Folding the Main Chain of Small Proteins with the Genetic Algorithm

Thomas Dandekar and Patrick Argos

European Molecular Biology Laboratory
Postfach 102209, 69012 Heidelberg, Germany

Grid-free protein folding simulations were effected using the genetic algorithm, a backbone representation and standard dihedral angular conformations. The topological folding of idealized four-helix bundles was investigated in detail to differentiate among the important protein folding forces used as fitness criteria. Hydrophobic interactions were the most significant while local forces and hydrogen bonds were far less effective in promoting folding. Stable secondary structural regions were also important as nucleating centers. Using the fitness parameters optimized in idealized simulations together with standard secondary structure predictions derived from the amino acid sequence alone, the proper main-chain folding of the four-helix bundle proteins cytochrome b_{562}, cytochrome c' and hemerythrin was achieved. In addition the backbone topology as predicted by the genetic algorithm for crambin, a mixed helix/strand protein with known structure, is presented and discussed.

Keywords: genetic algorithm; protein folding; protein structure prediction

1. Introduction

Genetic algorithms were originally introduced by John Holland (1975) to exploit the principles of natural adaptation in complicated computer tasks involving searches for optimized solutions over large combinatorial character spaces. A population of random bit strings is used for starting solution trials. Each trial is decoded for a particular application and the quality of its solution judged by fitness criteria. The better the solution, the higher the probability that the corresponding bit string (a "chromosome") is selected as a parent for the next generation. In addition, selected parents may recombine ("genetic crossover") with a certain probability; further, a low mutational frequency is introduced when children strings are copied from parent strings. Succeeding generations should thus encode increasingly better solutions ("fitter populations") through selection, mutation and recombination. The process is repeated to yield a near-optimal solution.

Genetic algorithms have been applied in numerous technical problems, such as gas pipeline flow or artificial intelligence (Goldberg, 1989). As the prediction of protein folding and structure represents a formidable optimization search task in atomic conformational space with near limitless possibilities (Sternberg & Thornton, 1978), we previously explored the potential of genetic algorithms in protein structure and have described three-dimensional grid-bound models and various sequence and folding applications (Dandekar & Argos, 1992). Subsequently genetic algorithms have also been demonstrated to be superior to Monte Carlo simulations in two-dimensional protein models (Unger & Moult, 1993). Sun (1993) has recently pioneered the use of full energy terms and genetic algorithms to calculate the tertiary structure of very small proteins such as mellitin. Our present work abandons the grid-bound models we described earlier and exploits the ability of the genetic algorithm to fold the main-chain of small proteins from a knowledge of the primary sequence and predictions of its secondary structure from the sequence alone. The effect of different folding forces (e.g. hydrophobic interactions) and their importance is examined through the study of idealized four-helix bundles. Fitness criteria, simulating the forces with appropriate parameters and weights determined in the idealized cases, were then used in a genetic algorithm to predict the tertiary backbone fold of proteins with experimentally known structures, including four-helix bundles and a mixed strand/helix protein, all with sizes up to 120 residues.

2. Methods

(a) *The genetic algorithm*

Pascal programs were written (T. Dandekar) for this study utilizing a basic genetic algorithm as described by Goldberg (1989). Grid-free model representations of the proteins were included and subprograms analyzing the quality of the 3-dimensional protein structure effected.

0022-2836/94/080844–18 $08.00/0

The conformations were represented by binary digits (see below), which offers the maximum number of schemata per bit for any coding (Goldberg, 1989). The crossover procedure was kept simple; i.e. 2 parents in the population were selected for crossover with a probability of 0·2 and only 1 random crossover site in their chromosomes was allowed. Sufficient repetitions of such single crossovers are able to yield most recombination events achievable by more sophisticated crossover algorithms (Goldberg, 1989). The standard simulation was run for a generation time/population product of 400,000 (632 individuals for 632 generations), requiring 14 h computer time on a VAX 3100 work station. The number of generations for the simulation was set high to allow convergence of the simulation to a stable protein structure.

(b) *Model representation*

A simplified grid-free protein representation was used: internal (angular) coordinates connected the main-chain atoms; standard peptide bond angles and distances (C^α–C′ 1·53 Å; C′–O 1·24 Å; C′–N 1·32 Å; N–C^α 1·47 Å) were used (Schulz & Schirmer, 1979); the dihedral Φ and Ψ rotation angles at the C^α atom were restricted to a set of 7 standard conformations described by Rooman *et al.* (1991) and shown to be sufficient to mimic the backbone topology in a wide range of proteins with known topology. Four conformations (Φ,Ψ) are possible for any of the 20 amino acids: α-helix (−65, −40), 3_{10} helix (−89, −1), β-strand (−117, 142) and extended conformation (−69, 140). Additional conformations were allowed for explicitly defined residue types: 2 for typical backbone topologies of glycine ((78, 20); (103, −176)) and a third for *cis*-proline turns (−82, 133). All main-chain atoms were modeled including the carbonyl oxygen atom (C′,O,C^α,N); side-chains were not explicitly modeled; however, parameters were attached to the C^α atoms in the simulation fitness function such as secondary structural preferences or hydrophobicity values (Manavalan & Ponnuswamy, 1978).

The idealized 4-helix bundle (Argos *et al.*, 1977) used in this work possessed the topology depicted in Figure 1(a) and was characterized as ($a_{10}L_5a_9L_5a_9L_5a_{10}$) where a refers to a helix followed by its length in number of residues and L indicates similarly a loop region. The amino acid residues in the idealized bundle were assumed to be distributed according to a perfect amphipathic wheel (Figure 1(b)); for instance, in a helix of 10 amino acid residues, the distribution of strong hydrophobic residues (A) to other amino acids (a) follows the pattern AaaAaaaAaa (Schiffer & Edmundson, 1967). Besides this standard example, bundles with different loop and helix lengths were examined.

The atomic coordinates of experimentally determined protein structures used for comparison with the simulation results were taken from the Brookhaven data bank (Bernstein *et al.*, 1977). The following Brookhaven files were utilized: 1HMD (hemerythrin from sipunculid worm (Stenkamp *et al.*, 1982)); 256B (cytochrome b_{562} from *Escherichia coli* (Lederer *et al.*, 1981)); 2CCY (cytochrome c′ from *Rhodospirillum molischianum* (Finzel *et al.*, 1985)); and 1CRN (crambin from Abyssinian cabbage (Hendrickson & Teeter, 1981)). Structures resulting from the simulation as well as those experimentally verified were visualized with the protein characterization system of Chelvanayagam & McKeaig (1991). A superposition program originally developed by McLachlan (1979) was utilized to superpose the observed and predicted C^α positions.

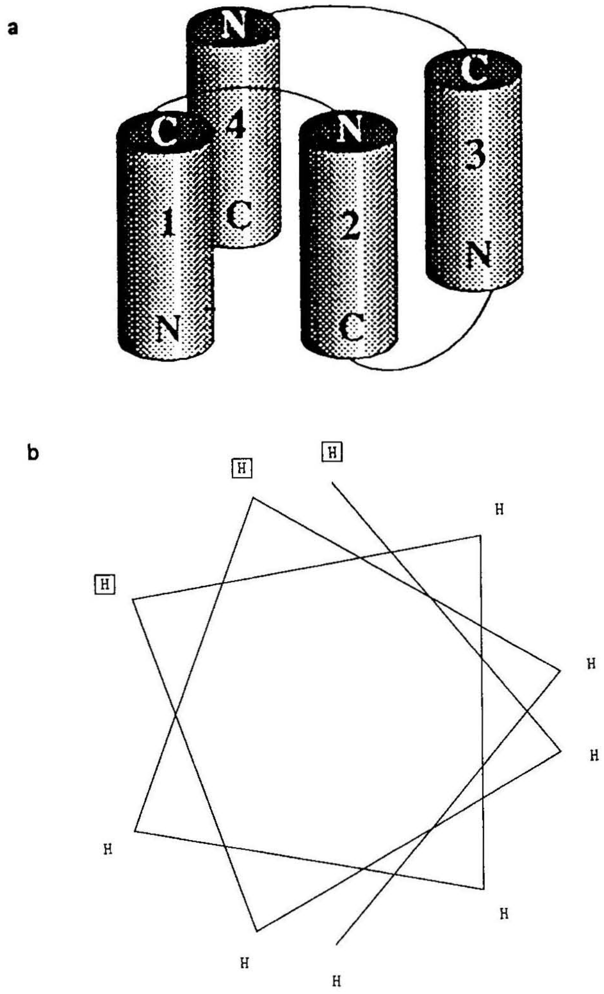

Figure 1. a, Schematic topology of the idealized 4-helix bundle where helices are shown as cylinders with respective N and C termini along the primary sequence. Connecting loops are shown as thin, arched lines. b, Illustration of an α-helical 10-residue amphipathic wheel, which is a projection of C^α atom positions at intersections of lines. The peak hydrophobic residues are boxed.

(c) *Secondary structure prediction*

The simulations of non-idealized proteins used secondary structural predictions as nucleation spans. The prediction technique of Ptitsyn & Finkelstein (1983) was selected as representative, though others could have been utilized with similar overall results. In their approach, local interactions were evaluated from a stereochemical theory based on the relative stabilities of α and β-structures for different residues in synthetic polypeptides. Long-range effects were approximated by the interaction of each chain region with an averaged hydrophobic template.

(d) *Fitness criteria*

Survival of the solution trials in the genetic algorithm simulation is proportional to fitness (Goldberg, 1989). The proper selection criteria are critical for the outcome. The fitness function incorporating the different criteria must reach a maximum positive value. Criteria values are viewed as rewards (plus terms) or punishments (minus

Table 1
Fitness function criteria

Criteria	Des	Term	Specific parameters	Details
Constant	C	$\text{weight}_C \times \text{lchrom}$	$\text{weight}_C = +350$ $l_{\text{chrom}} = 2 \times$ residue length of protein	See eqn (2)
Clash	cl	$\text{weight}_{\text{cl}} \times \sum \sum \text{overlap}(i,j)$	$\text{weight}_{\text{cl}} = -500$	See eqn (3)
Secondary structure	co pf	$\text{weight}_{\text{ss}} \times$ cooperativity $\text{weight}_{\text{ss}} \times$ (struct.pref.-38)	$\text{weight}_{\text{ss}} = +14$ 38 residue test bundle	See eqn (4)
Tertiary structure	ghs	$\text{Weight}_{\text{ghs}} \times$ global hydrophobic scatter	$\text{weight}_{\text{ghs}} = -24$	See eqn (6)
	phs	$\text{weight}_{\text{phs}} \times$ peak hydrophobic scatter	$\text{weight}_{\text{phs}} = -19$ peak $= \{\text{M,I,L,V,Y,C,F}\}$	
local burial of hydrophobics	lb	$\text{weight}_{\text{lb}} \times \sum_{\text{peak}}$ closest neighbor	$\text{weight}_{\text{lb}} = -5$ peak $= \{\text{M,I,L,V,Y,C,F}\}$	See the text
hydrogen bonds	hy	$\text{weight}_{\text{hb}} \times \sqrt{\sum (\text{bondclass } (j))^2}$	$\text{weight}_{\text{hb}} = +10$	See eqn (5)

The term Des refers to an abbreviated designation for the involved criterion. The detailed calculation of each parameter is described in Methods. For the optimal case, the standard fitness function included the terms $C + \text{cl} + \text{co} + \text{pf} + \text{ghs} + \text{phs}$.

terms). The fitness function for the idealized 4-helix bundle simulation (optimized criteria and weights, Table 1) was composed by criteria relying on steric overlaps and secondary and tertiary structural characters.

