Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 4436–4440, May 1994
Biophysics

An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function

JAMES U. BOWIE AND DAVID EISENBERG

Department of Chemistry and Biochemistry, University of California, Los Angeles–Department of Energy Laboratory of Structural Biology and Molecular Medicine, Molecular Biology Institute, University of California, Los Angeles, CA 90024-1570

Contributed by David Eisenberg, November 12, 1993

ABSTRACT **Three short protein sequences have been guided by computer to folds resembling their crystal structures. Initially, peptide fragment conformations ranging in size from 9 to 25 residues were selected from a database of known protein structures. A fragment was selected if it was compatible with a segment of the sequence to be folded, as judged by three-dimensional profile scores. By linking the selected fragment conformations together, hundreds of trial structures were generated of the same length and sequence as the protein to be folded. These starting trial structures were then improved by an evolutionary algorithm. Selection pressure for improving the structures was provided by an energy function that was designed to guide the conformational search procedure toward the correct structure. We find that by evolution of only 400 structures for fewer than 1400 generations, the overall fold of some small helical proteins can be computed from the sequence, with deviations from observed structures of 2.5–4.0 Å for Cα atoms.**

In protein folding by computer, it is not possible to exhaustively test all possible protein conformations because of the huge number of structures a protein chain can adopt (1). Consequently, some sort of directed search to the correct structure is required. We employ two strategies to direct the search of conformation space. First, we build trial structures using fragment conformations selected from a database of known protein structures. Because local conformations are restricted by these initial fragment selections, the number of possible conformations is reduced. The second strategy is to design an energy function that decreases gradually as the true conformation is approached, because an energy function that maintains a gradient in the direction of the correct conformation can in essence direct the search of conformation space (Fig. 1). This can be appreciated by considering an alternative potential energy function that is flat everywhere except at the true conformation. With a flat energy function, an exhaustive search of conformation space would be required since the energy provides no information about how close the conformation is to the true structure. We describe here a method to generate a guiding energy function.

METHODS

Extracting Compatible Fragments from a Structure Database. We built the initial trial structures for later evolution from a mixture of small fixed-length (9 residues) and longer variable-length (15–25 residues) fragments, as shown in Fig. 2*A*. These fragments were extracted from a database of known protein structures by two enrichment procedures based on the three-dimensional (3D) profile method of Bowie *et al.* (2), which assesses the compatibility of a sequence for a structure. In the 3D profile method, each residue receives a score according to how well it fits its environment in the structure. For the present application, we determine the score by replacing a sequence segment in a protein of known structure with the sequence of the protein to be folded.

Short Fragments. For fixed-length fragments, 9-residue segments were extracted from the sequence to be folded and scored for compatibility with all possible 9-residue environment patterns in a database of protein structures. Proteins in this database that are homologous [as defined by Sander and Schneider (3)] to the protein to be folded were excluded. The 15 most compatible fragments were selected and the conformations were stored. This procedure indeed enriches for fragments that are more likely to be correct. Only 17% of random 9-residue fragments in a database of 71 nonhomologous proteins were found to be within a 1.75-Å distance matrix error (DME) (4) on Cα atoms, whereas in the enriched pools, 31% were within a 1.75-Å DME.

Long Fragments. Up to 100 longer fragments ranging in size from 15 to 25 residues were selected by finding the best alignments, without gaps, of the sequence to be folded with the environment patterns of structures in a structure database (2). Fragments longer than 25 residues were not found to be useful. Initially, 500 top-scoring alignments were taken from the full Brookhaven Protein Data Bank (5) of structures, excluding only those homologous to the protein to be folded [by the criteria of Sander and Schneider (3)]. In a second step, fragments homologous to higher scoring fragments were removed from the list, and we saved a maximum of 100 top-scoring nonhomologous fragments. Again this procedure enriches for fragments that are more likely to be correct. In the length range of 15–19 residues, 12% of fragments extracted by this procedure are within a DME of 2 Å and 28% are within a DME of 3 Å, compared with 4% within a DME of 2 Å and 17% within a DME of 3 Å for random fragments of this length. In the length range of 20–25 residues, 3% of extracted fragments are within a DME of 2 Å and 12% are within a DME of 3 Å, compared with 1% within a DME of 2 Å and 4% within a DME of 3 Å for random fragments of this length.

