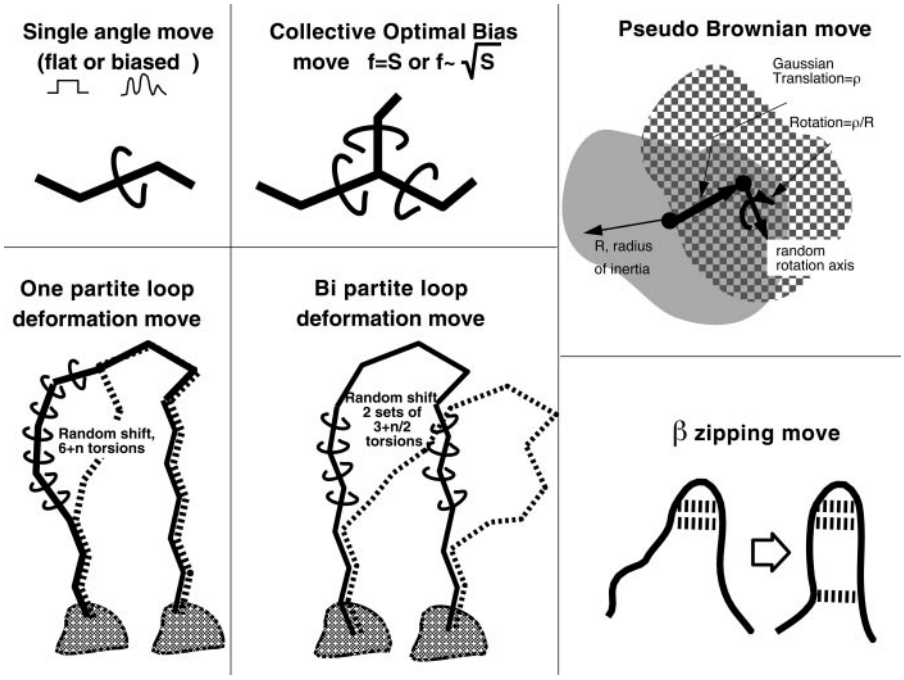

FIG. 2. The types of random moves used in ICM.

An example of the static random move is changing a randomly chosen torsion with an amplitude of 90° [43]. An example of a dynamic move is changing torsions with probabilities depending on the current secondary structure. It is also important to discriminate between *local* and *distributed* moves (changing one local group of variables or two or more groups at a time, respectively), and *correlated* and *uncorrelated* moves (angles in one or several groups are changed in concert or independently, respectively). Finally, random moves may be either *discrete* or *continuous*.

While the zone and loop moves, which are correlated and continuous, greatly improve the rate of convergence for peptide and loop simulations [44], they do not optimally cover the whole range of essential molecular rearrangements, since they change only one local group of variables at a time. In the first class the group consists of several adjacent unconstrained torsion angles in the tree, while in the second case a larger, but still contiguous, chain fragment is deformed under the constraint that the loop ends do not change their position. These moves therefore are local and static, since they only change conformation of a local contiguous fragment and do not depend on the current conformation.

The local correlated moves are efficient in folding α-helices and short compact loops. However, beta sheets or long loops which require simultaneous movement of the two non-adjacent chain fragments cannot be efficiently sampled without correlated and distributed moves. Two types of such moves are described here: a bipartite loop move and a beta-zipping move (Fig. 2).

In the one-partite loop deformation move n adjacent angles ($n > 6$, $n \leq N_{\text{loop torsions}}$) are chosen and a random biased move is performed according to the residue-dependent ψ-ϕ probability distributions, followed by a loop closure procedure. In the bipartite loop move, the n ψ-ϕ angles are divided into two groups separated by a random number of intervening