An optimized positive constant C (see below) was added to the calculated fitness value:

$$\text{fitness} = C + \text{clashes} + \text{secondary structure} + \text{tertiary structure}. \quad (1)$$

The constant is used to normalize fitness values such that about 10% of the random population has no positive fitness in the first generation and thus selection is sufficiently strong for evolution to progress rapidly but not so dominant that the population becomes homogeneous where only a few fit individuals, unable to change further, survive (Goldberg, 1989). The same value for the constant is used throughout the simulation but needs to be adjusted with the protein or encoding chromosomal length (l_{chrom}), since the number of residues affects the absolute values of the remaining criteria. The constant used in the optimized (see Results) fitness function was:

$$C = +350 \times l_{\text{chrom}}. \quad (2)$$

Residue clashes were considered to occur if main-chain C^α atoms, n in total, overlapped; i.e. if they were closer than 3·8 Å, which is the idealized distance in protein structures (Schulz & Schirmer, 1979). The clash criterion was defined as:

$$\text{clash} = \text{clashweight} \times \sum_{i=1}^{n} \sum_{j=i}^{n} \text{overlap } (i,j) \quad (3)$$

where

$$\text{overlap } (i,j) = 1 \text{ if } \sqrt{\sum_{k=x,y,z} (C^\alpha(res_i)_k - C^\alpha(res_j)_k)^2} \leq 3{\cdot}8 \text{ Å};$$

otherwise, overlap $(i,j) = 0$.
An exception is the first clash found in a trial structure where the overlap value is taken as 3, which is critical to remove the final clash in further stages of the simulation. The optimal clashweight for the idealized 4-helix bundles is -500 (see Results).

Finding suitable criteria for secondary structure and tertiary structure was more complex. The subparameters and weight values given here were found by many different simulation trials (see Results). They worked best for finding the topology of the standard structure; the effects of choosing other parameters or weights are illustrated in Results.

Two terms were chosen to reflect the desired secondary structural content and sequence positioning; namely, cooperativity and secondary structure preference such that:

$$\text{secondary structure fitness} = +14 \times (\text{cooperativity} + \text{secondary structure preference} - 38\ ). \quad (4)$$

The values for the secondary structure preference for each residue (1 if the preferred and trial state were the same, otherwise zero) were added and then 38 was subtracted to normalize relative to the small idealized 4-helix bundle protein with 38 α-helical residues. The resulting value summed with cooperativity fitness was then multiplied by +14 to allow sufficient weight value relative to the tertiary structure fitness.

"Cooperativity" of 2 successive residues, which by virtue of being in the same conformation initiate that state independent of structural type, was set to 10, which was then increased by 1 for every further and successive residue in the same conformation and in the C-terminal direction. The cooperativity values were then summed for each initiation site, yielding the final cooperativity value in the fitness function.

Nucleation regions of secondary structure were tested in the simulations. The nucleating residues were kept fixed in their conformational state. If such regions were derived from predictions based on the actual protein sequence, flanking residues were rewarded with a preference value of 20 if they had the same secondary structural state as that predicted in the nucleation region. Cooperativity for succeeding residues in the same structural state was rewarded as previously explained (for any state chosen and also in loop regions). No extension was allowed to be closer than 3 residues from a neighboring nucleation region or an extension of it. In the idealized 4-helix bundle simulations without pregiven secondary

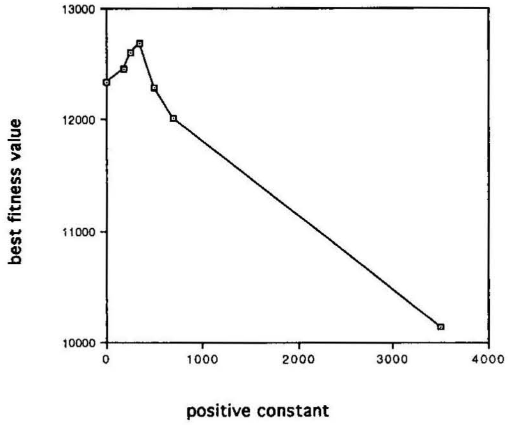

Figure 2. Plot of best fitness value obtained at the end of a simulation *versus* the positive constant C used as an additive term in the fitness function. At the optimal value (+350), 10% of the random start population had a negative fitness. Too harsh a selection (high negative constant added) would yield the fittest individual with a negative fitness while too many survivors (high positive constant added) would not allow evolution to proceed with sufficient speed (Goldberg, 1989). Non-optimal constants yield folds such as those illustrated in Figures 4(d) and (e).

structure, residues of high helix-forming potential were modeled by a fitness value of 20 if their state in the simulation trial was α-helical.

Hydrogen bonds were also tested as a selection criterion. Their formation was judged by the distance between the main-chain carbonyl oxygen atom of one residue and the main-chain nitrogen atom of another residue. In the selection scheme, the closest nitrogen atom to each main-chain carbonyl oxygen atom was taken and the value of all such distances for all backbone oxygen atoms. The sum was in turn multiplied by a negative weight such that simulation trials with larger distances were penalized by subtraction from the fitness value. A second selection scheme emphasized hydrogen bonding further by using the square of the distances in the sum.

Since hydrogen bonds often occur in repeating patterns such as in α-helices between the (i)th and $(i+4)$th residues, another selection procedure was investigated that rewarded formation of hydrogen bonds with a positive weight$_{\mathrm{hb}}$; in this case, the bond length had to be closer than 10 Å and all possible bond classes for a residue i to a residue $(i+j)$ spanning j residues were considered; i.e. i to $i+2$, i to $i+3, \ldots$ i to $n-2$ where there are n total residues in the protein. The 10 Å value was used as it is slightly larger than 1 hydrogen bond distance of 3 Å plus the 3 radii of the carbonyl carbon, oxygen and main-chain nitrogen atoms; above this distance the hydrogen-bonding forces should be negligible.

Hydrogen bond fitness =

$$\text{weight}_{\mathrm{hb}} \times \sqrt{\sum_{j=1}^{n-2} (\text{bondclass}(j))^2}. \quad (5)$$

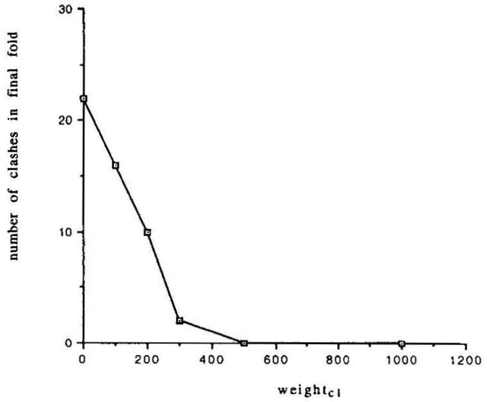

Figure 3. Plot of the number of clashes observed in the final fold of the simulation *versus* different values of a positive clash weight. A clash value of −500 is optimal. A very strong negative weight leads to non-compact structures; selection against clashes must nevertheless be sufficiently strong to remove all of them during the simulation.

Bondclass(j) is the number of acceptable hydrogen bonds spanning j residues in the protein. A further selection scheme simply rewarded each $(i,i+4)$ bond formed to encourage helix formation.

Tertiary structure fitness was composed of 2 terms; namely, global hydrophobic scatter (ghs) and peak hydrophobic scatter (phs):

tertiary structure fitness =

$$(\text{weight}_{\mathrm{ghs}} \times \text{global hydrophobic scatter}) + (\text{weight}_{\mathrm{phs}} \times \text{peak hydrophobic scatter}). \quad (6)$$

Global hydrophobic scatter was calculated by the overall distribution of residues around the center of mass (c_{m}) with coordinates $(x_{c_{\mathrm{m}}}, y_{c_{\mathrm{m}}}, z_{c_{\mathrm{m}}})$ such that:

global hydrophobic scatter =

$$\sum_{i=1}^{n} \sqrt{\sum_{k=x,y,z} (\mathrm{C}^{\alpha}(res_i)_k - k_{c_{\mathrm{m}}})^2}. \quad (6a)$$

Loop regions, predicted (only defined if there were at least 3 secondary structure elements predicted) or observed, were excluded from the calculation of the scatter criteria. The optimized weight$_{\mathrm{ghs}}$ was −24 (see Results) such that, the larger the scatter, the less able is the fitness function to achieve its maximum positive value. The scatter of the peak hydrophobic residues {M,I,L,V,Y,C,F} (according to Manavalan & Ponnuswamy, 1978) was amplified by the weight$_{\mathrm{phs}}$ with value −19. A lower weight did not result in the expected final fold for the idealized 4-helix bundle, while a higher weight distorted the evolving structure by blocking movement of secondary structures containing 1 or more hydrophobic residues near the protein center. The latter rule incorporates in a straightforward way information from the sequence into the tertiary structure fitness. These criteria and the complete optimized fitness function are summarized in Table 1.

Figure 4. *Ab initio* folding simulation of a small idealized four-helix bundle as described in Methods. A representative individual is shown from the random start population (a), and fittest individuals are shown after (b) 1, (c) 10, (d) 30, (e) 100 generations and (f) for the optimal fold at the simulation end. Virtual bonds connecting successive C$^{\alpha}$ atoms are given as connecting lines. Double images of each case are shown for a stereo perspective. Terminal residues are indicated as Gly.