Building a Population of Trial Structures from the Selected Fragments. Hundreds of initial trial structures, with conformations described by dihedral angles, were built from the enriched pools of 9-residue fragments and the list of up to 100 longer fragments by the following procedure. First, conformations were built from the N terminus to the end using nonoverlapping 9-residue fragments. Starting at the N terminus, a fragment was randomly selected for residues 1–9 from the fragment pool for this segment of sequence. Next, a fragment was randomly selected for residues 10–18, then residues 19–27, and so forth until the C terminus. The dihedral angles from these fragments became the dihedral angles describing the trial structure. If the sequence length was not an exact multiple of 9, the end was filled in with a smaller fragment. No effort was made to adjust the dihedral

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: 3D, three dimensional; DME, distance matrix error.

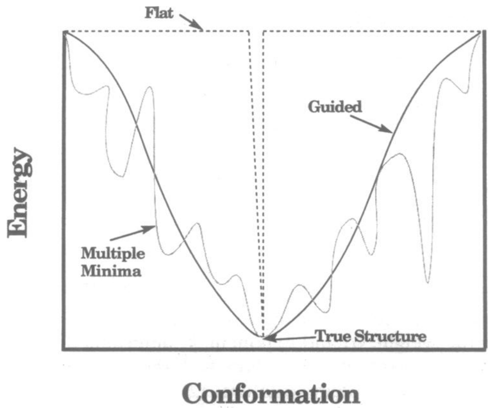

FIG. 1. Three hypothetical one-dimensional potential functions for the folding of a protein, illustrating the computational problems in protein folding. The flat potential has a minimum at the true structure but is not helpful in finding the minimum. This defect is removed in the guided potential. The multiple-minimum potential is more realistic and introduces the problem of barriers between local minima and the true structure. The evolutionary algorithm is capable of overcoming these local barriers.

angles at the fragment joints. This procedure was repeated with other random selections to build up the other trial structures. For half the trial structures, the dihedral angles were overwritten by fragments offset by 4 residues. Thus, residues 1–4 were unaffected, but the remaining residues were built from fragments for residues 5–14, residues 15–23, etc. At this point then, half the trial structures were built from nonoverlapping 9-residue fragments starting at the N terminus and half were built mostly from nonoverlapping 9-residue fragments starting a residue 5. Next, for each trial structure, two nonoverlapping large fragments were randomly selected and the conformations for those regions were replaced by the conformation of the large fragments. Thus, up to 50 residues could possess the dihedral angles of these large fragments. This starting population of initial trial structures was then optimized using an evolutionary algorithm.

The Evolutionary Algorithm. The energy (see below) of the population was minimized with an evolutionary algorithm (6). Others have recently applied the genetic algorithm to problems in protein folding (7–10). The algorithm is outlined in Fig. 2*B*. Each trial structure of a given protein was encoded as a set of dihedral angles. An initial population of hundreds of parental trial structures was generated by linking fragments of known protein structures, as described above. In the analogy to genetics, each dihedral angle can be considered as a gene and the set of dihedral angles that make up a trial structure can be considered as a chromosome. Each parental chromosome then generated one child, by duplication, which was then subject to mutations and recombinations. A mutation was a change of a dihedral angle that could either be a random selection from a database of rotamers or a slight adjustment of its current angle. In recombination, a segment of a child's chromosome was replaced by the dihedral angles of the same sequence segment from another randomly chosen parent. To preserve the local fragment conformations for enough generations to be tested in a variety of tertiary structure environments, mutations, and recombination end points had a higher probability at the fragment junctions.

