# An Evolutionary Strategy for All-Atom Folding of the 60-Amino-Acid Bacterial Ribosomal Protein L20

A. Schug and W. Wenzel
Forschungszentrum Karlsruhe, Institut für Nanotechnologie, 76021 Karlsruhe, Germany

ABSTRACT We have investigated an evolutionary algorithm for de novo all-atom folding of the bacterial ribosomal protein L20. We report results of two simulations that converge to near-native conformations of this 60-amino-acid, four-helix protein. We observe a steady increase of "native content" in both simulated ensembles and a large number of near-native conformations in their final populations. We argue that these structures represent a significant fraction of the low-energy metastable conformations, which characterize the folding funnel of this protein. These data validate our all-atom free-energy force field PFF01 for tertiary structure prediction of a previously inaccessible structural family of proteins. We also compare folding simulations of the evolutionary algorithm with the basin-hopping technique for the Trp-cage protein. We find that the evolutionary algorithm generates a dynamic memory in the simulated population, which leads to faster overall convergence.

## INTRODUCTION

De novo protein structure prediction remains one of the outstanding challenges of biophysical chemistry. Much biomedical information would be gained if the presently available sequence information could be efficiently translated into three-dimensional structure (1). Although homology-based methods for protein structure prediction (2,3) have shown consistent progress (4), all-atom folding methods remain in their infancy. Recent studies for small proteins document both the feasibility and limitations of this approach (3,5–11), in particular regarding the simulation of the folding process (12,13).

Our approach to all-atom structure prediction is based on the thermodynamic hypothesis (14), which postulates that many proteins are in thermodynamic equilibrium with their environment. For these systems the native conformation corresponds to the global minimum of their free-energy landscape (15,16). Using the prevailing funnel paradigm (17,18) for protein folding, thermodynamically inspired optimization methods can locate the global optimum of the free-energy surface as the native conformation. In contrast to Levinthal's folding-path scenario (19), an optimization method need not follow the dynamics of the system, which makes this approach potentially much faster. We developed an all-atom protein force field (PFF01) (7,20,21) with an area-based implicit solvent model that approximates the free energy of peptide conformations under physiological conditions. Using this free-energy force field we were able to predict the tertiary structure of several two- and three-helix proteins: the 20-amino-acid Trp-cage protein (7,22–24), the 36 amino-acid villin headpiece (25), and the 40-amino-acid headgroup of the HIV accessory protein (9,26). Our method requires an accurate force field and an efficient stochastic optimization method to reliably locate the global optimum of the free-energy surface. Little is presently known about the increase of computational complexity with system size or the relative efficiency of different optimization strategies. Stochastic methods (27) map the search for the global optimum to a fictitious dynamical process that explores the free-energy surface with a bias toward low-energy conformations. The performance of methods with just one such dynamical process (7,20) is obviously limited by the speed of the energy/force evaluation for each proposed conformation. Efficient parallel methods could speed up the search process significantly, but their efficiency often saturates quickly with the number of dynamical processes (replicas) (26). Here we investigate a simple evolutionary strategy and demonstrate that it can overcome these limitations. Evolutionary strategies evolve an active population of many replicas. Their selection rules generate a dynamic memory of the overall process, which should speed up its convergence. We report two different folding simulations for the 60-amino-acid bacterial ribosomal protein L20 (28,29), both of which converge to low-resolution models of the native conformation. In the final population of these simulations, the energetically lowest conformation had approached the native state to 4.5 Å and 4.3 Å backbone root-mean-square deviation (RMSB), respectively (see Fig. 1). We find that six and five of the energetically lowest 10 conformations converge to near-native conformations within the constraints of the algorithm. The ''native content'' of the simulated ensemble, calculated as a suitably defined weighted average of the RMSB deviations of the population, increased as much as 60-fold during the simulations (see Fig. 2). These results demonstrate the feasibility of de novo protein structure prediction for a four-helix protein and transferability of our force field to this previously inaccessible structure class. As for the two- and three-helix proteins, we find that the entire low-energy landscape is dominated by conformations with nearly native secondary structure.

*Submitted July 11, 2005, and accepted for publication January 10, 2006.*

Address reprint requests to W. Wenzel, E-mail: wenzel@int.fzk.de.



0006-3495/06/06/4273/08 $2.00

doi: 10.1529/biophysj.105.070409

To rationalize the success of the evolutionary algorithm, we also performed folding simulations for the Trp-cage protein. We find that the dynamical memory of the active population speeds the convergence on average in comparison to the best simulation protocol previously available (24). These results motivate the use of evolutionary techniques for de novo folding studies of larger and more complex proteins in the future.