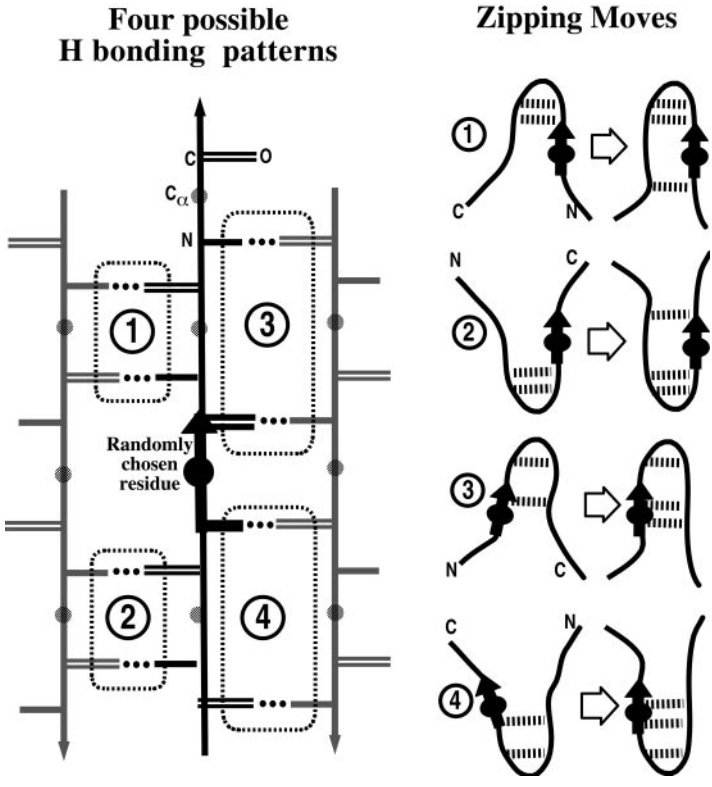

FIG. 3. Four types of the β-zipping move. If one of the patterns shown is identified, a temporary distance restraint is imposed to extend the β-hairpin.

angles and then the biased change and the closure procedures are performed. To facilitate formation of beta sheets we applied another type of random move, a beta-zipping move. After the random initial choice of the residue of interest, the hydrogen bonding neighbors are identified. The four situations of interest are shown in Fig. 3. If such a neighbor exists, a geometrical rearrangement of both strands is generated to form the missing hydrogen bond and extend the beta structure. This move is distributed since both strands are changed at once and correlated since the several ψ-ϕ angles are changed at once.

TEMPERATURE, ACCEPTANCE RATIO, AND DISADVANTAGES OF SIMULATING ANNEALING

The Metropolis acceptance criterion states that a new trial conformation with higher energy will be accepted with the probability of $\exp(-\Delta E / RT_{\text{sim}})$, where T_{sim} is the simulation temperature and ΔE is the energy increase. During global optimization the simulation temperature can be independently tuned for the best performance. The meaning of this parameter in stochastic global optimization is the energy accuracy required to recognize the global minimum. The Metropolis selection procedure is likely to reject the energy rise of RT_{sim} and, therefore, it "insists" that a new conformation in this energetic vicinity be generated and evaluated. Simulation at a higher temperature samples more widely and has a larger acceptance ratio, but spends less time in each location, while simulation at a lower temperature does the opposite.

The ultimate test of the MC global optimizers is their ability to identify as many as possible low energy minima in a minimal number of energy evaluations starting from a random conformation [45]. The optimal simulation temperature RT_{sim} can be found for a given representative benchmark [45]. In [45], a complete map of low energy states was built and then the number of local minima in a 20 kcal/mol range from the global minimum was counted as a measure of global optimization efficiency. The optimal simulation temperature was found to be about 600 K (RT_{sim} is about 1.2 kcal/mol). Both lower temperatures and higher temperatures are less efficient in identification of the energy minima.

In simulated annealing the temperature is gradually reduced according to a predetermined schedule. The high temperature part of the simulation allows crossing of large energy barriers and therefore samples broadly but without rigorous refinement in each vicinity. Conversely, the low temperature part of the simulation constrains the molecule to the vicinity identified by the high temperature calculation, thus allowing determination of its exact conformation and its energy. However, this scheme is extremely vulnerable. First, it depends on preexisting knowledge regarding the exact *time* required to identify the global minimum. If we underestimate the time and cool the system off too quickly, the molecule becomes prematurely frozen in an unrelated conformation, while if the cooling schedule is too slow the simulation is conducted at an inefficient high temperature, which impedes its ability to identify deep local energy minima. Second, simulated annealing relies on the assumption that the high temperature coarse-grained calculation in its walk through different low energy valleys will end up in a valley containing the lowest energy minimum, which is not clear until the lower temperature refinement is done. The global optimization procedure at a moderately elevated but constant temperature (e.g., 600 K) gives an equal chance to each valley visited and, especially with a proper history-feedback mechanism, does not suffer from this impediment.

HISTORY-FEEDBACK MECHANISMS

Both the full simulation history and its recent history can be used to enhance the global optimization efficiency. We will consider three undesirable situations that are commonly encountered during optimization, and will describe actions that improve the performance.