A new population of parental and child trial structures was thereby generated that was twice as big as the starting population and with all children being different from their parents. The energy score, for each trial structure in the new population, was determined and then each trial structure was made to compete with 10 randomly chosen trial structures in the population. A win was assigned to the trial structure with the lower energy score. The 50% of the trial structures with the most wins survived to become the next population and the

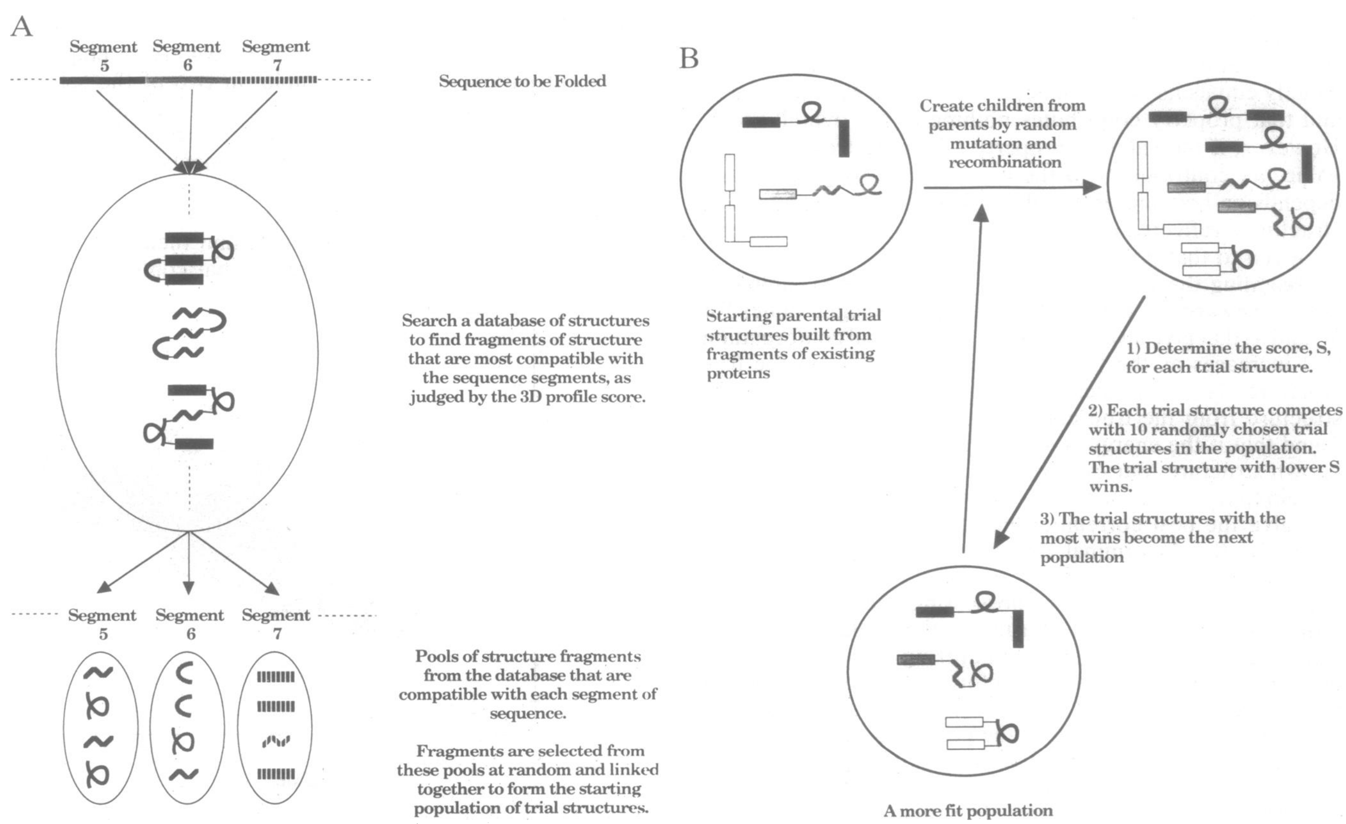

FIG. 2. Evolutionary algorithm. (*A*) The initial stage in which promising fragments are selected. (*B*) The stage of evolutionary cycling in which conformations are replicated with mutations and recombinations and the fittest is selected for the next cycle.

