J. Mol. Biol. (1997) **269**, 240–259

JMB

Protein Folding Simulations with Genetic Algorithms and a Detailed Molecular Description

Jan T. Pedersen and John Moult*

Center For Advanced Research

We have explored the application of genetic algorithms (GA) to the

models (Kolinski & Skolnick, 1994), extending to one lattice point per two residues (Park & Levitt, 1996). It is clear that there must be a price to pay for these simplifications, but so far there have been few studies (Mounge *et al.*, 1995; DeBolt & Skolnick, 1996; Huang *et al.*, 1996) to establish to what degree the different simplifications affect the ability to identify the functional conformation reliably. We have taken a relatively conservative approach. An implicit solvent description is used, [redacted]

may not have converged to the lowest free energy conformation possible. The fact that the experimental structures consistently had slightly lower values supported the hypothesis that this last factor played a significant role. We therefore sought a more effective way of searching in torsion space in the compact states of a protein molecule.

Genetic Algorithms (GA) offer one way of searching more effectively in crowded spaces. They were first introduced by Holland (1975) to simulate [redacted]

1.7 Å within the final population in the GA simulation. These results were impressive, although the proteins simulated were present in the dictionary which was used to derive the GA. The search method is said to be approximately 100 to 200 times more efficient than MC simulated annealing protocols.

A more detailed review of the use of GAs for protein structure prediction may be found in Pedersen & Moult (1996).

The search method, presented here, is tested on a set of 28 fragments of protein structures, up to 14 residues long. The fragments were selected on the basis of experimental data and energetic criteria indicating a preference to adopt a native-like structure independent of the presence of the rest of the protein. They were identified from a survey of the current literature on the folding of proteins and protein fragments. These fragments have the advantage of being relatively small, and therefore present a more tractable search problem than complete proteins. A second advantage is that generation of a native-like conformation in an objective search provides added information about the con-

intramolecular electrostatic interactions; and the solvation free energy:

$$\Delta G = \Delta G_{\text{local}} + \Delta G_{ES} + \Delta G_{\text{solv}} \quad (1)$$

$$\Delta G = \sum_k S_{R(k)} E_k^L + \sum_k \sum_l K_{i(k)j(l)} E_{kl}^C + \sum_k \sum_l \sigma_{t(l)} \Delta A_l^n \quad (2)$$

Local backbone electrostatic energy

The first term:

$$\Delta G_{\text{local}} = \sum_k S_{R(k)} E_k^L \quad (3)$$

is a sum over the relative main-chain conformational free energy for all residues in the structure. E_k^L is the Coulomb energy of residue k, arising

Solvation free energy

The third term in equation (2):

$$\Delta G_{\text{solv}} = \sum_k \sum_l \sigma_{t(l)} \Delta A_l^n \qquad (5)$$

provides an area based solvation free energy. The sum Σ_k is over all polar, charged and non-polar groups in the structure. Σ_l is a sum over the differ-

the de-solvation free energy of these groups is effectively represented by the K_{ij} scaling.

A strong main-chain hydrogen bond has an energy of about −2.0 kcal/mol relative to the unfolded state, a little larger than has been suggested from the analysis of mutagenesis experiments (Shirley *et al.*, 1992). The local electrostatic interaction screening parameters result in a residue secondary structure preference close to that de-

structed. When a clash was encountered, a new set of angles for the current residue were drawn from the library and applied to the chain. If the construction of a residue failed ten times, the previous residue was reconstructed, before trying again. This procedure is in the worst case exponential with the length of the sequence to be built, but in practice has been found to scale approximately linearly with the number of residues in the peptide.

Generation of an initial population for a GA

[redacted]

the temperature in the previous interval, and $k = 0.9$.

Genetic Algorithm

In a GA, a set of genes represent different possible states of the system, and diversity is obtained through mutations within a gene and crossovers between genes. In the GAs used in this work, a gene consists of the information describing a conformation of a peptide, encoded as the set of ϕ, ψ

[redacted]

formation from all the 50 joint trials is evaluated for acceptance using a Metropolis test. If the free energy of the new conformation is lower than either of the parents the conformation is accepted, else the conformation is accepted with the probability:

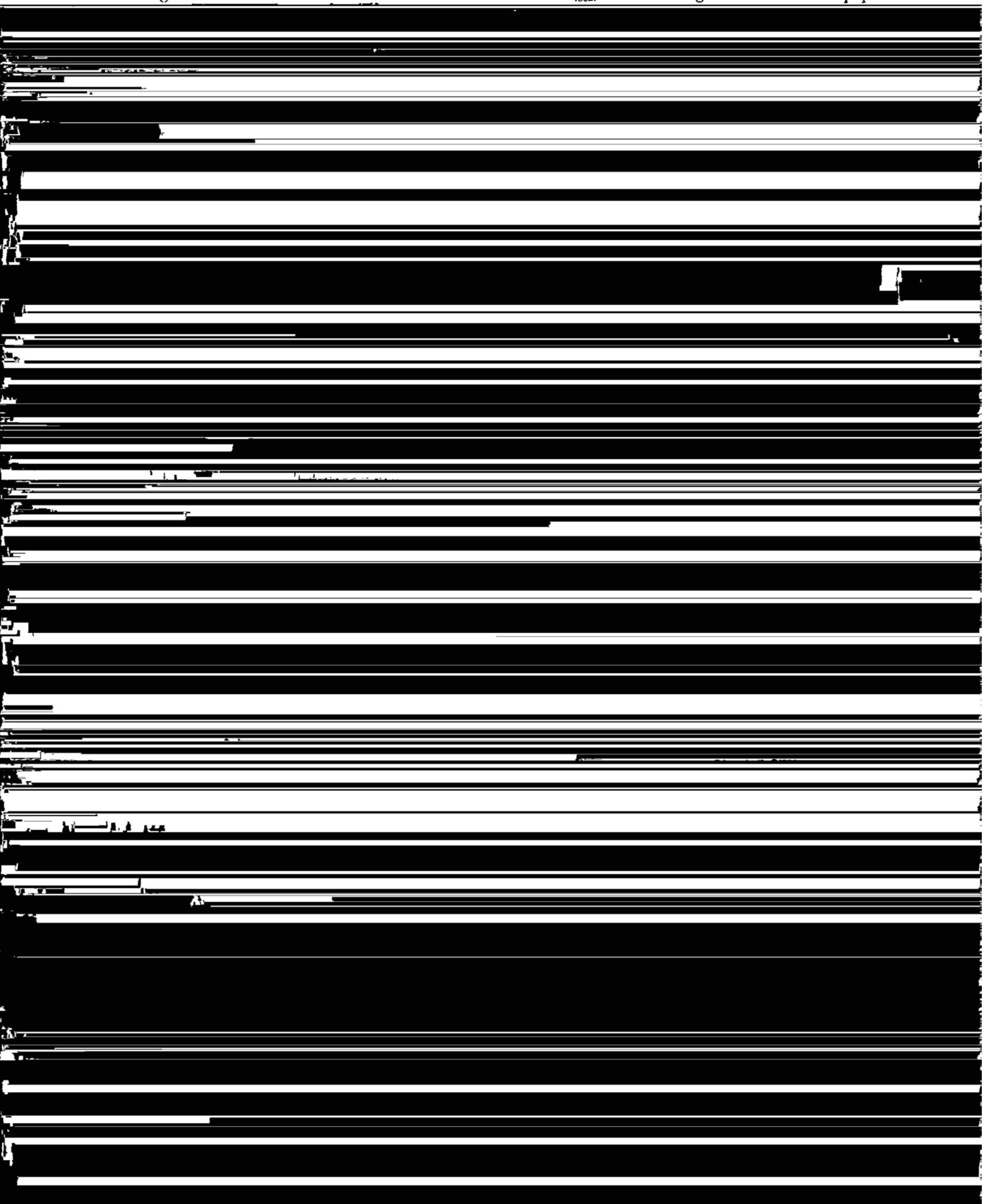
$$P_{accept} = e^{-\left(\frac{\Delta\Delta G_1 + \Delta\Delta G_2}{RT}\right)}$$

Selection of protein fragments to test the method

The selection of independent folding units (IFUs) for testing the GA is based on three criteria. (1) Experimental evidence for a context independent conformation in the part of the sequence which is to be simulated. (2) Large local burial of non-polar