# METHODS

## Biophysical model

We applied the evolutionary optimization strategy to fold the 60-amino-acid bacterial ribosomal protein L20 (PDB code 1GYZ (29,30), sequence: WIARI NAAVR AYGLN YSTFI NGLKK AGIEL DRKIL ADMAV RDPQA FEQVV NKVKE ALQVQ), using PFF01 as the underlying biophysical model. PFF01 stabilizes the native structure of several small helical proteins against independently generated decoys (31). It represents all atoms individually (with the exception of hydrogen in $CH_n$ groups). The energy is parameterized as:

$$V(\{\theta_i\}) = V_{LJ} + V_C + V_{HB} + V_S, \quad (1)$$

using physically motivated contributions with an emphasis on the simplicity of their numerical evaluation. $V_{LJ}$ designates the Lennard-Jones interaction derived in a potential-of-mean-force approach by fitting the experimentally observed short-range (2–5 Å) radial distributions of a set of 138 proteins that are believed to span a wide range of possible folds (32). $V_C$ implements an established electrostatic parameterization for proteins (33) and $V_{HB}$ denotes a potential of mean force for hydrogen bonding. $V_S$ designates the solvent-accessible-surface-based implicit solvation model. The full parameterization of PFF01 was reported by Herges and Wenzel (21).

During the folding process, we consider only variations of the dihedral angles $\{\theta_i\}$ of the backbone and the side chains, keeping all other angles and bond-lengths fixed. In our simulation, we therefore propose only moves around the side-chain and backbone dihedral angles, which are attempted with 30% and 70% probability, respectively. The moves for the side-chain angles are drawn from an equidistributed interval with a maximal change of 5°. Half of the backbone moves are generated in the same fashion, and the remainder are generated from a move library that was designed to reflect the natural amino-acid-dependent bias toward the formation of $\alpha$-helices or $\beta$-sheets. The probability distribution of the move library was fitted to experimental probabilities observed in the PDB database (34). Although it drives the simulation toward the formation of secondary structure, the move library contains no bias toward helical or sheet structures beyond that encountered in nature.

## Optimization methods

The low-energy region of the free-energy landscape of proteins is extremely rugged because the packing of atoms in collapsed conformations is quite dense. Efficient optimization methods must speed up the simulation by avoiding high-energy transition states, adapt large-scale moves, or accept unphysical intermediates. The basin-hopping technique has proved to be a reliable workhorse for many complex optimization problems (35), including protein folding (9,36–38), but it employs only one dynamical process.

Here we generalize the basin-hopping method to a population of size $N$, which is iteratively improved by $P$ concurrent dynamical processes (we used $P = 50$–$100$). The whole population is guided toward the optimum of the free-energy surface with a simple evolutionary strategy. This strategy must balance energy improvement and diversity of the population. Conformations are drawn from the population and subjected to an annealing cycle. At the end of each cycle the resulting conformation is either integrated into the active population or discarded. Similar strategies, employing a conformation stack, were explored in simulations of the 23-amino-acid BBA5 protein (32,36).

This algorithm was implemented on a distributed master-client model in which idle clients request a task from the master. The master maintains the active conformations of the population and distributes the work to the clients. Each step in the evolutionary algorithm has three phases:

### *Selection*

In this step, a conformation is drawn randomly from the active population. We have used two different probability distributions for the simulations of bacterial ribosomal protein L20. Simulation A used a uniform distribution. In simulation B, the selection probability fell linearly with the energetic rank of the conformation, the energetically best conformation was $N$ times as likely to be chosen as the worst replica.

### *Annealing cycle*

We used a geometric cooling schedule with $T_{start} = 600$ K, $T_{end} = 2$ K. The number of steps per cycle increased as $10^5 \times \sqrt{N_{cycle}/P}$, where $N_{cycle}$ is the total number of cycles of all simulations (maintained by the master process). Toward the end of the simulation, a single annealing cycle comprised as much as $2.3 \times 10^6$ steps.

### *Population update*

The acceptance criterion for newly generated conformations must balance the diversity of the population against the enrichment with low-energy decoys.

The new conformation replaces a member of the active population in the following cases:

1. There are no similar active conformations and the new conformation has a lower energy than the energetically worst active conformation. In this case, the worst conformation is replaced by the new conformation.
2. The new conformation is similar to one or more active conformation(s) and its energy is lower than the energy of the closest of these similar conformations. In this case the new conformation replaces the closest (by RMSB) similar conformation.

New conformations that do not fit these selection rules are discarded. The decision tree for this process is shown in detail (see Fig. 6). Two conformations are considered similar if their mutual RMSB is below $R_C$. We used $R_C = 3$ Å. The selection rules insure the diversity of the population, whereas the replacement criteria insure that structurally different low-energy conformations are always accepted.

## Seed population

The evolutionary algorithm is best seeded with a wide variety of competitive starting conformations. We started with 100 random conformations obtained from short Monte Carlo simulations of the completely stretched ''stick'' conformation. These conformations had an average RMSB deviation of 12.2 Å from the experimental conformation. For each of these conformations, we performed high-temperature (500 K) Monte Carlo simulations of 50,000 steps until a total of 17,000 distinct decoys had been gathered.

We performed these simulations with a 20% reduced strength of the solvent interactions ($V_S$) to facilitate the rapid formation of secondary structure. It has been argued that hydrophobic collapse competes with secondary-structure formation in protein folding. In the collapsed conformational ensemble, large-scale conformational changes, such as those required for secondary-structure formation, occur only rarely.

The decoy set was ranked according to total energy as well as the individual energy terms ($V_S$, $V_{LJ}$, $V_{HB}$, and $V_C$ (side chain) and $V_C$ (backbone)). For each criterion, we selected the best 50 conformations and eliminated

duplicates to arrive at a population of 266 distinct conformations (by the criterion defined above), which seeded the population for the evolutionary algorithm.