Table 1. Fitness function for a protein structure

Property	σ	Set 1		Set 2		Set 3	
		W	X	W	X	W	X
Profile score	0.27	0.0	0.0	0.0	0.0	1.53	0.90
Hydrophobicity contrast C	0.31	7.79	0.84	9.97	0.59	3.08	1.44
Total accessible surface area	0.06	3.72	1.6	1.67	1.52	5.13	1.11
Overlap penalty I	0.05	0.64	1.26	2.38	0.36	3.35	0.45
Overlap penalty II	0.05	3.39	0.87	1.45	1.19	2.04	1.03
Fatness	0.13	3.57	0.97	0.95	1.60	0.0	0.0

Fitness function, $S = \Sigma_{i=1}^{6} W_i|(P_i/\langle P_i\rangle - 1.0)/\sigma_i|^{X_i}$, where $P_i/\langle P_i\rangle$ is the ratio of the ith property for the current conformation P_i to its target value $\langle P_i\rangle$, determined from a database of known protein structures. The value of σ gives the standard deviation of property i in the database. The three sets give the optimal values for the weight W_i and exponent X_i determined from three sets of 63-residue conformations having the 434 repressor sequences. The profile score is how well the environment at each side chain matches the environment observed for that side chain in known protein structures (2). Target value for profile score = $10^{(-3.0674+1.0847 \cdot \log M)}$, where M is the molecular weight. Hydrophobicity contrast C is a global measure of segregation of polar groups and apolar groups in a protein structure (12). $C = \Sigma_i \Delta\sigma_i r_i - \langle\Delta\sigma\rangle\langle r\rangle$, where $\Delta\sigma$ is the atomic solvation parameter of an atom i at a distance r from the center of mass. Target value = $-4.3148 \times 10^{-4} M - 0.47444$. The total accessible surface area is from ref. 13. Target value = $10^{(0.38137+0.83203 \cdot \log M)}$. Overlap penalty I is the number of atom pairs within 50% of the sum of van der Waals radii. Target value = $17.123 + 0.14286 \cdot M$. Overlap penalty II is the number of atom pairs between 50% and 80% of the sum of van der Waals radii. Target value = $-4.8257 + 0.21560 \cdot M$. Fatness is the ratio of longest to shortest axis of inertial ellipsoid for the structure (14). Target value = 1.54.

remaining trial structures were extinguished. A new population was thereby generated that was the same size as the first but contained trial structures that were fitter on average than the members of the parental population, as judged by the energy score. Because there is a tendency for a child to resemble its parent, a population can quickly become dominated by a single highly fit conformation and its offspring. To minimize this effect, we incorporated a sharing procedure as described by Palmer and Smith (11). The cycle of duplication with variation and selection was repeated until the lowest energy score did not improve for a preset number of generations. Details of the evolutionary algorithm are accessible via anonymous file transfer protocol (FTP).*

The Energy Function. The potential energy or fitness functions of this paper are a sum of six terms (Table 1), each of which describes a global property of a protein chain. Notice that if the value of property i for a conformation, P_i, exactly equals its target value, $\langle P_i\rangle$, the ratio $P_i/\langle P_i\rangle$ equals 1 and that property contributes nothing to the overall score. However, if the value of property i deviates from its target value, its contribution to the overall score is positive and depends on how uniform that property is for known protein structures, reflected in σ_i, the weight, W_i, and the exponent, X_i, given to that property.

Weighting the Fitness Function to Drive the Evolutionary Search Toward the Correct Structure. We sought values for W_i and X_i that give an energy score, S, for a trial conformation that increases as the trial structure diverges from the true structure. A useful measure of the divergence of a trial structure from the true conformation is DME. In fact, we found that if the energy is simply taken as the DME from a known structure, our evolutionary algorithm readily folds a protein to that structure. Thus we sought values of W_i and X_i that give the best correlation between the score S and the DME for the N-terminal domain of 434 repressor, a 63-residue helical protein (15).

Groups of conformations possessing the 434 repressor sequence were generated by running the evolutionary algorithm described above, but minimizing the DME of the population. Representative conformations were selected throughout the evolution. Because good conformations tend to increase their proportion of the population exponentially, we selected an exponentially decreasing number of conformations as the generations proceeded. Three separate sets of conformations were generated. The total conformations in each group between 0- and 10-Å DME were 1705, 1505, and 1526.