1. Too many unproductive trials. A frequent "recent-history" problem is inability of the procedure to make a move, i.e., find an acceptable new conformation at a given temperature. Although a certain degree of persistence is necessary, it may be beneficial to provide an "escape mechanism" to avoid excessive searches in an unproductive region of conformational space [45]. By counting the number of sequentially rejected moves, one can determine whether the search procedure has stalled and take appropriate action. Several escape mechanisms can be envisioned. We found that a temporary temperature raise after the achievement of the upper limit of unaccepted trials (N_1) improves the performance. The temperature is doubled after N_1 unproductive steps and is reset to normal after a new conformation is accepted.

Before describing the other two scenarios, we need to introduce the concept of a "conformational stack."

Conformational stack of the best representative conformations. In any sampling procedure going through millions of conformations it is useful to accumulate a condensed set of the best representative conformations. Such a conformational stack, first described in

[45], is useful for visualization and analysis, but, more importantly, it can also provide a useful feedback to the search procedure in a difficult situation. To build the stack, we need to define the metric in the conformational space. The difference between each two conformations must be quantitatively expressed by a relevant distance function. We use three types of comparison depending on a particular modeling task: (1) the root-mean-square deviation (rmsd) of the essential torsion angles (appropriate in peptide folding calculations); (2) Cartesian coordinate rmsd of essential atoms (appropriate in protein loop simulations or docking against a static receptor); and (3) Cartesian rmsd upon optimal superposition of the essential atoms (appropriate in small molecule sampling or in peptide simulations).

The conformational stack, which retains only the best energy conformation within a given conformational vicinity, evolves gradually during the simulation. Each new conformation is compared with the previously stored stack conformations; if it is not in the vicinity of any stack conformation, the new conformation is added to the stack. If it is similar to an existing stack conformation, however, it must compete with the existing conformation and will replace it only if it possesses a lower energy. It is indeed like biological evolution in which fitness is determined by the energy and only the fittest conformation survives each location. If the vicinity radius is too small the number of stack conformations will be exceedingly large. Limitation of the number of slots will result in only one family's dominating the stack. If the radius is too large, one conformation will absorb all the others and the diversity will not be visible. In peptide simulation we use the radius of 30°.

2. Too many visits to the same vicinity. A search stuck in a small region of conformational space stays in the vicinity of the same stack conformation but does not improve it. In other words, new conformations are found and accepted (as opposed to the previous case), but these conformations are around a known conformation and are energetically inferior to it. The conformational stack counts the number of fruitless visits to the same slot and resets this number to zero if a better energy conformation is found to replace the previously stored one. A good remedy in this case is to force the escape from this over-visited vicinity by randomizing free torsion angles. We use the amplitude of 30°.

3. Inability to improve the stack. The stack should be kept at a relatively small size during the simulation (around a hundred slots), since a comparison with each stack member must be performed for every new conformation that passes the Metropolis criterion. Therefore, the evolution of conformational species will take place within the limited number of slots. Every conformation in a new geometrical vicinity has a chance to replace the highest energy stack conformation, if the new energy is lower. If it does not happen and the search wanders in some high energy areas, not generating any changes in the stack (neither new slots, nor improvements of existing slots, nor increase of the number of visits in the existing stack conformation), the "high energy walk" is interrupted. We established the optimal limit of 50 high energy steps. Once the limit is exceeded the procedure returns to one of the under-visited conformations from the stack.

THE SETUP OF THE OBMCM OPTIMIZATION PROCEDURE

Each simulation starts from a completely random conformation (i.e., random numbers between -180° and 180° are assigned to each variable torsion angle) and ends when the number of function evaluations reaches its limit or the termination criterion is met (see below). The energy optimization of a single peptide consists of the following procedures:

1. Biased random move in a randomly chosen group of angles (OB-moves and beta-zipping moves).
2. Local energy minimization of the whole molecule
3. Calculation of the solvation energy and entropy for the minimized conformation
4. Metropolis selection (temperature increase if the rejection limit is reached)
5. Comparison of the accepted conformation with the stack (possible history feedback)
6. Back to step 1.