### Performance measure: native score

To judge the performance of the algorithm it is important to note that it is not possible for the entire population to converge to the native conformation. We measured the progress of the simulation by monitoring the quality of the lowest conformation and the number and rank of near-native conformations (within a threshold 3 Å RMSB with respect to the native conformation). To quantify the latter we defined a native score $\sum 100(N-R+1)/N$ of the population. The sum runs over all near-native conformations in the population. $N$ is the size of the population and $R$ designates the rank of the conformation in the population. A score of 100 corresponds to a native decoy placed at the top position, whereas a near-native decoy at the bottom of the population contributes nothing to the score.

## RESULTS

### Folding the bacterial ribosomal protein L20

We believed that the evolutionary algorithm would perform best with a large and diverse population. We started a simulation for the seed population described above ($N = 266$). Configurations were drawn according to a uniform selection probability. In the course of this simulation, comprising 50 annealing cycles per replica, the native score rose only slowly (see Fig. 2, *lower*), indicating an overall slow convergence. In the limit $P = N$, the best conformations are only drawn relatively seldom. An equidistributed selection probability concentrates much effort on comparatively high-energy conformations.

To overcome this difficulty, we pruned the simulation to the best $N = 50$ decoys (by energy) and continued simulation A for another 5500 annealing cycles. Simulation B was started from the same subpopulation, but the computational effort was further biased toward the improvement of the ''best'' structures using a selection probability that fell linearly with the energetic rank of the conformation.

At the end of the simulations, the respective lowest-energy conformations had converged to 4.6 and 4.3 Å RMSB with respect to the native conformation. Simulation B had reached a slightly lower energy than simulation A. Table 1 demonstrates that of the 10 lowest structures in each simulation, 5 and 6, respectively, had independently converged to near-native conformations of the protein. The first nonnative decoy appears in position 2, with an energy deviation of only 1.8 kcal/mol (in our model) and a significant RMSB deviation.

The good agreement between the folded and the experimental structure is evident from Fig. 1, which shows the overlay of the native and folded conformations. The good alignment of the helices illustrates the importance of hydrophobic contacts to correct folding of this protein. An independent measure to assess the quality of these contacts is to compare the $C_\beta$-$C_\beta$ distances in the folded structure to those of the native structure. These correspond to the nuclear Overhauser effect constraints of the NMR experiments that determine tertiary structure. The $C_\beta$-$C_\beta$ distance matrices in Fig. 1 demonstrate ~60% (75%) coincidence of the $C_\beta$-$C_\beta$ distances to within 1 (1.5) standard deviation of the experimental resolution for both simulations. The dark diagonal blocks indicate intrahelical contacts. These are, perhaps not

**TABLE 1 RMSB and secondary structure of the 10 lowest energy decoys of the final populations of simulations A and B**

| Name | Energy | RMSB | Three-state secondary structure |
|---|---|---|---|
| 1GYZ | | | ccHHHHHHHcccccccHHHHHHHHHHHcccccccccHHHHHHcHHHHHHHHHHHHHHHcccc |
| A1 | −167.87 | 4.64 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccHHHHHcHHHHHHHHHHHHHHHcc |
| A2 | −166.15 | 8.25 | ccHHHHHHHHHHcccHHHHHHHHHHHcccccccccccHHHHHcHHHHHHHHHHHHHHHcc |
| A3 | −165.91 | 4.41 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccHHHcHHHHHHHHHHHHHHHcc |
| A4 | −164.11 | 5.54 | ccHHHHHHHHHHcccHHHHHHHHHHHcccccccccccHHHHHcHHHHHHHHHHHHHHHcc |
| A5 | −163.99 | 3.79 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccHHHHHcHHHHHHHHHHHHHHHcc |
| A6 | −163.93 | 4.04 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHcc |
| A7 | −163.45 | 8.52 | ccccHHHHHHHHcccHHHHHHHHHHHcccccccccccHHHHHcHHHHHHHHHHHHHHHcc |
| A8 | −163.20 | 4.37 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccHHHHHcHHHHHHHHHHHHHHHcc |
| A9 | −162.67 | 5.55 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccHHHHccHHHHHHHHHHHHHHHcc |
| A10 | −162.52 | 3.78 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHccc |
| 1GYZ | | | ccHHHHHHHcccccccHHHHHHHHHHHcccccccccHHHHHHcHHHHHHHHHHHHHHHcccc |
| B1 | −169.41 | 4.30 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHcc |
| B2 | −168.08 | 5.50 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHcc |
| B3 | −167.80 | 8.98 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHccc |
| B4 | −167.61 | 4.58 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHcc |
| B5 | −167.37 | 9.42 | ccccHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHccc |
| B6 | −167.33 | 4.29 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccHHHHcccccHHHHHHHHHHHHHHHcc |
| B7 | −167.02 | 9.20 | ccHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHccc |
| B8 | −167.00 | 3.93 | ccHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHcc |
| B9 | −166.80 | 9.25 | ccccHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHcc |
| B10 | −166.77 | 3.96 | cHHHHHHHHHHHcccHHHHHHHHHHHcccccccccccccccccHHHHHHHHHHHHHHHcc |

RMSB indicates the deviation from the experimental structure. The first row designates the secondary structure content of the experimental structure.