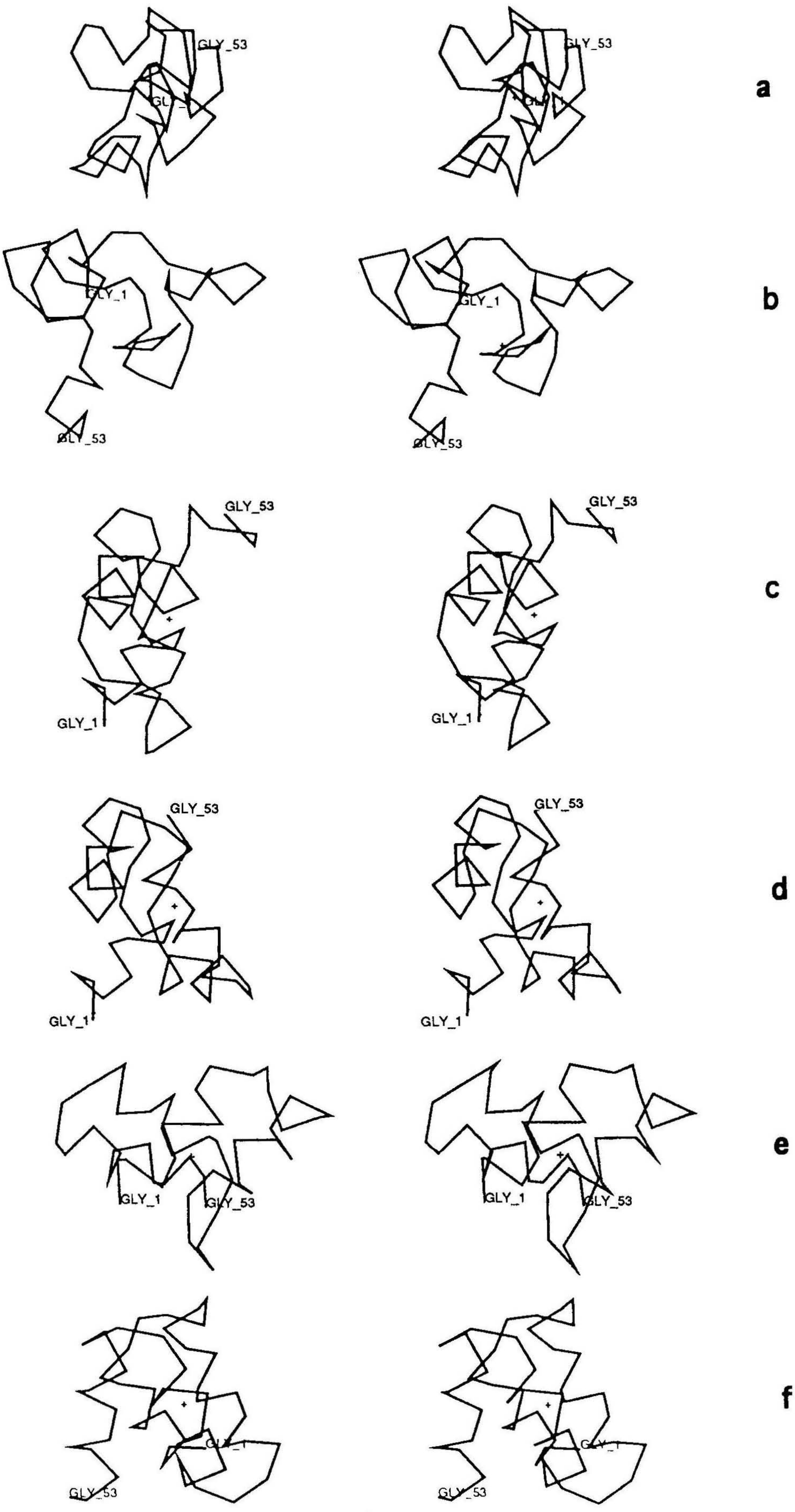
GLY_53
GLY_1
a
GLY_1
GLY_53
b
GLY_53
GLY_1
c
GLY_53
GLY_1
d
GLY_1
GLY_53
e
GLY_1
GLY_53
f

Fig. 4.

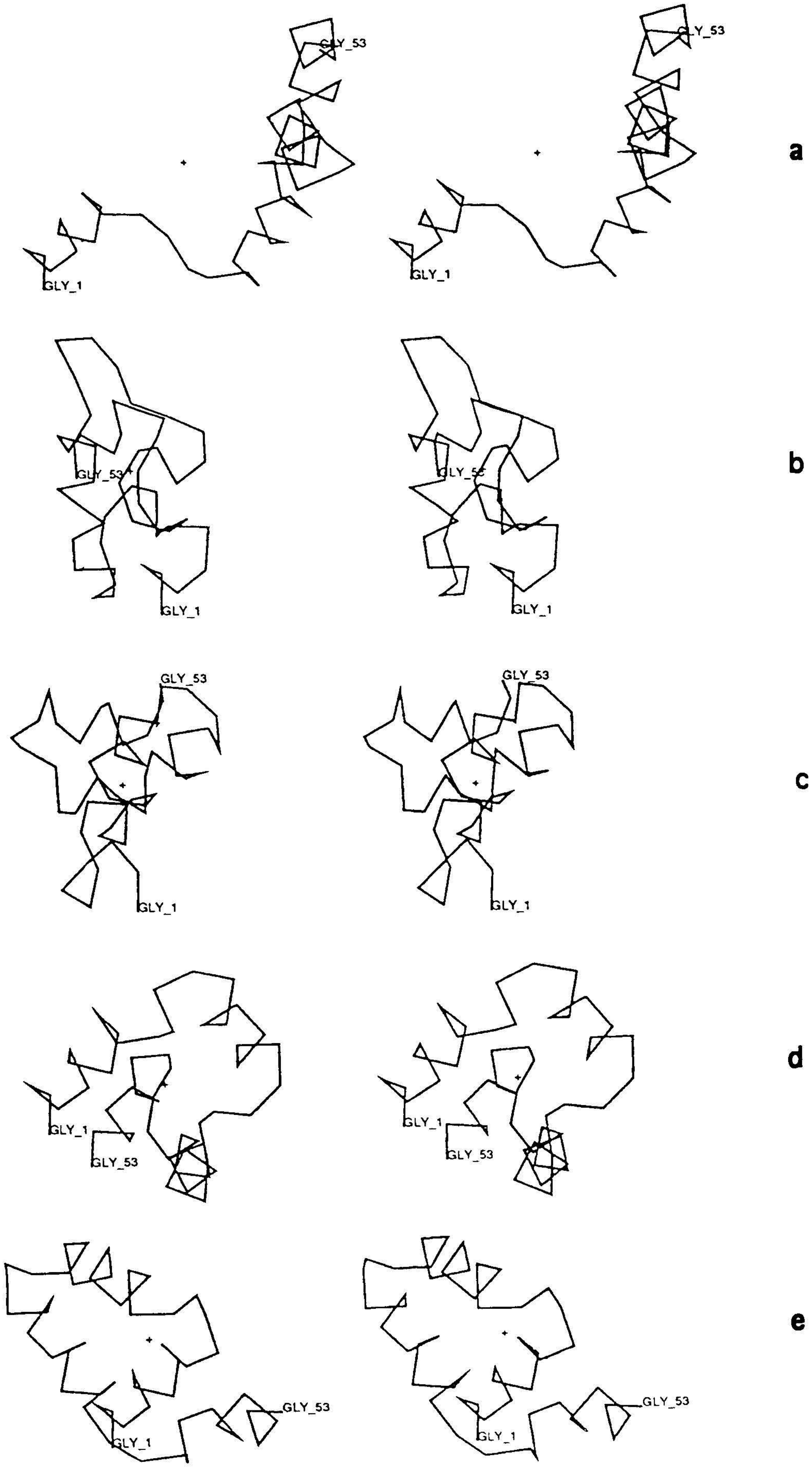
GLY_53
GLY_53
GLY_1
GLY_1
a
GLY_53
GLY_53
GLY_1
GLY_1
b
GLY_53
GLY_53
GLY_1
GLY_1
c
GLY_1
GLY_53
GLY_1
GLY_53
d
GLY_53
GLY_1
GLY_53
GLY_1
e

Fig. 5.

3. Results

(a) *Establishing constant and clash fitness terms*

The optimal value for the positive normalizing constant C of the simulation fitness results in the worst individuals having a negative fitness in the random start population; a value of +350 is best for the application in the plot of Figure 2 where different values are explored. The constant has been substracted from each four-helix simulation result for comparison. Similarly the optimum for the clash value can be found in that clashes should not be tolerated in a final fold. The clash value has to be sufficiently large to allow selection of structures that have no atom overlap and yet are compact, similar to natural proteins (Schulz & Schirmer, 1979). In Figure 3 a plot of the number of clashes in the final fold *versus* clashweight is given for the idealized four-helix bundle; clearly, 500 is optimal, which just removes the last clash. A larger value would open the structure and not maintain the required compactness.

Figure 4 shows the grid-free *ab initio* folding of an idealized four-helix bundle $a_{10}L_5a_9L_5a_9L_5a_{10}$ (a = helix and L = loop and the number following indicates the residue length) using the fitness parameters detailed in Methods. There is no pregiven helical or loop nucleation site in this simulation, but helical conformations in expected regions were rewarded by a positive fitness payoff (see Methods); however, other main-chain conformations were allowed and could result if a higher total fitness value could be achieved for the overall protein conformation. The optimized fitness function detailed in Table 1 was used. In Figure 4 an individual from the random start population is shown first. In the next generations the removal of clashes is the most prominent feature of the simulation; after this, formation of secondary structure becomes more apparent and in the later stages of the simulation the tertiary fold appears.

(b) *Testing different folding forces in idealized bundles*

The folding of the idealized four-helix bundle can be exploited to test dominant forces in protein folding. The following theoretical models or fitness criteria (see Methods and Table 1) were investigated to test for their potential importance in protein folding.

(i) *Hydrophobicity*

The dominance of hydrophobicity was considered through the global hydrophobicity scatter terms in the fitness function where only loop residues of the simulation trial were excluded.

(ii) *Burial of hydrophobic residues*

The burial of the most hydrophobic residues was tested under two conditions. The global scatter was considered as in (i) but only for peak hydrophobic amino acids {M,I,L,V,Y,C,F} in the sequence and with a specific weight$_{\text{phs}}$ (Table 1). Another fitness term based on the local burial (lb) of peak hydrophobic residues involved a punishment summing distances between each peak hydrophobic C^{α} atom and its nearest peak hydrophobic C^{α} atom in the given simulation without repeating any given pair. Associated with this term was also a weight designated weight$_{\text{lb}}$ (Table 1). Alternative selections investigated for this criterion involved punishment by the summation of squares of respective distances or implementations rewarding buried pairs closer than a given cutoff distance.