We used the following parameters: simulation temperature of 600 K, a set of local amino acid-dependent probability distributions for the backbone ϕ-ψ angles and the side chain torsion angles as previously described [7]. Conformations were compared with the angular rms of ψ, ϕ angles and the vicinity threshold of 30° was applied. The simulation temperature was doubled upon 10 sequential rejections; the local randomization with 30° amplitude was applied after the same stack conformation was visited more than 40 times; and the current conformation was reset to the least visited stack conformation upon 50 unsuccessful trials to modify the stack. These parameters were found by trial and error in a large number of peptide simulations. The ECEPP/3 force field was used with the modifications described in the preceding section. Following local minimization, the atomic accessibility-based solvation energy was calculated (the atomic radii and solvation energy densities σ_i are described elsewhere [8]). The side chain entropy parameters $\Delta S^{\max}$ and $a^{\max}$ for each residue were taken from Ref. [6]. Simulations converged after about 5–7 days using a single SGI R10000 processor (250 MHz).

TERMINATION CRITERION

The OBMCM procedure for global optimization is still a stochastic procedure and, as we said earlier, one does not know in advance how long it takes to find the global minimum. This average convergence time, even at the same number of independent variables, depends strongly on the geometry of the native conformation. The interaction energy between molecular fragments may be sufficient to create rather strained and unlikely local conformations, which aggravate the search. Therefore, the best setup of the global optimization procedure is the one which does not rely heavily on the expected convergence time.

One way to achieve this setup is to start several parallel calculations which periodically save their conformational stacks. Since every calculation starts from a completely random configuration and the size of the conformational space is exceedingly large, close similarity, both in terms of the energy values and in terms of geometry, between the lowest energy conformations retained in two or more stacks is a good indicator of convergence. Typically we execute 5–10 runs for a given peptide; if about half of these reach the same lowest conformation, we assume that the global minimum has been reached.

COMPARISON WITH SOME OTHER MC METHODS

To compare the global optimization efficiencies one could estimate the average number of energy evaluations required to identify a native-like conformation with sufficiently low conformational energy. First, we compared four zero-order algorithms of random step generation: changing one randomly selected variable at each step with various amplitudes; changing two or more coupled variables with 180° amplitude; and MD-like random

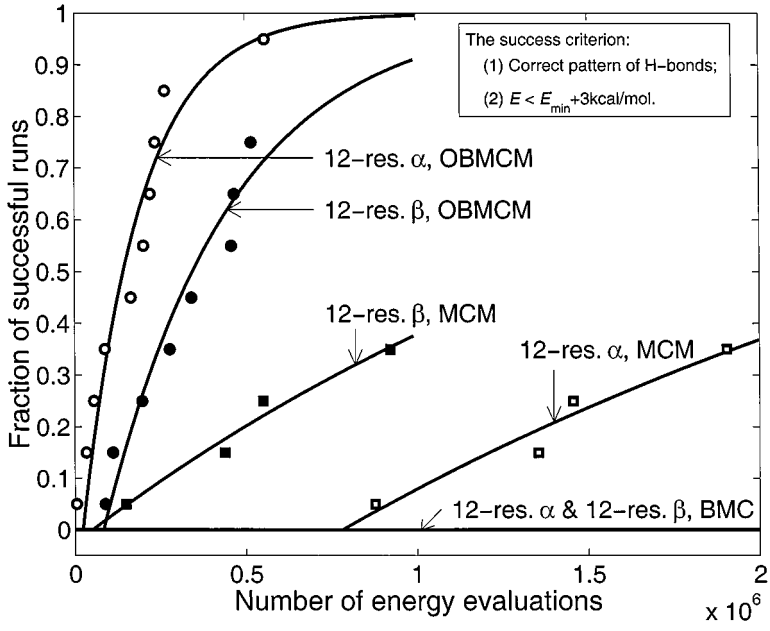

FIG. 4. The fraction of successful runs of OBMCM [7], BMC [1], and MCM [6] simulations as a function of the number of energy evaluations. The success is defined as identification of a conformation with the correct hydrogen bonding pattern and the energy gap from the global minimum energy $E_{\min}$ less than 3 kcal/mol. For the BMC simulations, no successful run was found within 4×10^6 energy evalutions.

movements of all variables with small amplitudes. Simulations of the α-helix and a β-hairpin peptide indicated that a zero-order optimal-bias MC algorithm yielded both larger rmsd's and larger acceptance ratios than all the unbiased categories (Table 2 in [38]).

To analyze the effect of the optimal bias on the simulation efficiency three representative MC global optimization methods were analyzed recently in [38]. In that work, the *ab initio* simulation of a 12-residue α-helix and a 12-residue β-hairpin were used to compare OBMCM, BMC zero-order optimization [1], and MCM first-order optimization [6]. For each method the results of 10 independent simulations, starting from a random set of dihedral angles, were averaged. The performance of OBMCM, measured as the time required to reach the 50% success rate, was one order of magnitude better than the performance of MCM and about two orders of magnitude better than that of BMC (Fig. 4).