Weights and powers for each property were adjusted separately for each group, also by means of an evolutionary algorithm, to fit the line: $S = 10 \cdot$DME. The quantity minimized was RMSDev-All + 10·RMSDev-Min. RMSDev-All is the rms deviation of all data points to the line. RMSDev-Min is the rms deviation of the lowest score values in each 0.5-Å DME bin. Thus, deviations below the line were penalized more heavily than deviations above the line. This was done to make it more difficult for conformations to settle into a good local minimum far from the correct conformation. Sets of weights and powers where the correct structure was not at the global minimum were eliminated from the population. The three resulting sets of weights and exponents are given in Table 1.

The different sets of weights and exponents emphasize different parameters. For example, in two of the sets, the profile score contributes nothing to the overall fitness score, and in one, the fatness measure contributes nothing. While any of these sets of weights and exponents could be used separately, we found the most effective scoring function was to determine three scores for each conformation using the three sets of weights and exponents and then to assign the highest of the three as the score for that conformation. By

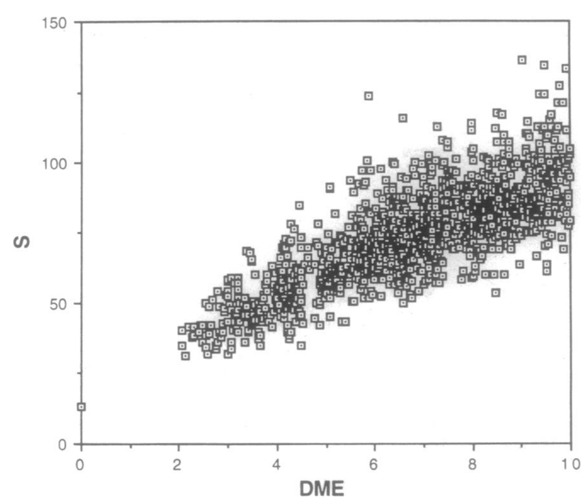

FIG. 3. Plot of the energy or fitness score, S, vs. DME relative to 434 repressor for a set of 63-residue conformations. The weights were not optimized on the set of conformations shown here.

*fourier.mbi.ucla.edu (128.97.39.21), directory pub/evolution.

analogy to an evolutionary selection, we test the ability of the organism (the trial structure) to survive in three environments. Fig. 3 displays a good correlation between DME and S, suggesting that minimizing S by varying the conformation of a small helical protein will tend to drive the conformation toward the true structure. Details of the energy function generation are accessible via anonymous FTP.*

The Structure Databases. The databases used are available via anonymous FTP.*

RESULTS

Folding 434 Repressor. As a first test of the algorithm, we attempted to fold 434 repressor, although our results with this protein are biased by information from the crystal structure that we had used to optimize our empirical energy function S. By starting from a population of 400 conformations generated from randomly assembled fragments (Fig. 2*A*), the population was optimized (Fig. 2*B*) until 200 generations passed without an improvement in the energy of the top scoring conformation. The evolutionary algorithm dropped the energy S of the top conformation in the population from >50 in the first generation to final values of 13.6, 10.4, and 10.1 for the three runs. The energy of the crystal structure of 434 repressor is 13.2. The structure in the three final populations that was closest to the crystal structure of 434 repressor is shown in Fig. 4. For the first and third runs, the best structures do indeed fold into structures that closely resemble the true structure, both having DMEs of only 3.0 Å on $C\alpha$ positions.