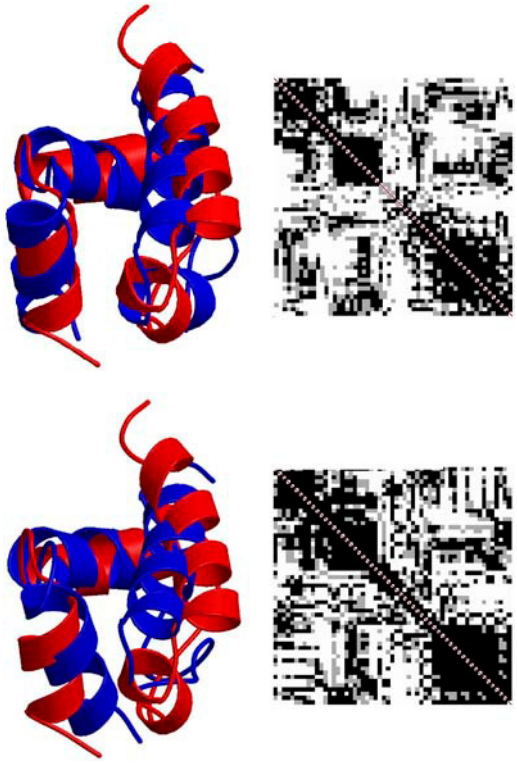

FIGURE 1 Overlay of the folded and the native conformation of the bacterial ribosomal protein L20 in simulations A and B (*upper* and *lower*, respectively) with the corresponding $C_\beta$-$C_\beta$ matrices. The upper triangle of the $C_\beta$-$C_\beta$ matrix shows absolute, the lower relative deviations between the folded and the experimental structure, respectively. Each square encodes the deviation between the $C_\beta$-$C_\beta$ distance of two amino acids in the experimental structure to the $C_\beta$-$C_\beta$ distance of the same amino acids in the folded structure. Black (*gray*) squares, deviation of $<1.50$ Å (2.25 Å); white squares, large deviations.

too surprisingly, resolved to very good accuracy. The off-diagonal dark blocks indicate the formation of a large fraction of correct long-range native contacts.

## Analysis of the final populations

Table 1 demonstrates that all low-energy conformations have essentially the same secondary structure, i.e., position and length of the helices are always correctly predicted, even if the protein did not fold correctly in all cases. Conformations with near-native secondary structure dominate the low-energy landscape of the protein (see also Fig. 5). This is quite remarkable, because the acceptance criterion does not favor the occurrence of similar conformations.

Not all, but many, conformations of the final population managed to approach the native conformation in good overall correlation with their energy. Fig. 4 shows the similarity of the members of the final populations ordered by energy. The prevalence of light areas near the top of the figure, particularly in comparison to the native structure in the top row, indicates that the final population contains many conformations approaching the native state. The selection criterion of the evolutionary algorithm also stabilizes nonnative conformations in the populations as long as they have a competitive energy (see Fig. 6). The degree of secondary-structure content and similarity decreases for the decoys with higher energy in good correlation with their energy. This demonstrates that the evolutionary algorithm strikes a good balance between energy improvement and diversity of the population.

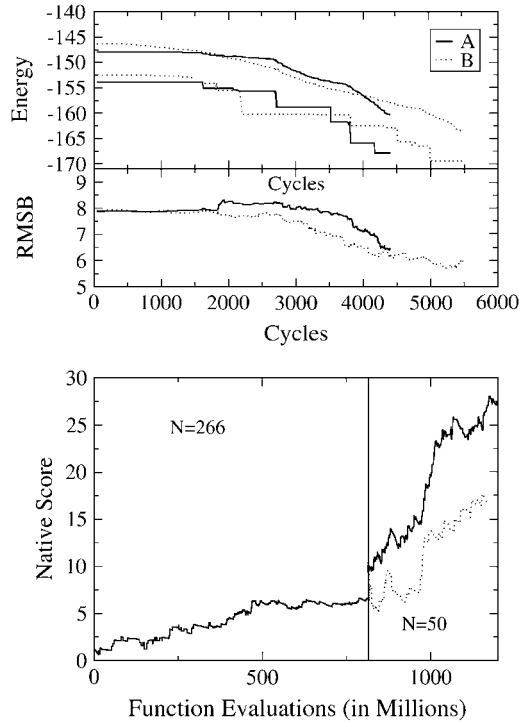


FIGURE 2 (*Top*) Average and minimal energy (*upper part*) and average RMSB deviations (*lower part*) as a function of iteration number for simulations A and B with $N = 50$. (*Bottom*) Native score in the phases with $N = 266$ and $N = 50$ of both simulations versus the number of function evaluations (*solid line*, simulation A; *dotted line*, simulation B).