We also compared the OBMCM performance with that of the diffusion process-controlled MC algorithm (DPCMC) [36]; however, the comparison was complicated by substantial differences in the molecular representation and the energy function between the two methods. We implemented the DPCMC three-step conformation generation scheme [36] and performed the simulation for the all-atom model of the 16-residue peptide [39]. The preliminary results using DPCMC (Fig. 5) did not show a drastic performance enhancement over the OBMCM step generation scheme; however, the inferior performance of DPCMC in our test may be due to the implementation details.

SIMULATION OF α, β, AND $\beta\beta\alpha$ PEPTIDES

After some efforts to improve the ECEPP3 energy function and the solvation model we were able to predict peptides with different secondary structures using identical procedures

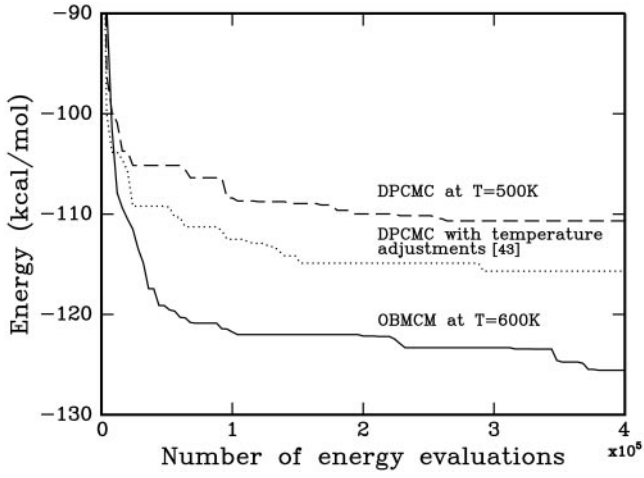

FIG. 5. The best energies achieved in the DPCMC and OBMCM simulations for a 16-residue α-helix peptide as a function of the number of energy evaluations. We performed the DPCMC simulations at 500 K with and without temperature adjustments proposed in [44]. The adjustments algorithm temporarily doubles the temperature if more than 10 trial conformations were rejected in a row. Each presented curve was the average of four independent simulations from random starting conformations.

and energy functions. The examples include 12- and 16-residue helical peptides [7, 38], 9- and 12-residue β-hairpins [8, 38], and the originally designed 23-residue $\beta\beta\alpha$ peptide [8] containing D-proline in the fourth position and a phenanthrol side chain in the sixth position [40]. Here we report a structure prediction of a new 23-residue $\beta\beta\alpha$ peptide, which contains only the standard amino acids [9].

Figure 6 shows a series of snapshots taken at different time points during one of the simulations. Each of the four simulations started from a different totally random conformation. Up to 13,000,000 energy evaluations were allowed. Three of four simulations found the lowest energy conformation with an accuracy of 3 kcal/mol. The fourth simulation was stuck at higher energy conformations. The temperature was doubled from 650 to 700 times (i.e., the consecutive rejection limit was reached, on average, once per 100 random moves) during each simulation. In all simulations we observed that the C-terminal helix found the lowest energy conformation earlier than the β-hairpin. The amount of helicity varies in the set of low energy conformations but most of the conformations have at least a part of the C-terminal helix. The short beta hairpin at the N-terminus is less pronounced and is present only in a fraction of the low energy conformations, including the lowest energy one. We found many low energy conformations in which the hairpin adopts different conformations or packs differently. The stabilization gap between the lowest energy $\beta\beta\alpha$ conformation, and the lowest energy conformation with a different topology in which both the β-hairpin and a turn of the C-terminal helix lose their secondary structures, is 6.8 kcal/mol.