(iii) *Secondary structure nucleation*

Some residues were kept fixed in a particular secondary structural conformation to test the importance of nucleating sites in achieving a proper main-chain fold. Fitness weights for residues immediately flanking such a region were set to 20 if an identical state was assigned. Together with the cooperativity rule, secondary structural regions were thus allowed extension providing there was no counterselection, notably by clash or tertiary structural criteria. The nucleation spans could cover an entire helical region, for instance, or only a portion of the known secondary structural span such as the N-terminal one-third. The latter was used to test for the minimal length required for such regions and the ability of the cooperativity rule to extend appropriately the substructure or to overcome errors in secondary structure prediction.

(iv) *Hydrogen bonds*

Hydrogen bonds (hb) could also be critical in promoting the tertiary fold. Accordingly, the tertiary structural terms in the fitness function (Table 1) were appropriately replaced.

(c) *Significance of folding forces*

Several model representations were tested to elicit the significance of the just described characters in protein folding simulations. Several combinations of the criteria were used in fitness functions (Table 1) as well as several weight values. Even the ideal four-helix bundles were allowed individual helical lengths

Figure 5. Fittest fold (stereo view) from simulations of an idealized 4-helix bundle using different hydrophobic criteria. Conditions investigated include: (a) no hydrophobic forces; (b) very strong hydrophobic forces; (c) strong local hydrophobic forces; (d) weak local hydrophobic forces plus global hydrophobic forces; and (e) optimal global hydrophobic forces. Table 1 shows the exact fitness criteria used for each simulation. Virtual bonds connecting successive C^{α} atoms are shown as connecting lines. Terminal residues are given as Gly.

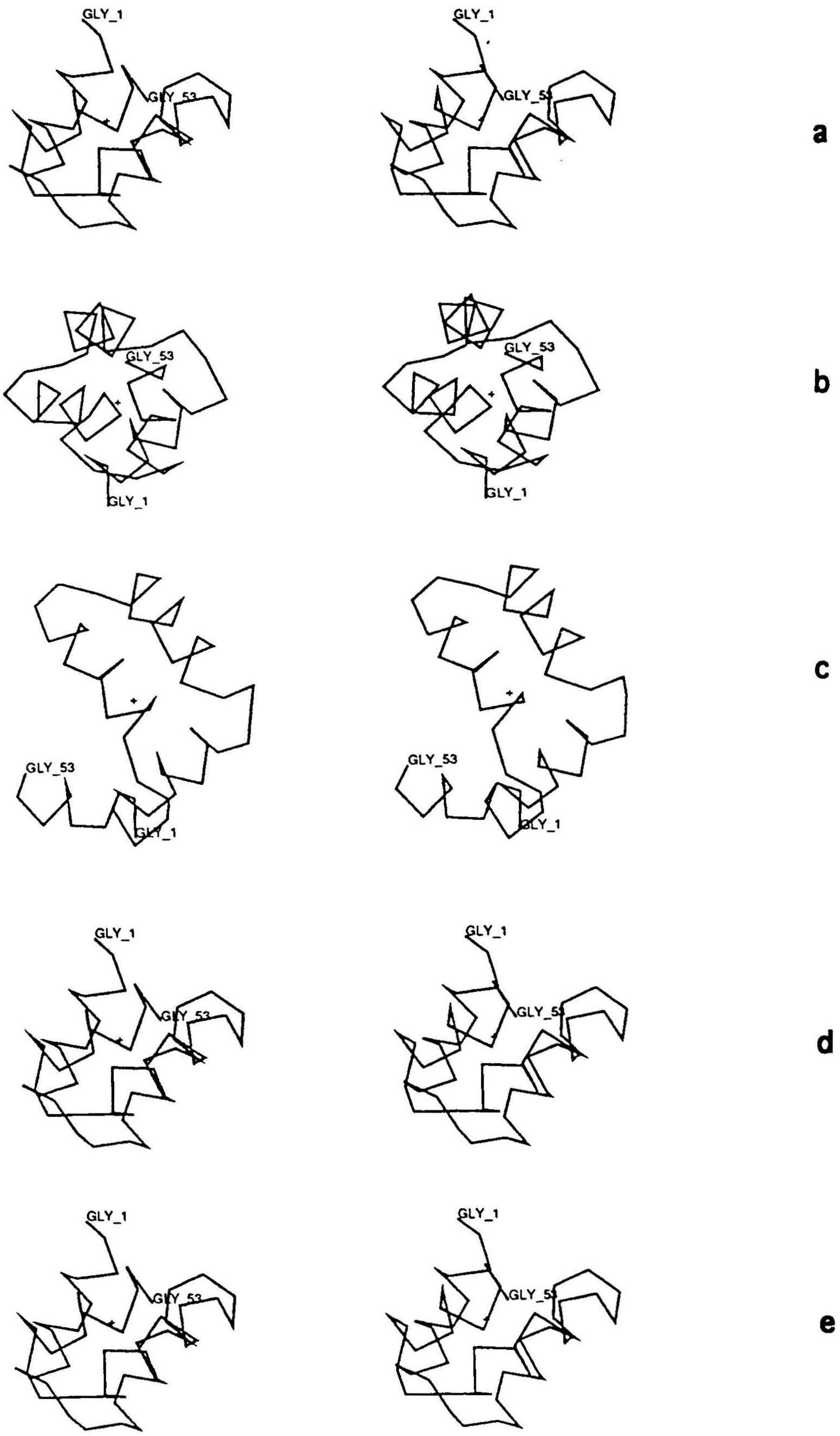

Figure 6. Fittest fold (stereo view) from simulations of an idealized 4-helix bundle using different hydrophobic criteria but always including fixed secondary structural conformational states. The criteria used in the fitness function included: (a) global and local hydrophobic forces; (b) very strong local hydrophobic forces; (c) no peak hydrophobic residues; (d) very strong peak hydrophobic residues; and (e) optimized selection (Table 1) with secondary structure nucleation sites. Virtual bonds connecting successive C^{α} atoms are shown as connecting lines. Terminal residues are shown as Gly.

Table 2

Effect of different fitness parameters and weights

Fitness function	Model	Representative fold
A. *Test of basic parameters*		
x^*C + cl + co + pf + ghs + phs	Survival fraction	Figure 2
C + x^*cl + co + pf + ghs + phs	Clash removal	Figure 3
C + cl + co + pf + ghs + phs	Standard simulation	Figure 4
B. *Tertiary structure selection*		
C + 2cl + co + pf	No tertiary selection	Figure 5(a)
C + 2cl + co + pf + 5ghs + 5phs	Strong global hydrophobicity	Figure 5(b)
C + 2cl + co + pf + 10lb	Strong local hydrophobicity	Figure 5(c)
C + 2cl + co + pf + ghs + phs + lb	Local and global hydrophobicity	Figure 5(d)
C + 2cl + co + pf + ghs + phs	Standard global hydrophobicity	Figure 5(e)
C. *Secondary structure nucleation*		
C + cl + co + pf [+nuc] + ghs + phs + lb	Local and global hydrophobicity	Figure 6(a)
C + cl + co + pf [+nuc] + 10lb	Strong local hydrophobicity	Figure 6(b)
C + cl + co + pf [+nuc] + ghs + lb	No peak hydrophobicity	Figure 6(c)
C + cl + co + pf [+nuc] + ghs + phs	Strong hydrophobicity	Figure 6(d)
C + cl + co + pf [+nuc] + ghs + phs	Standard with nucleation	Figure 6(e)

The short designation for the criteria used and their weights are explained in Table 1. The term x^* denotes various tested weights as plotted in Figures 2 and 3. The term [+nuc] is not an additive one in the fitness function but denotes simulations with fixed secondary structure nucleation sites; in these simulations over all generations only individuals with appropriate secondary structural conformation were allowed to remain in the population. At least 10 random start populations and subsequent simulations were tested for each condition. Typical optimal individuals at the end of a simulation are shown in Figures 5 and 6.

ranging from 8 to 16 residues and loop lengths from 3 to 6 residues.

The results are illustrated in Figures 5 and 6 as well as in Tables 1 and 2 for the idealized bundle $a_{10}L_5a_9L_5a_9L_5a_{10}$. The peak hydrophobic residues were assigned with alternate spacings of two and three residues in length to satisfy the perfect amphipathic condition (Schiffer & Edmundson, 1967) illustrated in Figure 1. Table 1 lists the terms, weights and constants of the optimized fitness function and of those with investigated alternatives. The various weights tested included factors of 10, 5, 2, 0·5, 0·2 and 0·1 relative to the optimal weights listed in Table 1. Near-optimal weights, notably for clash and tertiary structure, were fine-tuned in smaller steps. Table 2 shows various combinations of criteria used in the fitness function; in each case ten random starting folds were attempted for each fitness combination. Figure 5 illustrates representative structures of the fittest achieved in simulation sets of ten where all folding forces but secondary structural nucleation sites and hydrogen bonds were applied with various weights. Clearly the fold in Figure 5(e) is closest to that of an idealized bundle, which contains four helices with nearly parallel axes as illustrated in Figure 1. This fold was achieved by using criteria based on an additive constant, elimination of clashes, preferred secondary structural states for the entire helical length, cooperativity for any state and residue site, and global hydrophobic scatter with loop residues excluded. The bundle is somewhat open as the clash value in the simulations was doubled (in comparison to Table 1) to counteract the strong hydrophobic force component used in trials illustrated in Figure 5(b) and 5(c). Eliminating hydrophobic forces leads to a fully open structure (Figure 5(a)); secondary structures nonetheless appear as they are formed during the selection. Simulations where the weight for secondary structure fitness is increased and that for tertiary structure fitness correspondingly decreased evolve similarly with mixed orientations of the helical axes; yet, the secondary structures are stable and already observed in early generations. In contrast, enhancing the weight by a factor of 5 for global hydrophobicity relative to the optimal value yields a compact structure but with distorted helices and remaining clashes (Figure 5(b)); enhancing the secondary structure preference and clash weights provides a better force balance such that structures similar to Figure 5(e) are achieved. The effect of local hydrophobic forces was also investigated such that a fitness value was given for locally close hydrophobic residues. The outcome for all such trials was similar: local hydrophobic forces are not sufficient to find the proper fold and distort the structure if accompanied by a strong weight (Figure 5(c)). Nonetheless, they do promote the overall fold in conjunction with global forces if they are given a low weight (Figure 5(d)).