The occurrence of many light areas in Fig. 4 suggests that the entire population is dominated by only a few distinct decoy families (39). This finding is visually confirmed in

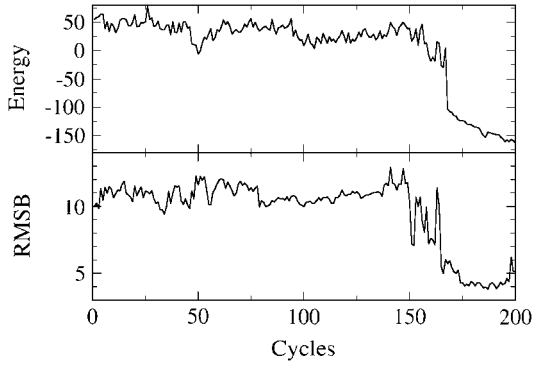


FIGURE 3 Energy (*upper*) and RMSB deviation (*lower*) of the best decoy in the final population of simulation B as a function of iteration number, indicating a continuous convergence of the simulation toward the native conformation.

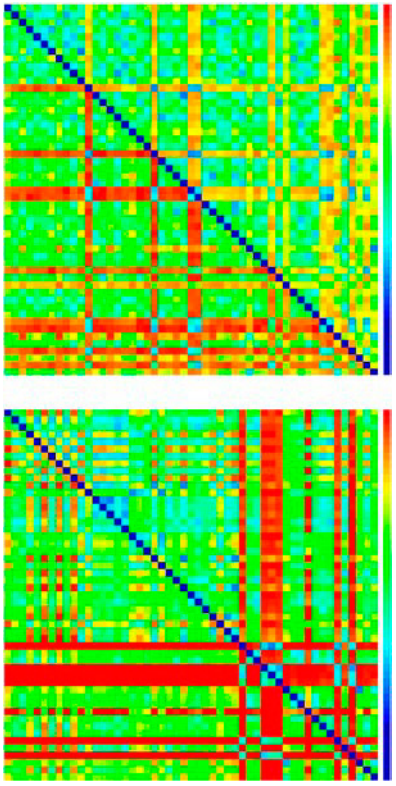

FIGURE 4 Color-coded distance matrix of the final conformations of simulations A (*top*) and B (*bottom*). In each panel, the upper right (lower left) triangle encodes the backbone (full) RMSB deviation between the members of the population. The top row and leftmost column in each figure show the native conformation. Blue/green (1- to 4-Å range), similar structures; red (deviations of 8–10 Å), large deviations. The conformations are sorted by energy, starting with the best from the top.

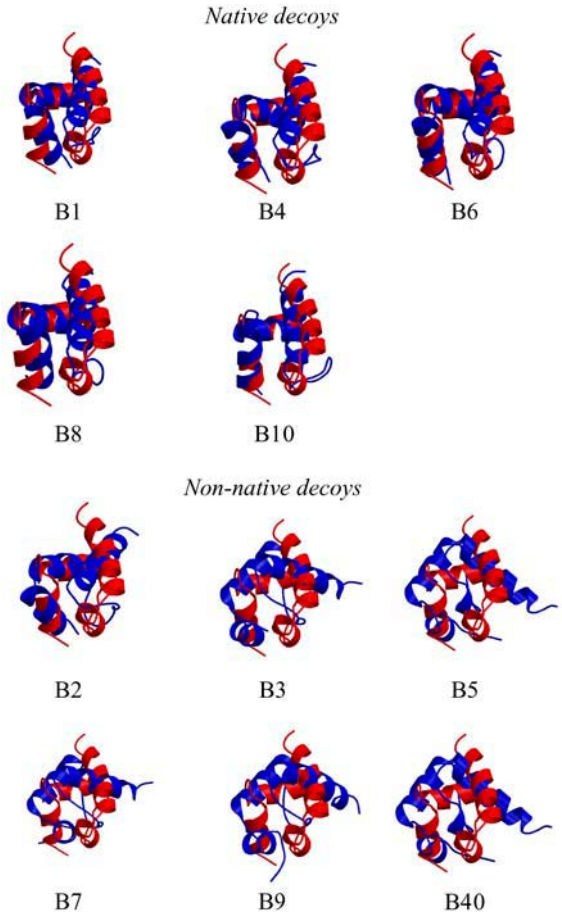


FIGURE 5 Overlay of the native and the energetically best decoys in the simulation B native family: B1, B4, B6, B8, B10, and nonnative family: B2, B3, B5, B7, B9. The first substantially different decoy (B40) is shown in the bottom row.

Fig. 5, which compares the energetically best five native and five nonnative decoys of simulation B. The first truly different decoy is found at position 40 of the list (corresponding to the bright red bars in Fig. 4. Even this decoy still has much of the correct secondary structure. The distribution of the low-energy conformations is in agreement with the funnel-hypothesis for protein folding. Our results point to the existence of many structurally similar low-energy conformations: if the folding process of bacterial ribosomal protein L20 is adiabatic, such low-energy conformations are populated with high probability in the late stages of the folding process. The final collapse to the native conformation can then be viewed as a set of transitions among the low-energy conformations such as those shown in Fig. 5.