Unfortunately, the experimental structure of the BBA5 peptide is not available for direct comparison. In the minimal energy conformation (i) the C-terminal helix spans residues from 12 to 20 (exactly the same range is observed in the experimental structure, the backbone rmsd with BBA1 is 0.4 Å, all-atom rmsd is 1.8 Å); (ii) the β-hairpin spans residues 2 to 7 (the same as the range in the experimental structure), and the rmsd with BBA1 values are 0.6 and 1.0 Å for the backbone and all atoms, respectively. However, the packing of

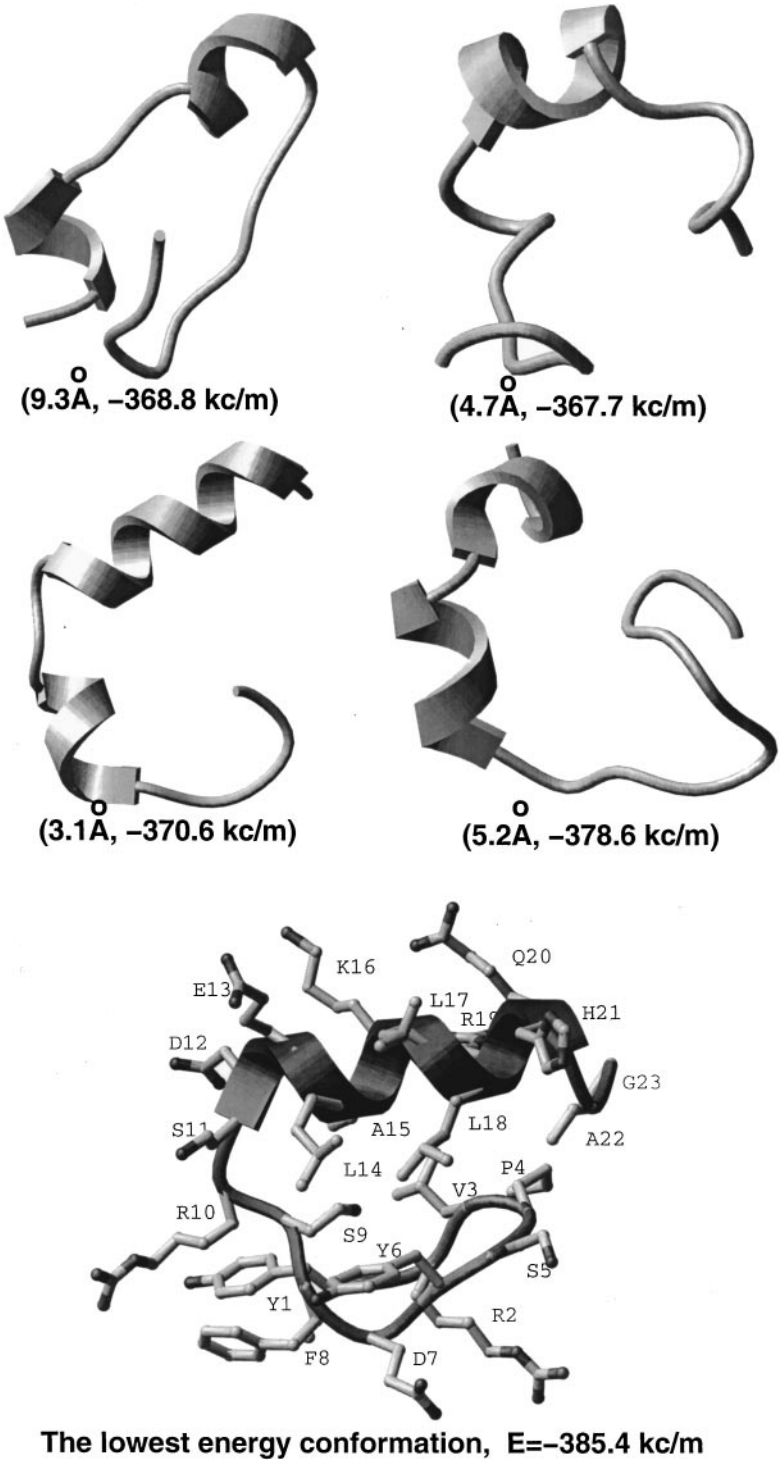

FIG. 6. A series of low energy conformations of the 23-residue peptide (BBA5), and the lowest energy $\beta\beta\alpha$-conformation. The rms deviation from the minimal energy conformation is shown (the experimental structure of BBA5 is not available).

the hairpin onto the helix in the predicted BBA5 conformation is shifted compared to the BBA1, with the global rmsd equal to 3.8 Å. Despite the packing problems, the observed similarity between the minimal energy conformation for BBA5 and the NMR structure of BBA1 [9] is encouraging.