Secondary structure nucleation regions clearly promote the fold; this effect is illustrated in Figure 6. The helix regions in these simulations of the idealized bundle were kept fixed in the helical state while all other fitness parameters and simulations were similar to those of Figure 5. The standard bundle with appropriate shape is now achieved providing global and peak hydrophobic forces are also used (Figure 6(e)). Without peak hydrophobicity the structure opens (Figure 6(c)). Replacing global hydrophobicity by local hydrophobic forces again distorts the structure (Figure 6(b)). Moderate

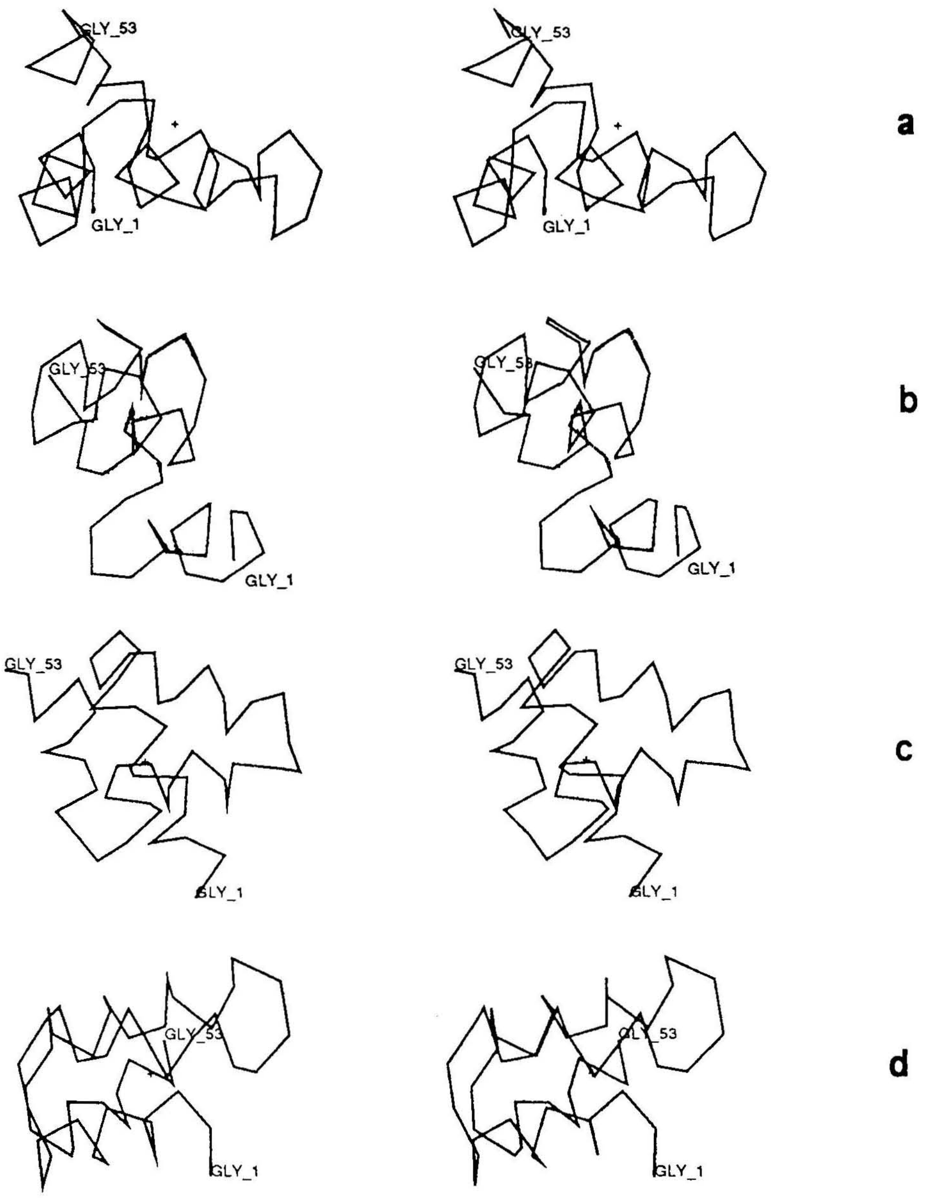

Figure 7. Illustrative backbone topology (shown in stereo) using C-terminal residues as nucleation sites for each helix in an idealized bundle. A representative individual was taken (a) from the random start population, (b) 10 generations later, (c) 30 generations later and (d) after 60 generations. Virtual bonds connecting successive C^{α} atoms are shown as connecting lines. Terminal residues are given as Gly.

increase of the global hydrophobicity to the standard weight (Table 1) did not change the structure (Figure 6(d)) nor did no or low local hydrophobic forces (Figure 6(a)).

The influence of nucleation length within secondary structures was also investigated. Different portions of the helices were fixed: three residues in the middle or at the C or N-terminus of each helix of the bundle. These nucleation sites within secondary structure proved to be effective in promoting folding if they were at the termini of the helices. With C-terminal helices as nucleation sites, the fold proceeded within 60 generations to the four-helix bundle (Figure 7); N-terminal residues were slightly less effective (100 generations required). Middle nucleation was much less efficient in that a bundle-like structure resulted only after 500 generations. Hydrogen bonds were equally ineffective to guide folding under all schemes (see Methods). Neither global nor local (selection for bonds in certain regions) maximization of hydrogen bonds improved the evolution toward the four-helix bundles. Helical regions could be formed, but only by direct selection for (i, $i+4$) helical bonds. Still a proper bundle fold was not achieved; illustrative folds were similar to that shown in Figure 5(b). The optimal or standard fitness function (Table 1) is thus $C+\mathrm{cl}+\mathrm{co}+\mathrm{pf}+\mathrm{ghs}+\mathrm{phs}$ and corresponds to the simulation results shown in Figure 4; with addition of fixed regions of secondary structure nuclea-

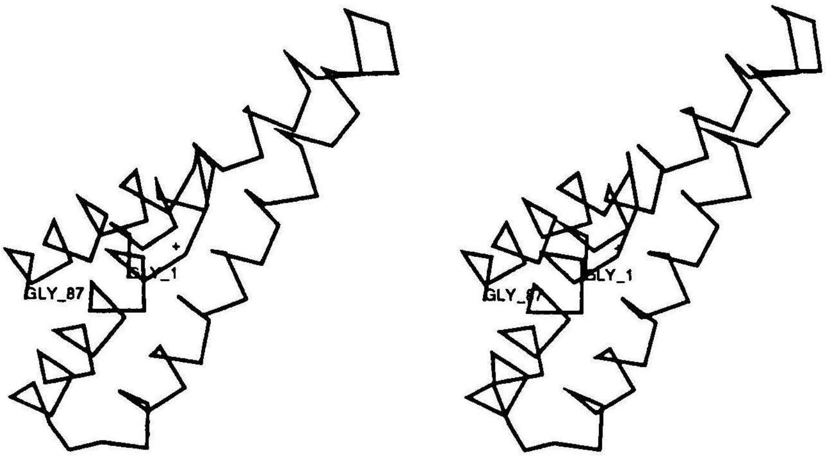

Figure 8. Stereo views for the folding of an idealized hemerythrin. The exact length of the observed helices and loops are used and hydrophobic residues are distributed ideally according to an amphipathic model. Virtual bonds connecting successive C^{α} atoms are shown as connecting lines. Terminal residues are indicated as Gly.

tion, results as illustrated in Figure 6(e) are achievable.

(d) *Folding of non-idealized four-helix bundles*

Given the promise of the idealized bundle simulations, we investigated the folding of experimentally and structurally characterized four-helix proteins using the optimized fitness parameters determined from the idealized case (Table 1). As a first step the idealized length of the four helices and terminally connecting loops were replaced by those of hemerythrin (from sipunculid worm (Stenkamp *et al.*, 1982)). However, the distribution of hydrophobic amino acid residues was maintained according to that of the idealized hydrophobic wheel (Figure 1) and without correspondence to the hemerythrin primary structure. These folding trials were successful and illustrate the ability of the genetic algorithm to fold bundles composed of different helical lengths (Figure 8). It must be emphasized that the fitness criteria involving fixed secondary structural sites and sequence information to elicit peak hydrophobic residues were essential in achieving the ideal hemerythrin fold. Without the latter, the resulting fold was considerably distorted as in Figure 5(c); without secondary structure nucleation sites, incompletely folded or incorrect structures resulted as in Figure 5(b). Thus the basic folding parameters used in the idealized case are also essential in identifying a real structure and knowledge of the sequence and secondary structure elements are fundamental.

We next tested if the absence of idealized information (ideal helix amphipathicity and knowledge of the secondary structure elements as elicited from the tertiary structure) and substitution of readily available sequence information and secondary structure predictions from it was sufficient to identify the correct fold. The forces and fitness criteria can in principle delineate boundaries of different secondary structures and can partly correct missing predictions through hydrophobic forces and cooperativity. From the primary sequence alone, it is possible to delineate strongly hydrophobic patterns and to predict secondary structural spans, albeit with some inaccuracy. Three different four-helix bundle proteins with known structures were tested to check the consistency of the genetic algorithm performance. In a single trial only the structure corresponding to the fittest individual obtained during the entire simulation was used for the prediction of the fold. Regions predicted by the representative technique of Ptitsyn & Finkelstein (1983) as helix or β-strand, independent of their correspondence with observed structures, were taken as fixed nucleation centers in the simulations. Their method was chosen since it is based on molecular theory and model structure templates, and thus is not biased by the particular sequence statistics of the proteins tested here. Three states (helix, turn and β-strand) are predicted by their approach and, if a given sequence state has insufficient probability, no prediction is made. The secondary structure predictions for cytochrome c', hemerythrin and cytochrome b_{562} together with the observed secondary structure and the distribution of peak hydrophobic residues are shown in Table 3. The fitness function used was that given in Table 1. Optimal folds for these proteins found by the simulations are shown for cytochrome b_{562} (Figure 9), cytochrome c' (Figure 10) and hemerythrin (Figure 11). It is noteworthy for hemerythrin that the folding pattern of an N-terminal β-strand, which appears as an attached tail to the bundle, could be achieved despite its length and symmetry-breaking properties. In all these simulations, the evolution succeeded in about one-half of ten random-start trials for each protein (Table 4). The simulation was considered successful if the protein structure found as the optimal or fittest solution had proper helical handedness, proper topology, proper helical orientation such that each helical axis was 20° within that observed, antiparallelity in successive helices