## Convergence

Fig. 2 demonstrates the convergence of both the energy and the average RMSB deviation as the function of the number of total iterations (basin-hopping cycles). Both simulations had an acceptance ratio of ~30%. The simulation using equidistributed probabilities converged faster regarding the overall energy, but not regarding the average RMSB of the population. Both simulations smoothly approach the native conformation, as is also indicated by the plot of the native score (see Methods), which increases over 60-fold in the course of these simulations. Fig. 3 traces the development of energy and RMSB deviation of the best decoy in the final population of simulation B. Most of the helical structure forms early in the simulations. Because our dynamics is artificial, this does not necessarily imply early helix formation in the physical folding process. We note that there is a sudden rapid drop in RMSB deviation to the native conformation, which is accompanied by a rapid drop in energy.

## Method comparison

To date, we have failed to make significant progress for de novo folding of the bacterial ribosomal protein L20 with other stochastic optimization methods. To rationalize the success of the evolutionary algorithm, we compare simulations with the evolutionary algorithm and the basin-hopping technique for the Trp-cage protein. In a recent comparison of

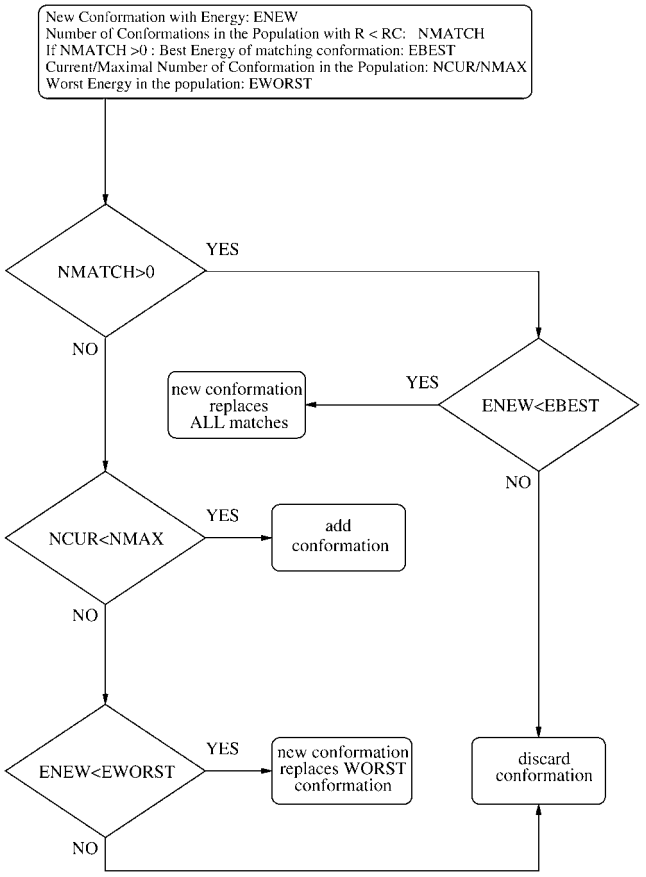


FIGURE 6 Chart of the decision-making process, when a newly generated conformation (with energy $E_{new}$) is presented to the master process. The worst matching conformation in the active population is replaced by the new conformation, if the latter differs from all present conformations and is lower in energy than the lowest conformation. If there are similar conformations, the closest (by RMSB) is replaced, if its energy is higher.

stochastic optimization methods (24), our version of the basin-hopping technique emerged as the most efficient technique for folding the Trp-cage protein. We first performed 20 independent basin-hopping simulations with 100 cycles per replica. We then ran an evolutionary algorithm using a population size of $N = 20$ and $P = 50$ processors that used the same total computational effort. The evolutionary algorithm used a uniform distribution for conformational selection (as simulation A for bacterial ribosomal protein L20). Both methods used exactly the same parameterization of the annealing cycle. Both simulations folded the protein, although the evolutionary algorithm obtained a marginally better energy at the end. The best and mean energies of both ''populations'' are shown in Fig. 7 as a function of the numerical effort.

Both simulations converge a ''population'' of 20 dynamical processes. Although these are independent in the basin-hopping method, they are coupled via the selection criteria in the evolutionary algorithm. We note two important differences between the methods: The best energy drops faster early in the simulation using the basin-hopping method. In

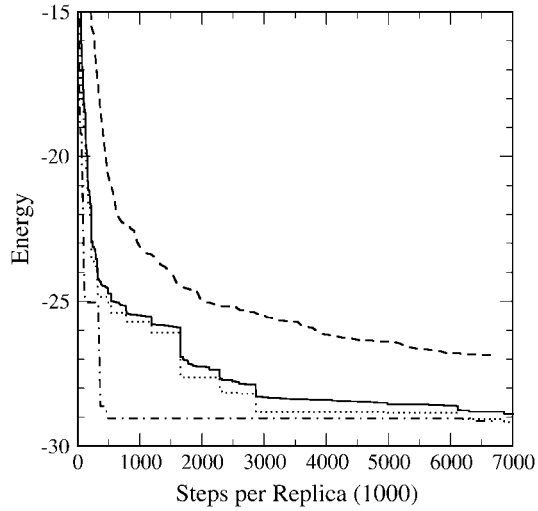


FIGURE 7 Minimal (*red*) and average energies (*black*) in kcal/mol for the basin-hopping simulations (*dashed lines*) and the evolutionary algorithm (*solid lines*) as a function of the numerical effort per population member (in thousands of function evaluations).

the basin-hopping method, the best conformation is selected with unit probability, which often leads to improvements in early cycles (high acceptance ratio). The evolutionary algorithm selects the best conformation only with probability $N/P \ll 1$; its best energy thus trails the basin-hopping method. This observation explains the difficulties of the evolutionary algorithm right after starting.