CONCLUSIONS

The Optimal Bias Monte Carlo-minimization, featuring an improved energy function and an extended set of random moves, can identify the unique global minimum of a 23-residue peptide (containing 70 essential torsion angles and 385 atoms) after starting from a set of random torsion angle values. Two principal parts of this minimal structure correspond with high accuracy to the known experimental three-dimensional structure of the peptide; however, the packing of these parts differs from the experiment. To our knowledge, no other procedure is capable of finding this conformation from a truly random start. One calculation takes about 200 h on a single R10000 processor. An identical energy function and the simulation procedure also predict an α-helical and a β-hairpin peptide [8, 38]. A minor improvement of the energy function combined with a teraflop supercomputer might be sufficient for *ab initio* predictions of small proteins.

ACKNOWLEDGMENTS

We thank NIH (Grant R01 GM55418-01) and DOE (Grant DE-FG02-96ER62268) for financial support (this does not constitute an endorsement by DOE of the views expressed in the article). We thank Alex Morrill and Yingyao Zhou for help during preparation of the manuscript.

REFERENCES

1. B. Lee, N. Kurochkina, and H. S. Kang, Protein folding by a biased Monte Carlo procedure in the dihedral angle space, *FASEB J.* **10**, 119 (1996).
2. F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions and intrinsic torsional potentials for the naturally occurring amino acids, *J. Phys. Chem.* **79**, 2361 (1975).
3. G. Nemethy, M. S. Pottle, and H. A. Scheraga, Energy parameters in polypeptides. IX. Updating of geometric parameters, nonbonded interactions and hydrogen bond interactions for the naturally occurring amino acid, *J. Phys. Chem.* **87**, 1883 (1983).
4. G. Nemethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga, Energy parameters in polypeptides. X. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides, *J. Phys. Chem.* **96**, 6472 (1992).
5. R. A. Abagyan, M. M. Totrov, and D. A. Kuznetsov, ICM: An efficient technique for structure predictions and design, *J. Comput. Chem.* **235**, 983 (1994).
6. Z. Li and H. A. Scheraga, Monte Carlo-minimization approach to the multiple-minima problem in protein folding, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6611 (1987).
7. R. A. Abagyan and M. M. Totrov, Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins, *J. Mol. Biol.* **235**, 983 (1994).
8. R. A. Abagyan, Protein structure prediction by global energy optimization, in *Computer Simulation of Biomolecular Systems*, edited by W. F. van Gunsteren *et al.* (Kluwer/Escom, London, 1997), Vol. 3, p. 363.
9. M. D. Struthers, J. J. Ottesen, and B. Imperiali, Design and NMR analyses of compact, independent foled BBA motifs, *Folding & Design* **3**, 95 (1998).
10. C. B. Anfinsen, Principles that govern the folding of protein chains, *Science* **181**, 223 (1973).