In none of the three attempts to fold 434 repressor was the best conformation the one with the lowest energy. The ranks of the three best conformations were 59, 400, and 220 out of 400 in the population. Moreover, the lowest energy conformations in the final populations had relatively poor DMEs of 4.0, 5.4, and 5.1 Å. Thus, our energy function cannot absolutely distinguish between good and poorer structures. Low-energy wells distant from the true structure clearly exist and are accessible to our conformational search. Nevertheless, our procedure is capable of finding the correct region of conformation space with high probability. Out of the 1200 structures in the three final populations, more than one-third are within a 4.0-Å DME of the correct structure. Moreover, these 1200 structures were arrived at after evaluating fewer than 1 million conformations. For comparison, we generated 1 million 434 repressor conformations by using dihedral angles selected at random from a database and the best conformation obtained had a DME of 4.4 Å. Given the huge number of possible conformations a 63-residue protein could adopt, we have achieved an enormous enrichment in structures that resemble the true structure.

Engrailed and Protein A Folding Attempts. We next tested the algorithm on two other small proteins: the 57-residue engrailed homeodomain (17) and a 50-residue fragment of the B domain of protein A (residues 8–57) (18). Both proteins represent favorable cases because both are small, largely helical, single-domain proteins lacking both disulfide bonds and metals. These tests are otherwise unbiased by knowledge of the structure. The evolutionary algorithm was provided with no information about the true structure, other than its amino acid sequence, and no parameters were changed from the 434 repressor folding attempts.

The crystal structure of the engrailed homeodomain consists of a three-helix protein core and an N-terminal arm that extends from the main body of the protein and inserts into the minor groove of the DNA to which it is bound (17). Because this arm is not packed against the rest of the protein, it is likely that the arm adopts a different structure in solution and may well be disordered. Although we kept the arm sequence in our folding attempts, it was almost invariably in a very different orientation than in the crystal structure, which resulted in overall poor DME measures even though the remainder of the protein was folded correctly. Consequently, we describe the results with reference to the core region only.

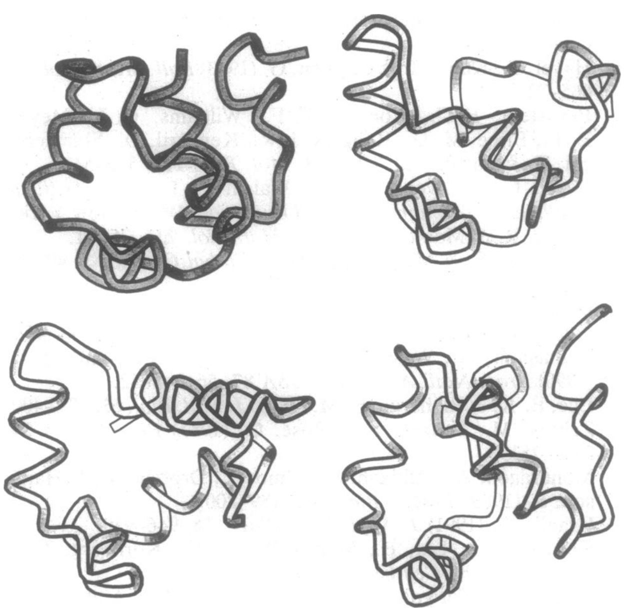

FIG. 4. Best conformations of 434 repressor, as judged by the DME measure of coordinate error, in three folding attempts. The true crystal structure is shown in darker grey in the upper left corner. The best conformation for run 1 is shown in the upper right corner (DME = 3.0 Å; S = 14.7; rank = 59), for run 2 it is shown in the lower left corner (DME = 4.5 Å; S = 27.4; rank = 400), and for run 3 it is shown in the lower right corner (DME = 3.0 Å; S = 11.9; rank = 220). Runs 1, 2, and 3 converged after 813, 718, and 940 generations, respectively. The figure was made using the program MOLSCRIPT (16).