The long-term superiority of the evolutionary technique becomes obvious when comparing the average energies of the population: The average energy of the evolutionary algorithm is much lower than for the basin-hopping method. In basin-hopping a certain fraction of the simulations go completely astray: they never find low-energy or near-native conformations. The selection process in the evolutionary algorithm efficiently eliminates such conformations from the active population, generating nonlocal, large-step moves that lead to improved overall performance.

## DISCUSSION

The results of this study provide impressive evidence that all-atom protein structure prediction with free-energy force fields is becoming a reality (40). Our simulations yielded final populations with significant native content. Near-native conformations were selected on the basis of the energy criterion as probable stable structures. To date we have folded several helical proteins with similar accuracy using the PFF01 force field (see Table 2). This study adds the first four-helix protein to this set and demonstrates the transferability of the force field to more complex systems than the two- and three-helix systems treated previously.

All-atom protein structure prediction requires a sufficiently accurate force field and efficient optimization methods that

**TABLE 2 Summary of folding simulations with PFF01 and various optimization techniques**

| Name (PDB code) | Amino acids | No. of helices | Resolution (Å) | Method |
|---|---|---|---|---|
| Trp-cage (1L2Y) | 20 | 2 | 3.4 | STUN (7), PT (22), ELP (23), BHT (24), EA |
| 1WQC | 26 | 2 | 3.2 | BHT (24) |
| Villin headpiece (1VII) | 36 | 3 | 4.5 | BHT (25) |
| HIV acc. protein (1F4I) | 40 | 3 | 2.3 | BHT (9), PT (26) |
| Protein A (1BDD) | 40 | 3 | 3.2 | BHT (A. Verma, S. Murthy, and W. Wenzel, unpublished data) |
| Bacterial ribosomal protein L20 (1GYZ) | 60 | 4 | 4.3 | EA |

BHT, basin hopping; PT, parallel tempering; ELP, energy landscape paving; EA, evolutionary algorithm (this study). For some proteins (1BDD, 1F4I) the lengths of the simulated fragments are given rather than the size of the PDB structure.

can reliably locate the global optimum of the free-energy surface. Recently, near-native conformations formed in folding simulations of protein A (8) in a similar approach, indicating that all-atom protein structure prediction may augment homology-based methods, at least for medium-sized proteins, in the foreseeable future.

The free-energy approach exploits the 30-year-old thermodynamic hypothesis (14), according to which the native structure of many proteins can be predicted using stochastic optimization methods. Our results demonstrate that the important influence of the solvent can be modeled with a relatively simple solvent-accessible surface approach, at least for some proteins. The stochastic exploration of the free-energy surface is much faster than direct simulation, because nonphysical moves can be attempted and nonphysical intermediates tolerated. Its natural drawback is the loss of kinetic and thermodynamic information. Recent years have seen the emergence of computational methods to explore the native conformation and the transition-state ensemble with great kinetic detail (13,41). Free-energy optimization methods and replica-exchange explicit water molecular dynamics thus offer complementary views of the protein folding process.

We are grateful to S. Gregurick, J. Moult, and J. Pedersen for discussions and portions of the code used in these simulations.

This work was supported by the Deutsche Forschungsgemeinschaft, the Bode Foundation, and the Bundesministerium für Wissenschaft und Forschung. We acknowledge support of the Supercomputational Materials Laboratory of Korea Institute of Science and Technology (Seoul, South Korea), where some of these simulations were performed.