Table 3

Secondary structure predictions of simulation examples

Cytochrome b562

```
  1  50     ....,....1....,....2....,....3....,....4....,....5
SEQ         MRKSLLAILAVSSLVFSSASFAADLEDNMETLNDNLKVIEKADNAAQVKD
                                    AAAAAAAAAAAAAAAAA   AAAAAA
                EEeEEeE  EEE tttU   HhhhHhhHhhhHhHHh       Hhh

 51 100     ....,....1....,....2....,....3....,....4....,....5
SEQ         ALTKMRAAALDAQKATPPKLEDKSPDSPEMKDFRHGFDILVGQIDDALKL
            AAAAAAAAAAAA     AAA       AAAAAAAAAAAAAAAAAAAAAAA
            hHhhHh   U      tttU    tt ttHhhHhhhHhHHHhhHhhhHhH

101 128     ....,....1....,....2....,....3....,....4....,....5
SEQ         ANEGKVKEAQAAAEQLKTTRNAYHQKYR
            AA   AAAAAAAAAAAAAAAAAAAAAA
            h tt U     hhhhHhh    U   U
```

Cytochrome c'

```
  1  50     ....,....1....,....2....,....3....,....4....,....5
SEQ         QQSKPEDLLKLRQGLMQTLKSQWVPIAGFAAGKADLPADAAQRAENMAMV
                AAAAAAAAAAAAAAAAAAAAAAAAAA         AAAAAAAAAAA
             ttttt HHhHhhhHHhhHhhhUU U  U      Hhhhhhhhh  HhHH

 51 100     ....,....1....,....2....,....3....,....4....,....5
SEQ         AKLAPIGWAKGTEALPNGETKPEAFGSKSAEFLEGWKALATESTKLAAAA
            AAAAAAAA             AAA    AAAAAAAAAAAAAAAAAAAAAA
            hhH tTtUttttttUttt      U      HHhhHhhH ttthhHhhhh

101 128     ....,....1....,....2....,....3....,....4....,....5
SEQ         KAGPDALKAQAAATGKVCKACHEEFKQD
            AA AAAAAAAAAAAAAAAAAAAAAA
               hhhHhhhhhhhhhHHhhHhhhH
```

Hemerythrin

```
  1  50     ....,....1....,....2....,....3....,....4....,....5
SEQ         GFPIPDPYCWDISFRTFYTIVDDEHKTLFNGILLLSQADNADHLNELRRC
                       AAA    AAAAAAAAAAAAAAAAAAA   AAAAAAAAAA
             U U  tEEEeEeE  EEeEEthhhhhHHhhHHHH tttttthHhhHhhH

 51 100     ....,....1....,....2....,....3....,....4....,....5
SEQ         TGKHFLNEQQLMQASQYAGYAEHKKAHDDFIHKLDTWDGDVTYAKNWLVN
            AAAAAAAAAAAAAA     AAAAAAAAAAAAAAA      AAAAAAAAAA
                UU    UU    U  U        hHHhhH  Ttt HhHhhhHHHh

101 113     ....,....1....,....2....,....3....,....4....,....5
SEQ         HIKTIDFKYRGKI
            AA    AAA
            hH  EeEeE   U
```

Crambin

```
  1  46     ....,....1....,....2....,....3....,....4....,....5
SEQ         TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN
            BBBB  AAAAAAAAAAAAA   AAAAAAAA BBBB   BBB
              UUttUU    U UU Utt hhhHHhhHttEEEEttt UtttU
```

Secondary structure predictions according to the method of Ptitsyn & Finkelstein (1983) are either β-strand (e or E), α-helix (h or H) and turn (t or T) or are not predicted (blank or U). It must be emphasized that only helical and strand predictions were used as nucleation regions with fixed dihedral angles; the turn predictions shown are only informative. Observed secondary structures (A, α-helix; B, β-strand) are also given. Peak hydrophobic residues according to the scale of Manavalan & Ponnuswamy (1978) are marked by capital letters in the predicted regions (H, E or T) or U if no prediction was given by Ptitsyn & Finkelstein (1983). The conformations of the first 23 residues of cytochrome b_{562} are not known and not used in this work; otherwise the simulations (Figures 9, 10 and 11) included all residues except those at the termini that were not predicted and assumed to be structurally disordered.

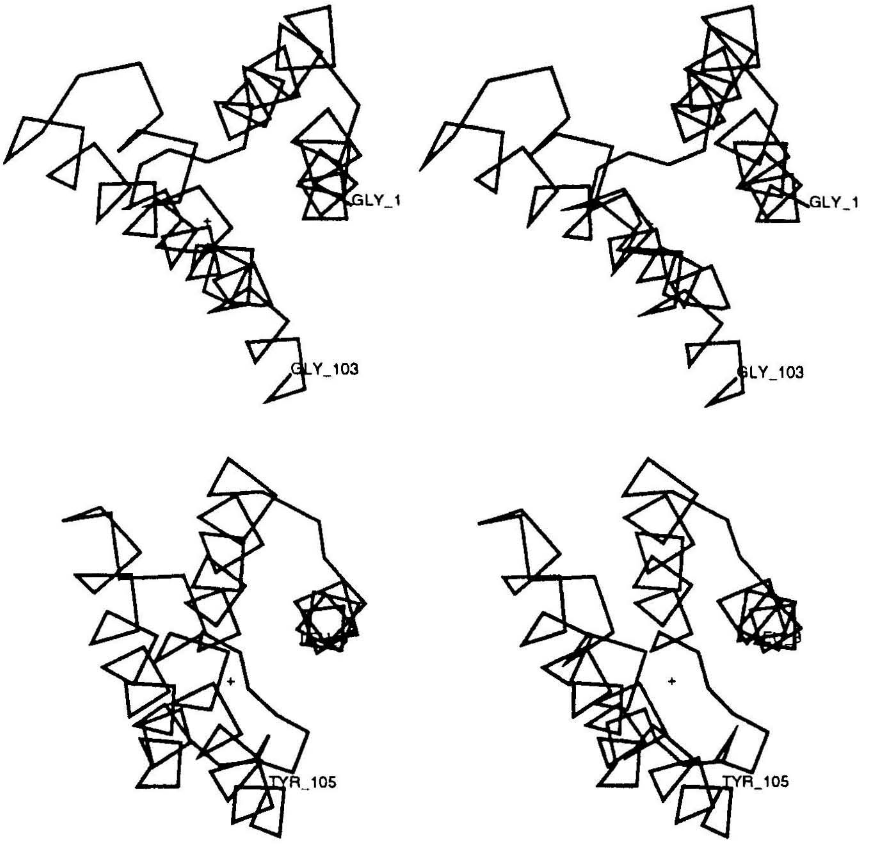

Figure 9. Fittest and observed folds (stereo view) of cytochrome b_{562}. The secondary structure predictions of Ptitsyn & Finkelstein (1983) are used as nucleation sites (Table 3). The top view illustrates the final fold in the simulation while the bottom is that of the experimentally determined fold (corresponding amino acid residues given). Virtual bonds connecting successive C^{α} atoms are shown as connecting lines.

along the sequence and proper length of the secondary structure (not more than 1 or 2 turns longer or shorter than that observed). Ten different simulations based on different random start populations are compared for each protein. The other final simulation states often only missed a turn of one or two of the secondary structural elements and, more importantly for the protein prediction of an unknown fold, would nevertheless be recognized as a helix bundle. An illustration is given in Figure 12 for hemerythrin. The second helix is broken by two missing helical turns in its middle region, preventing sufficient collapse of the bundle. The root-mean-square positional deviation (RMSD) of all the simulated C^{α} atoms from the observed structure is 8·6 Å, considerably larger than those for the fittest folds in the successful simulations (*vide infra*, Table 5).

It is clear that the genetic algorithm can achieve near-proper orientation of the helices as well as asymmetric positioning of β-strand extensions, though none of these features was an explicit part of the fitness function. Even for a simple feature such as the near parallelity of the helix axes observed in the simulations, many other outcomes are theoretically possible. For example, largely skewed helices could minimize the scatter around the center of mass. Only the combined effect of all the fitness criteria plus the prediction of nucleation sites allowed the successful fold. It is also noteworthy that the genetic algorithm was able to delineate near-proper helical and strand lengths despite secondary structure predictions that involved fewer than one-half of the residues in some helices, helices predicted with intervening turns or with non-contiguously predicted regions or overpredictions, or turn predictions at helical termini. However, in no case was at least one nucleating span within an observed helix completely missing.

Table 4

Folding trials for experimentally determined structures

	Folds found	Folds not found
Cytochrome b_{562}	6	4
Cytochrome c'	5	5
Hemerythrin	5	5
Crambin	4	6

The simulation was considered successful if the protein structure found as the optimal or fittest solution had proper helical handedness, proper topology, proper helical orientation such that each helical axis was 20° within that observed, antiparallelity in successive helices along the sequence and proper length of the secondary structure (not more then 1 or 2 turns longer or shorter than that observed). Ten different simulations based on different random start populations are compared for each protein.