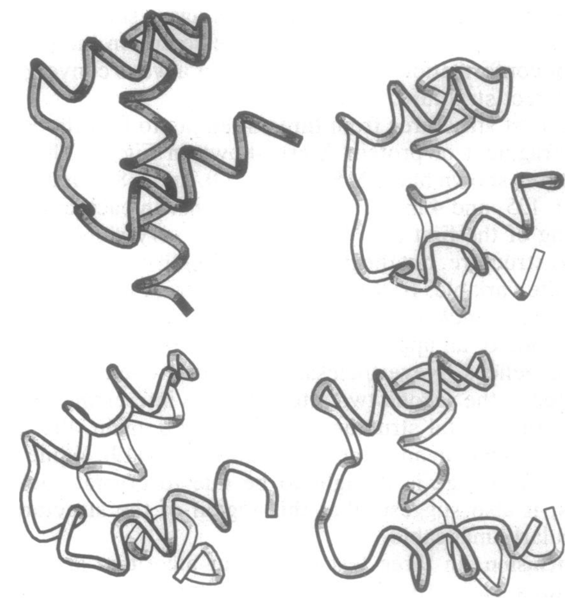

FIG. 5. Best conformations of engrailed homeodomain, as judged by the DME measure of coordinate error, in three folding attempts. Only residues 10–59 are shown, removing N-terminal arm. The true crystal structure is shown in darker grey in the upper left corner. The best conformation for run 1 is shown in the upper right corner (DME = 2.3 Å; S = 8.1; rank = 7), for run 2 it is shown in the lower left corner (DME = 3.3 Å; S = 17.1; rank = 91), and for run 3 it is shown in the lower right corner (DME = 3.4 Å; S = 16.8; rank = 378). Runs 1, 2, and 3 converged after 1375, 556, and 668 generations, respectively. The figure was made using the program MOLSCRIPT (16).

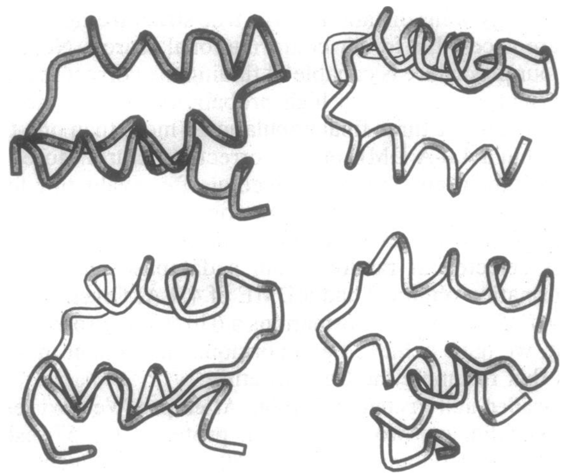

FIG. 6. Best conformations of the B domain of protein A as judged by the DME measure of coordinate error in three folding attempts. The true crystal structure is shown in darker grey in the upper left corner. The best conformation for run 1 is shown in the upper right corner (DME = 3.2 Å; S = 11.2; rank = 100), for run 2 it is shown in the lower left corner (DME = 3.0 Å; S = 11.9; rank = 125), and for run 3 it is shown in the lower right corner (DME = 3.8 Å; S = 7.4; rank = 4). Runs 1, 2, and 3 converged after 1057, 781, and 1294 generations, respectively. The structure shown for run 1 is not necessarily the best structure in the population, but DME is not good at distinguishing the incorrect topology shown, from the correct topology. The figure was made using the program MOLSCRIPT (16).

The best structures in the final population for three folding attempts are shown in Fig. 5. These structures had DMEs for the 50-residue core region of 2.3 Å, 3.3 Å, and 3.4 Å and ranks of 7, 91, and 378 in the final populations of 400. In each case, the basic fold is correct, including most of the secondary structure, but the third long helix is broken and distorted. We note that better-formed structures with intact third helices did appear earlier in several of these runs, but the distorted helices won out in the long run. This again shows that the energy function can rapidly focus the search in the correct area of conformation space but does not always converge on the correct structure.

The best structures from three attempts to fold the B-domain fragment of protein A are shown in Fig. 6. The best structures shown had DMEs of 3.2, 3.0, and 3.8 Å and ranks of 100, 125, and 4 in the final populations. In each case, the packing of the first and second helix is very similar to the crystal structure, but the position and length of the third helix are quite variable. In the best structure from the first run, a three-helix bundle forms, but the third helix is actually on the wrong side of an approximate plane formed by the first and second helices. Nevertheless, the correct overall fold was obtained in the second two attempts. It is interesting to note that in the crystal structure of this same domain, the third helix is actually in an extended conformation (19). While this structural change is thought to be due to crystal packing forces, it also suggests that this region of the structure is particularly malleable.