## REFERENCES

1. Baker, D., and A. Sali. 2001. Protein structure prediction and structural genomics. *Science*. 294:93–96.
2. Schonbrunn, J., W. J. Wedemeyer, and D. Baker. 2002. Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* 12:348–352.
3. Liwo, A., P. Arlukowicz, C. Czaplewski, S. Oldizeij, J. Pillardy, and H. Scheraga. 2002. A method for optimising potential energy functions by a hierarchichal design of the potential energy landscape. *Proc. Natl. Acad. Sci. U.S.A.* 99:1937–1942.
4. Moult, J., K. Fidelis, A. Zemia, and T. Hubbard. 2001. Critical assessment of methods of protein structure (CASP): round IV. *Proteins.* 45:2–7.
5. Snow, C. D., H. Nguyen, V. S. Pande, and M. Gruebele. 2002. Absolute comparison of simulated and experimental protein folding dynamics. *Nature*. 420:102–106.
6. Simmerling, C., B. Strockbine, and A. Roitberg. 2002. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* 124:11258–11259.
7. Schug, A., T. Herges, and W. Wenzel. 2003. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.* 91: 158102.
8. Vila, J., D. Ripoll, and H. Scheraga. 2004. Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc. Natl. Acad. Sci. U.S.A.* 100:14812–14816.
9. Herges, T., and W. Wenzel. 2005. Reproducible in-silico folding of a three-helix protein and characterization of its free energy landscape in a transferable all-atom forcefield. *Phys. Rev. Lett.* 94:018101.
10. Hansmann, U. H. E. 2002. Global optimization by energy landscape paving. *Phys. Rev. Lett.* 88:068105.
11. Lin, C., C. Hu, and U. Hansmann. 2003. Parallel tempering simulations of hp-36. *Proteins*. 52:436–445.
12. Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*. 282:740–744.
13. Garcia, A. E., and N. Onuchic. 2003. Folding a protein in a computer: an atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci. U.S.A.* 100:13898–13903.
14. Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science*. 181:223–230.
15. Onuchic, J. N., Z. Luthey-Schulten, and P. Wolynes. 1997a. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.
16. Hardin, C., M. Eastwood, M. Prentiss, Z. Luthey-Schulten, and P. Wolynes. 2003. Folding funnels: the key to robust protein structure prediction. *J. Comput. Chem.* 23:138–146.
17. Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997b. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.
18. Dill, K., and H. Chan. 1997. From Levinthal to pathways to funnels: The ''new view'' of protein folding kinetics. *Nat. Struct. Biol.* 4:10–19.
19. Levinthal, C. 1968. Are there pathways for protein folding? *J. Chem. Phys.* 65:44–45.
20. Herges, T., A. Schug, H. Merlitz, and W. Wenzel. 2003. Stochastic optimization methods for structure prediction of biomolecular nanoscale systems. *Nanotechnology*. 14:1161–1167.
21. Herges, T., and W. Wenzel. 2004. An all-atom force field for tertiary structure prediction of helical proteins. *Biophys. J.* 87:3100–3109.
22. Schug, A., and W. Wenzel. 2004. All-atom folding of the Trp-cage protein in an all-atom forcefield. *Europhys. Lett.* 67:307–313.
23. Schug, A., W. Wenzel, and U. Hansmann. 2005. Energy landscape paving simulations of the Trp-cage protein. *J. Chem. Phys.* 122:194711.
24. Verma, A., A. Schug, K. H. Lee, and W. Wenzel. 2006. Basin hopping simulations for all-atom protein folding. *J. Chem. Phys.* 124:044515.
25. Herges, T., and W. Wenzel. 2005b. Free energy landscape of the villin headpiece in an all-atom forcefield. *Structure*. 13:661–668.
26. Schug, A., T. Herges, and W. Wenzel. 2004. All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method. *Proteins*. 57:792–798.

27. Kirkpatrick, S., C. Gelatt, and M. Vecchi. 1983. Optimization by simulated annealing. *Science*. 220:671–680.

28. Raibaud, S., I. Lebars, M. Guillier, C. Chiaruttini, F. Bontems, A. Rak, M. Garber, F. Allemand, M. Springer, and F. Dardel. 2002. NMR structure of bacterial ribosomal protein l20: implications for ribosome assembly and translational control. *J. Mol. Biol.* 323:143–151.

29. Schug, A., and W. Wenzel. 2004b. Predictive in-silico all-atom folding of a four helix protein with a free-energy model. *J. Am. Chem. Soc.* 126:16736–16737.

30. Withers-Ward, E. S., T. Mueller, I. Chen, and J. Feigon. 2000. Biochemical and structural analysis of the interaction between the UBA(2) domain of the DNA repair protein HHR23A and HIV-1 Vpr. *Biochemistry*. 39:14103–14112.

31. Herges, T., A. Schug, and W. Wenzel. 2004. Exploration of the free energy surface of a three helix peptide with stochastic optimization methods. *Int. J. Quantum Chem.* 99:854–893.

32. Abagyan, R. A., and M. Totrov. 1994. Biased probability Monte Carlo conformation searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* 235:983–1002.

33. Avbelj, F., and J. Moult. 1995. Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*. 34:755–764.

34. Pedersen, J. T., and J. Moult. 1997. Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.* 269:240–259.

35. Wales, D. J., and J. P. Doye. 1997. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem.* 101:5111–5116.

36. Abagyan, R. A., and M. Totrov. 1999. Ab initio folding of peptides by the optimal bias Monte Carlo minimization procedure. *J. Comp. Phys.* 151:402–421.

37. Mortenson, P. N., and D. J. Wales. 2004. Energy landscapes, global optimization and dynamics of poly-alanine $Ac(ala)_8NHMe$. *J. Chem. Phys.* 114:6443–6454.

38. Mortenson, P. N., D. A. Evans, and D. J. Wales. 2002. Energy landscapes of model polyalanines. *J. Chem. Phys.* 117:1363–1376.

39. Levitt, M., and C. Chothia. 1976. Structural patterns in globular proteins. *Nature*. 261:552–558.

40. Garcia, A., and J. Onuchic. 2005. Folding a protein on a computer: hope or reality. *Structure*. 13:497–503.

41. Mayor, U., N. R. Guydosh, C. M. Johnson, J. G. Grossmann, S. Sato, G. S. Jas, S. M. V. Freund, D. O. V. Alonso, V. Daggett, and A. R. Fersht. 2003. The complete folding pathway of a protein from nanoseconds to microseconds. *Nature*. 421:863–867.