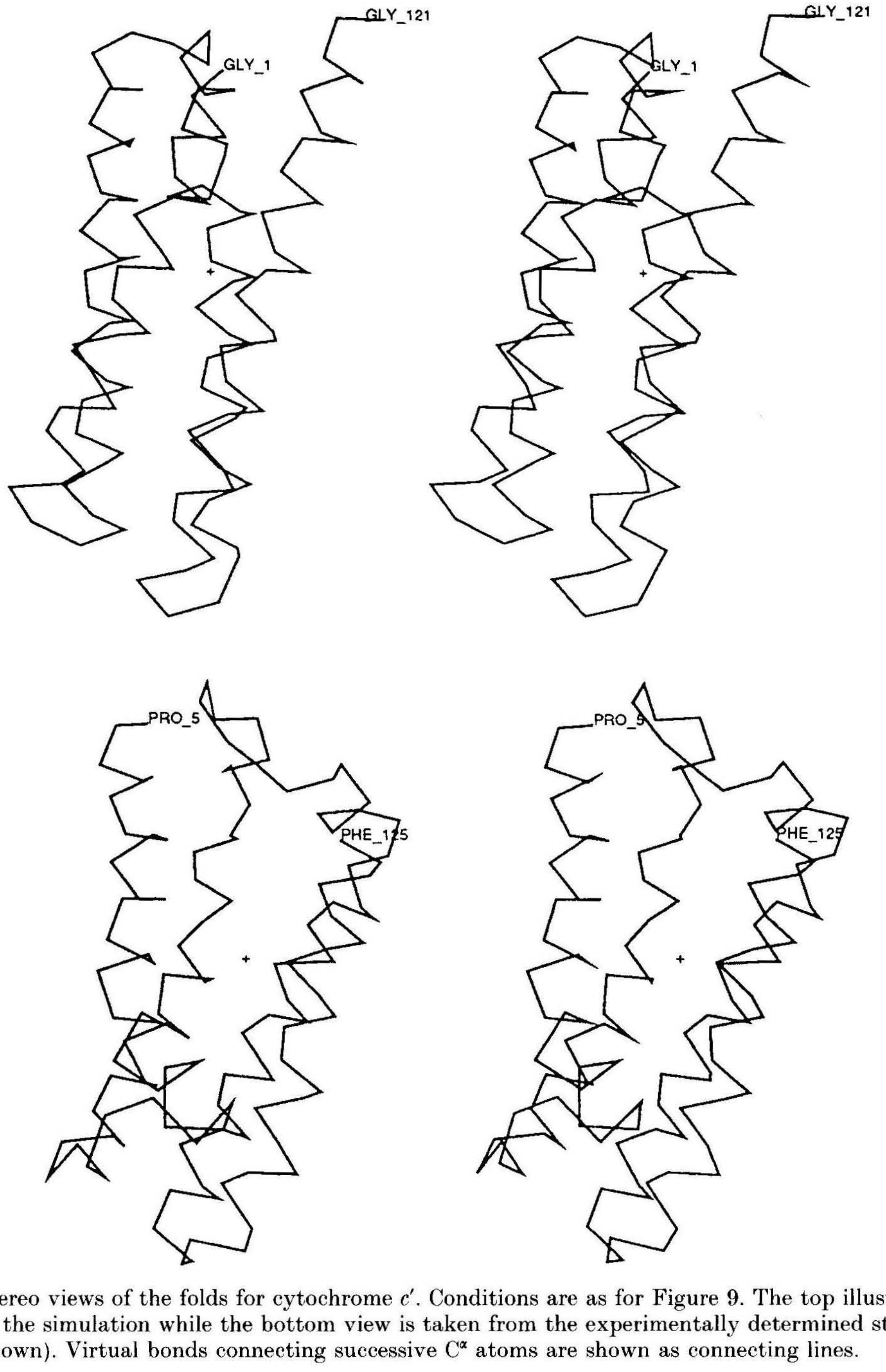

Figure 10. Stereo views of the folds for cytochrome c'. Conditions are as for Figure 9. The top illustration represents the final fold in the simulation while the bottom view is taken from the experimentally determined structure (terminal residue types shown). Virtual bonds connecting successive C^{α} atoms are shown as connecting lines.

The topology of the structures from the successful simulations can be compared with the observed crystallographic structures (Figures 9,10 and 11) by the more quantitative measurement based upon the overall RMSD values between the structurally equivalent C^{α} atoms of the simulated and experimental

Table 5

Comparison of structurally equivalent simulated and observed C^{α} atom positions using RMSD values given in Å.

Protein			Individual helices			
			(from N terminus)			
(fittest fold)	With loops	Without loops	I	II	III	IV
Cytochrome b_{562}	6·14	5·08	0·39	0·56	0·84	1·18
Hemerythrin	6·74	5·63	2·73	1·68	0·44	0·71
Cytochrome c'	6·14	3·87	1·40	1·20	0·41	0·90

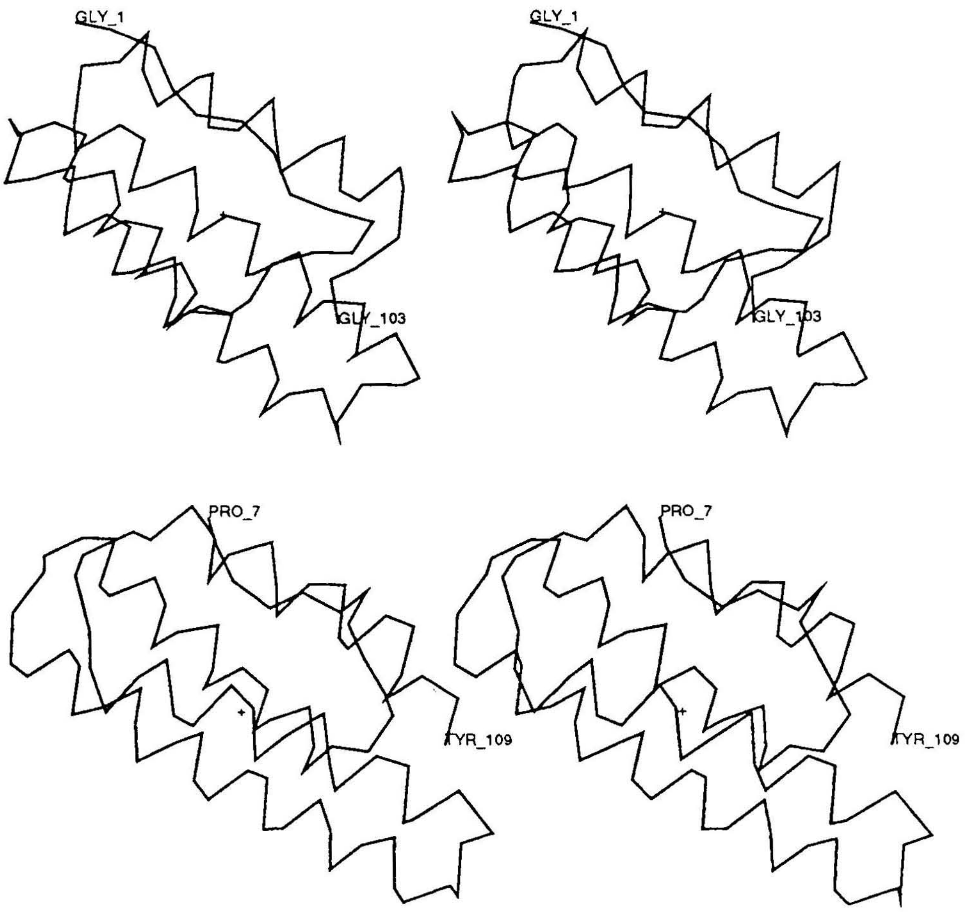

Figure 11. Stereo views of the folds for hemerythrin. Conditions are as in Figure 9. The top view represents the final fold in the simulation while the bottom illustration is that of the experimentally determined structure (terminal residue types shown). Virtual bonds connecting sequential C^{α} atoms are shown as connecting lines.

folds (Table 5). The RMSD values for the total proteins, including loop regions, are around 6 Å while excluding loop regions drops the values to the 4 to 5 Å range. When only the observed single helical regions are compared, the results are much closer to the observed structure.

The ability of the genetic algorithm to respond to the specific sequence associated with a single protein fold is illustrated in the cross comparisons between hemerythrin and cytochrome b_{562}, which have the same number of residues included in the prediction trials and thus allow a direct structural equivalencing of C^{α} atoms (Table 6). It is clear that the simulated and observed folds for a given protein are much closer at around 6 Å than are those, simulated or observed, for different sequences (roughly 12 Å).

The fittest individuals of the successful simulations displayed the lowest RMSD values of its C^{α} atoms relative to the observed structures. For example, in the simulation trials listed in Figure 9 for cytochrome b_{562}, the 6·14 Å RMSD of the optimal fold was followed by 6·91 Å and 7·01 Å deviation for the two next fittest folds in the same trial. For the hemerythrin simulation of Figure 11,

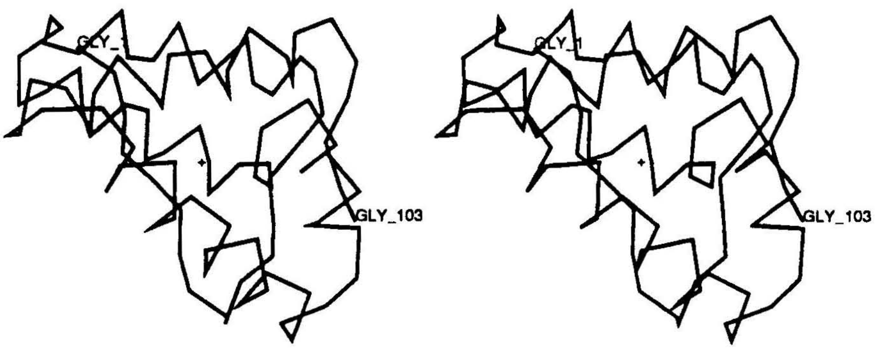

Figure 12. Example of failed simulation trial involving hemerythrin (see the text for a discussion). Virtual bonds connecting sequential C^{α} atoms are shown as connecting lines in the stereo view.

Table 6

Comparison of C^α RMSD (Å) fits amongst observed and optimal simulated folds including loop atoms

	Hemerythrin		Cytochrome b_{562}
	Simulated	Observed	Observed
Cytochrome b_{562} simulated	11·87	11·19	6·14
Cytochrome b_{562} observed	12·86	12·71	
Hemerythrin observed	6·74		

the fittest topology is at 6·74 Å while the two following both display 9·54 Å deviation.

The absolute fitness values associated with the optimal simulation folds of Figures 9, 10 and 11 are compared with the corresponding fitness values calculated directly from the crystallographic architectures; also listed are the average fitnesses from failed simulations for each of the three predicted proteins (Table 7). Fitness calculations for the observed structures assumed secondary structure predictions to be the same as those observed, important for the fitness terms expressed in equation (4). It is evident that the failed folds are the least fit while the observed structures are the most fit with the optimal simulations being only slightly less fit. The failed simulations' inability to better the observed or successfully simulated structures is supportive of the validity of the genetic algorithm approach.

(e) *Folding of a mixed helix/strand protein*

A further example was attempted on crambin where known secondary structure spans are predicted without nucleating segments or in the wrong conformational state. The Ptitsyn & Finkelstein (1983) crambin prediction is shown in Table 3: the N-terminal strand is predicted in part as a turn and the succeeding helix is not predicted. For four of the ten starting populations (Table 4), the simulations produced main-chain structures as exemplified in Figure 13. The N-terminal strand and helix are nonetheless found where the former is used in a sheet-like structure with the more C-terminal strand, though its orientation is not proper. The second crambin helix is not properly oriented though it is maintained and the final C-terminal helix has a reasonable orientation but is not sufficiently distorted as the observed span.