11. C. B. Anfinsen and H. A. Scheraga, Experimental and theoretical aspects of protein folding, *Adv. Protein Chem.* **29**, 205 (1975).
12. R. A. Abagyan, Towards protein folding by global energy optimization, *FEBS Lett.* **325**, 17 (1993).
13. R. E. Bruccoleri and M. A. Karplus, Conformational sampling using high-temperature molecular dynamics, *Biopolymers* **29**, 1847 (1990).
14. Y. Duan and P. A. Kollman, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution, *Science* **282**, 740 (1998).
15. M. Prevost and I. Ortmans, Refolding simulations of an isolated fragment of barnase into a native-like beta hairpin: Evidence for compactness and hydrogen bonding as concurrent stabilizing factors, *Proteins* **29**, 212 (1997).
16. M. R. Pear and J. H. Weiner, Brownian dynamics study of a polymer chain of linked rigid bodies, *J. Chem. Phys.* **71**, 212 (1979).
17. J. Wittenburg, *Dynamics of Systems of Rigid Bodies* (Wiley, New York, 1997).
18. W. W. Hooker and C. Margulies, *J. Astronaut. Sci.* **12**, 123 (1965).
19. A. M. Mazur and R. A. Abagyan, New methodology for computer-aided modelling of biomolecular structure and dynamics. 1. Non-cyclic structures, *J. Biomol. Struct. Dyn.* **6**, 815 (1989).
20. R. A. Abagyan and A. M. Mazur, New methodology for computer-aided modelling of biomolecular structure and dynamics. 2. Local deformations and cycles, *J. Biomol. Struct. Dyn.* **6**, 833 (1989).
21. A. K. Mazur, V. E. Dorofeev, and R. A. Abagyan, Derivation and testing of explicit equations of motion for polymers described by internal coordinates, *J. Comput. Phys.* **92**, 261 (1991).
22. L. M. Rice and A. T. Brunger, Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement, *Proteins* **19**, 277 (1994).
23. E. G. Stein, L. M. Rice, and A. T. Brunger, Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation, *J. Magn. Reson.* **124**, 154 (1997).
24. G. Kneller, Molecular dynamics simulation of 16-polyalanine in internal coordinates, in *Proceeding Am. Chem. Soc. Meeting*, Boston, MA (1996).
25. A. K. Mazur, Quasi-Hamiltonian equations of motion for internal coordinate molecular dynamics of polymers, *J. Comput. Chem.* **18**, 1354 (1997).
26. E. Barth and T. Schlick, Overcoming stability limitations in biomolecular dynamics. I. Combing force splitting via extrapolation with Langevin dynamics in LN, *J. Chem. Phys.* **109**, 1617 (1998).
27. T. Schlick, E. Barth, and M. Mandzink, Biomolecular dynamics at long timesteps: Bridging the timescale gap between simulation and experimentation, *Annu. Rev. Biophys. Biomol. Struct.* **16**, 179 (1997).
28. M. E. Tuckerman, G. J. Martyna, and B. J. Berne, Reversible multiple time scale molecular dynamics, *J. Chem. Phys.* **97**, 1990 (1992).
29. B. J. Berne and J. E. Straub, Novel methods of sampling phase space in the simulation of biological systems, *Curr. Opin. Struct. Biol.* **7**, 181 (1997).
30. B. Honig and A. Nicholls, Classical electrostatics in biology and chemistry, *Science* **268**, 1144 (1995).
31. A. H. Elcock, M. J. Potter, and J. A. McCammon, Application of Poisson–Boltzmann solvation forces to macromolecular simulations, in *Computer Simulation of Biomolecular Systems*, edited by W. F. van Gunsteren *et al.* (Kluwer/Escom, London, 1997). Vol. 3, p. 244.
32. D. Shalloway, in *Recent Advances in Global Optimization*, edited by C. A. Floudas and P. M. Pardalos (Princeton Univ. Press, Princeton, NJ, 1991), Vol. 1.
33. J. S. Evans, A. M. Mathiowetz, S. I. Chan, and W. A. Goddard 3rd, De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology, *Protein Sci.* **4**, 1203 (1995).
34. F. Avbelj and J. Moult, Determination of the conformation of folding initiation sites in proteins by computer simulation, *Proteins* **23**, 129 (1995).
35. D. Hoffman and E-W. Knapp, Polypeptide folding with off-lattice Monte Carlo dynamics: The method, *Eur. Biophys. J.* **24**, 387 (1996).
36. P. Derreumaux, A diffusion process-controlled Monte Carlo method for finding the global energy minimum of a polypeptide chain. I. Formulation and test on a hexadecapeptide, *J. Chem. Phys.* **106**, 5260 (1997).

37. H. A. Scheraga, Recent developments in the theory of protein folding: Searching for the global energy minimum, *Biophys. Chem.* **59**, 329 (1996).
38. Y. Zhou and R. Abagyan, Efficient stochastic global optimization for protein structure prediction, in *Rigidity Theory and Application*, edited by M. F. Thorpe and P. M. Duxbury (Plenum, New York, 1999), p. 345.
39. J. M. Scholtz, E. J. York, J. M. Stewart, and R. L. Baldwin, A water-soluble, α-helical peptide: The effect of ionic strength on the helix–coil equilibrium, *J. Am. Chem. Soc.* **113**, 5102 (1991).
40. M. D. Struthers, R. P. Cheng, and B. Imperiali, Design of a monomeric 23-residue polypeptide with defined tertiary structure, *Science* **271**, 342 (1996).
41. S. Kirkpatrick, C. D. Gellatt, and M. P. Vecchi, Optimization by simulated annealing, *Science* **220**, 671 (1983).
42. B. A. Berg, Locating global minimum in optimization problems by a random-cost approach, *Nature* **361**, 708 (1993).
43. S. R. Wilson and W. Cui, Applications of simulated annealing to peptides, *Biopolymers* **29**, 225 (1990).
44. T. V. Borchert, R. A. Abagyan, R. Jaenicke, and R. K. Wierenga, Design, creation, and characterization of a stable, monomeric triosephosphate isomerase, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 1515 (1994).
45. R. Abagyan and P. Argos, Optimal protocol and trajectory visualization for conformational searches of peptides and proteins, *J. Mol. Biol.* **225**, 519 (1992).