Conclusion. In three test cases, our algorithm arrives at structures close to the correct structure among a small collection of possible structures. Our approach does have a number of obvious limitations, however. First, although our energy function is capable of moving the search of conformation space into fruitful areas, it is not able to drive the search all the way to the final structure because some poor conformations have low-energy scores. Moreover, the energy function is currently applicable only to small helical proteins since it includes no properties, such as a long-range hydrogen-bonding term, that would favor β-sheet formation. Nevertheless, we have described an approach by which properties can be added that might improve the stringency and increase the generality of the energy function. Second, while the size of conformation space is greatly reduced by restricting local conformations to ones that have been selected from known structures, if the correct fragment conformation does not exist in some structure in the database, we have a diminished chance of evolving to the correct conformation.

Despite these limitations, we have been able to fold a few small helical proteins to conformations close to their true structures after evaluating fewer than half a million conformations—a tiny fraction of all the structures these protein chains can adopt. This limited success suggests that it may be possible to direct a search of conformation space, without resorting to explicit descriptions of the folding pathway.

We thank the Larry, David, and Gary Fogel, Tom Terwilliger, Scott LeGrand, and Robert Weiss for helpful discussions about evolutionary algorithms, Roland Lüthy for generating the three-dimensional–one-dimensional scoring table used in this work, and Scott LeGrand for sharing his work on accessible surface area calculations with us. Programming was assisted by routines from the Genetics Computer Group (Madison, WI) and PROTEUS (20) subroutine libraries. This work was supported by the National Institutes of Health and the Lita Annenberg Hazen Charitable Trust. J.U.B. was supported by a fellowship from the American Cancer Society and the National Science Foundation Program in Mathematics and Molecular Biology.

1. Levinthal, C. (1969) in *Mossbauer Spectroscopy in Biological Systems*, eds. Debrunner, P., Tsibris, J. C. M. & Munck, E. (Univ. Illinois Press, Champaign), pp. 22–24.
2. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991) *Science* **253**, 164–170.
3. Sander, C. & Schneider, R. (1991) *Proteins Struct. Funct. Genet.* **9**, 56–68.
4. Havel, T., Kuntz, I. & Crippen, G. (1983) *Bull. Math. Biol.* **45**, 665–720.
5. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.
6. Fogel, L. J., Owens, A. J. & Walsh, M. J. (1966) *Artificial Intelligence Through Simulated Evolution* (Wiley, New York).
7. Unger, R. & Moult, J. (1993) *J. Mol. Biol.* **231**, 75–81.
8. Le Grand, S. & Merz, K. (1993) *J. Global Optimiz.* **3**, 49–66.
9. Sun, S. (1993) *Protein Sci.* **2**, 762–785.
10. Dandekar, T. & Argos, P. (1992) *Protein Eng.* **5**, 637–645.
11. Palmer, M. & Smith, S. (1991) *Complex Systems* **5**, 443–458.
12. Yamashita, M., Wesson, L., Eisenman, G. & Eisenberg, D. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 5648–5652.
13. Lee, B. K. & Richards, F. M. (1971) *J. Mol. Biol.* **55**, 379–400.
14. Zehfus, M., Seltzer, J. & Rose, G. (1985) *Biopolymers* **24**, 2511–2519.
15. Mondragon, A., Subbiah, S., Almo, S., Drottar, M. & Harrison, S. (1989) *J. Mol. Biol.* **205**, 189–200.
16. Kraulis, P. (1991) *J. Appl. Crystallogr.* **24**, 946–950.
17. Kissinger, C., Liu, B., Martin-Blanco, E., Kornberg, T. & Pabo, C. (1990) *Cell* **63**, 579–590.
18. Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y. & Shimada, I. (1992) *Biochemistry* **31**, 9665–9672.
19. Deisenhofer, J. (1981) *Biochemistry* **20**, 2361–2370.
20. Pabo, C. & Suchanek, E. (1986) *Biochemistry* **25**, 5987–5991.