4. Discussion

The most important result of our study is the reasonable identification of the main-chain topology of small proteins using the genetic algorithm with surprisingly simple rules and with only a knowledge of the generally available primary sequence from which secondary structure predictions are made and the distribution of hydrophobic residues noted. Our aim was to achieve the approximate topology of the secondary structure elements; it is clear that their orientations (helical axes now within 20° of that observed) are in need of further refinement, which is likely to result with the use of additional fitness criteria. Tertiary structure is also represented in a very simplistic way by maintaining tight packing, in particular for peak hydrophobic residues, during the whole simulation while secondary structure and the bundle are built up and clashes are removed (Table 7). We stress that the model representation, fitness criteria and their respective weights developed here were optimized for α-helical pro-

Table 7

Comparison of absolute and component values in the fitness functions

	Secondary structure (minus clashes)	Tertiary structure		Total fitness (+C, as in Table 1)
		ghs	phs	
A. *Cytochrome* b$_{562}$				
Start fold	1330	−22,788	−4504	10,088
Generation 10	29,950	−23,064	−4806	38,130
Simulation end	41,316	−20,027	−4071	53,268
Observed fold	42,910	−21,323	−4186	53,451
Failed simulations	38,343	−19,043	−4024	51,326
B. *Hemerythrin*				
Simulation end	32,900	−21,271	−5228	42,451
Observed fold	33,072	−22,006	−4053	43,063
Failed simulations	31,958	−21,629	−4545	41,834
C. *Cytochrome* c′				
Simulation	52,604	−25,782	−6353	62,819
Observed fold	53,480	−26,958	−5371	63,501
Failed simulations	42,010	−24,544	−5706	54,110

The absolute values for the fitness function components (Table 1) as well as the total values obtained in the simulations for cytochrome b_{562}, cytochrome c' and hemerythrin (illustrated respectively in Figs 9, 10 and 11) are given. Values are also given for the observed folds where predicted and observed secondary structures are taken as the same. Further, fitness values averaged over the failed simulation trials are listed.

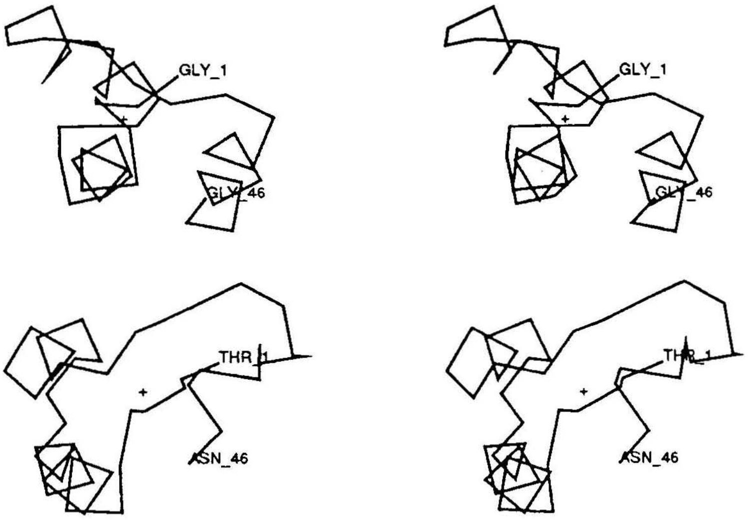

Figure 13. Stereo views of the folds for crambin. Conditions are as for Figure 9. The final fold in the simulation is shown at the top while the experimental structure is given at the bottom where terminal residue types are illustrated. Virtual bonds connecting successive C^{α} atoms are shown as connecting lines.

teins. β-Strands are also amphipathic and can be predicted with nearly the same accuracy as α-helices, thus providing nucleation sites for our genetic algorithm. However, since there are no simple and small β-strand rich structures and since most display disulfide bonds, we chose helical bundles for our initial folding trials. Our genetic algorithm-based strategy can be applied to the prediction of mixed (helical and strand) proteins as illustrated by our results on crambin. However, the results with the mixed model are not as successful using the fitness function optimized for small all-helical folds. Thus, prediction of β-strand-rich and mixed topologies will likely require additional fitness criteria and suitable weights. Such fitness terms could include generally available experimental data such as cysteine residues involved in disulfide bridging, crosslinking data, spatial proximity of catalytic residues, monoclonal antibody epitopes (exposed peptides) and distance constraints from cofactor ligands (e.g. 2 histidine residues for the globin heme). Further, more structurally resolved fitness terms incorporating, for example, residue size, shape and contacts can be used to improve the topology predictions.

It is possible to incorporate exact biophysical force fields and energy calculations into the genetic algorithm's fitness function to model all main and side-chain atoms with the aim of predicting tertiary structure with low atomic coordinate deviation from that of an experimentally determined structure. An exciting and valuable approach to use the genetic algorithm in this area involves the work by Sun (1993). However, such a technique is calculation intensive and appears for the present to be feasible only for very small proteins such as mellitin. In contrast, focussing on secondary structures and their spatial orientation can be effective for larger and less-resolved structures and is much easier to apply. After the main-chain trace has been sufficiently outlined, it is possible to utilize homology modeling techniques to place the side-chains based on full energy calculations (see Eisenmenger *et al.* (1993) for a review) and in this way attempt to bridge the gap from secondary structure prediction to a final protein fold. No doubt such methods are still in need of further development, especially regarding some relaxation of the assumed but partly erroneous main-chain fold.

The basic topology of the proteins examined in this work could be delineated using a small number of simple key forces focussing on global hydrophobic packaging and compactness, secondary structural nucleation and cooperativity of successive residues in its extension. The illustrative folding pathway given in Figure 4 agrees with recent nuclear magnetic resonance experiments (Baldwin & Roder, 1991) on folding intermediates, which suggest secondary structural nucleation in the early folding phase, followed by extension and association of the substructural spans, first locally and then globally. The effectiveness of such simple forces in eliciting the main-chain topology is also consistent with the molten globule (Kuwajima, 1989), which is characterized by secondary structure formation and association yet lacks specific side-chain interactions, thus implying that the backbone fold is achieved by relatively non-specific interactions of the type found in our fitness function. Our work also represents a nice extension of the results of Chan & Dill (1990, 1993) and Skolnick & Kolinski (1990), who use Monte Carlo grid-bound simulations in two and three dimensions and who also emphasize the importance of simple forces in achieving the proper fold. Our work is nonetheless free from fold biases due to particular grid topologies (Gregoret & Cohen, 1991). Hydrogen bonds were not sufficient to induce the proper fold in our genetic algorithm simulations. Since such bonds are also formed by an extended or unfolded structure in solution, the ineffectiveness of this parameter may be explained; the significance of such bonds, however, remains controversial (Creighton & Kim, 1991).

Though missing secondary structure elements can only sometimes be found by our genetic algorithm approach as in crambin, it is nonetheless able to bridge gaps and allow for proper extension of secondary structures. If the size of the predicted element is not too large relative to the observed substructure, the algorithm can often still point to the proper topology since the chances for outgrowth of this incorrect site are low due to the counter selection of the other parameters. The algorithm and fitness criteria have not yet been developed to correct robustly for wrongly conformed nucleation sites.

We have translated in this work the complex physicochemical forces between side-chains, secondary structural propensity and overall co-operativity of protein folding into abstract and simple rules that rely on knowledge of only the protein amino acid sequence. For the foreseeable future until the exact physicochemical forces are known and can be modeled in sufficient detail, our approach provides a way to bridge the gap between secondary structure predictions and tertiary fold.

References

Argos, P., Rossmann, M. G. and Johnson, J. E. (1977). A four-helical supersecondary structure. *Biochem. Biophys. Res. Commun.* **75**, 83–86.

Baldwin, R. L. & Roder, H. (1991). Characterizing protein folding intermediates. *Curr. Biol.* **1**, 218–220.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Chelvanayagam, G. & McKeaig, L. (1991). Stereo viewing on the PC/AT with EGA graphics. *J. Mol. Graph.* **9**, 111–114.

Chan, H. S. & Dill, K. A. (1990). Origins of structure in globular proteins. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 6388–6392.

Chan, H. S. & Dill, K. A. (1993). The protein folding problem. *Phys. Today*, **46**, 24–32.

Creighton, T. E. & Kim, P. S. (1991). Folding and binding. *Curr. Opin. Struct. Biol.* **1**, 3–4.

Dandekar, T. & Argos, P. (1992). Potential of genetic algorithms in protein folding and protein engineering simulations. *Protein Eng.* **5**, 637–645.

Eisenmenger, F., Argos, P. & Abagyan, R. (1993). A method to configure protein side-chains from the main-chain trace in homology modeling. *J. Mol. Biol.* **231**, 849–860.

Finzel, B. C., Weber, P. C., Hardman, K. D. & Salemme, F. R. (1985). Structure of ferricytochrome c′ from *Rhodospirillum molischianum* at 1·67 Å resolution. *J. Mol. Biol.* **186**, 627–643.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley Publ., Reading, MA.

Gregoret, L. M. & Cohen, F. E. (1991). Protein folding. Effect of packaging density on chain conformation. *J. Mol. Biol.* **219**, 109–122.

Hendrickson, W. A. & Teeter, M. M. (1981). Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature (London)*, **290**, 107–113.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor.

Kuwajima, K. (1989). The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins: Struct. Funct. Genet.* **6**, 87–103.

Lederer, F., Glatigny, A., Bethge, P. H., Bellamy, H. D. & Mathews, F. S. (1981). Improvement of the 2·5 Å resolution model of cytochrome b_{562} by redetermining the primary structure and using molecular graphics. *J. Mol. Biol.* **148**, 427–448.

Manavalan, P. & Ponnuswamy, P. K. (1978). Hydrophobic character of amino acid residues in globular proteins. *Nature (London)*, **275**, 673–674.

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79.

Ptitsyn, O. B. & Finkelstein, A. V. (1983). Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, **22**, 15–25.

Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1991). Prediction of protein backbone conformation based on seven structural assignments. *J. Mol. Biol.* **221**, 961–979.

Schiffer, M. & Edmundson, A. B. (1967). Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.* **7**, 121–135.

Schulz, G. E. & Schirmer, H. R. (1979). *Principles of Protein Structure*. Springer Verlag, New York.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science*, **250**, 1121–1125.

Stenkamp, R. E., Sieker, L. C. & Jensen, L. H. (1982). Restrained least-squares refinement of *Themiste dyscritum* methydroxohemerythrin at 2·0 Å resolution. *Acta Crystallogr. sect. B*, **38**, 784–792.

Sternberg, M. J. E. & Thornton J. M. (1978). Prediction of protein structure from amino acid sequence. *Nature (London)*, **271**, 15–20.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

Unger, R. & Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.* **231**, 75–81.

Edited by A. R. Fersht

(Received 31 August 1993; accepted 22 November 1993)