Figure 1a to f show the evolution of the free

the ΔG_{local} distribution gives a measure of popu-



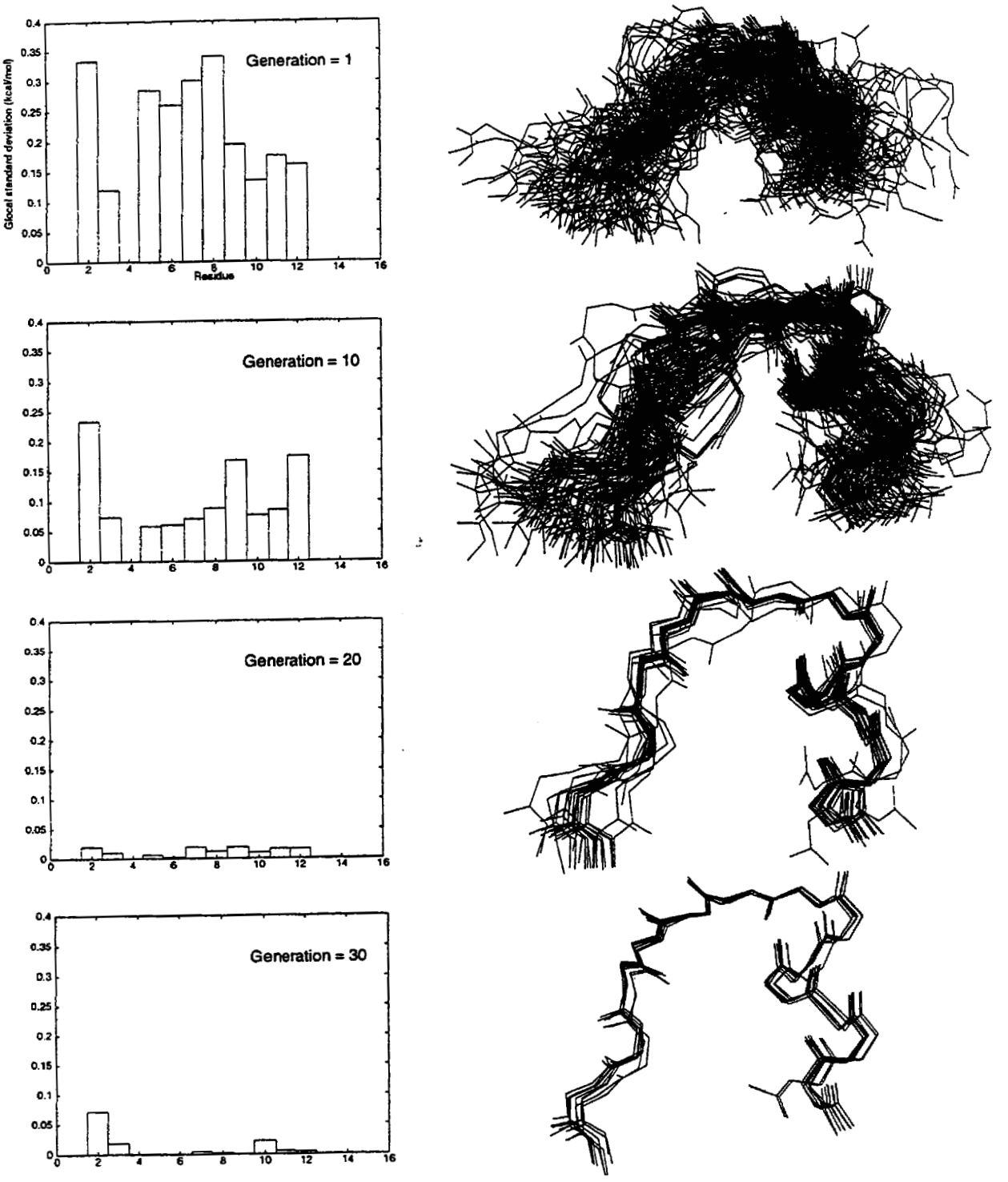


Figure 2. Snapshots from a GA simulation (residues 10 to 22 of Barnase (Baudet & Janin, 1991)). For each snapshot, the ΔG_{local} standard deviation is shown for each residue along the sequence together with the family of structures. The rapid reduction of the standard deviations shows how the population of structures converge in the GA. A start population of 100 conformations is used, and a carry forward fraction of 10%. The C^{α} RMS deviation between the final structure and the experimental structure is 2.4 Å.

(IBM590) processors. Simulations are also performed on the NIST IBM SP-2 parallel computer, using a ten-node processor network. A typical simulation of a 14 residue peptide takes four to six hours on ten processors.

The genetic algorithms are implemented in the molecular mechanics program J, which is written in C. The program provides a flexible platform for the implementation and testing of potential functions and search algorithms.

Table 1. Results for GA simulations on proposed independent folding units up to 14 residues long

No	Protein	Range	Len	Energy	RMS (Cα)	Structure (DSSP)	Sequence
1	1ALC	21-32	12	-24.6	2.07	---HHHHHHHT-	IALPELICTMFH
				-20.1	0.80	--HHHHHHHHH-	
2	1BGS	10-22	13	-29.2	2.41	-HHHHHS-SSS--	VADYLQTYHKLPD
				-25.9	0.50	-HHHHHHHSS---	
3	1BGS	88-98	11	-26.0	1.40	-EE-TT--EE-	ILYSSDWLIYK
				-23.9	0.10	-EE-TT--EE-	
4	1FKF	27-38	12	-27.1	2.44	--BTTTTB-TT-	TGMLEDGKKFDS
				-22.2	0.34	-EE-TTS-EEE-	
5	1FKF	46-59	14	-34.6	4.66*	---HHHHHHHHHH-	FKFMLGKQEVIRGW
				-18.4	2.59	----TTTT-S-TT-	
6	1HGF	100-113	14	-31.3	1.97	---HHHHHHHHH-	YDVPDYASLRSLVA
				-27.5	1.34	---TTHHHHHHHH-	
7	1HRC	7-18	12	-26.0	3.03*	-----HHHHHH-	KKIFVQKCAQCH
				-21.3	1.54	-HHHHHHTTTT-	
8	1HRC	92-103	12	-30.8	0.42	-HHHHHHHHHH-	REDLIAYLKKAT
				-26.7	1.04	-HHHHHHHHHH-	
9	1ILB	99-110	12	-29.0	6.04*	--HHHHHHHH-	FNKIEINNKLEF
				-19.2	2.32	----EETTEE--	
10	1LMB	15-26	12	-30.2	0.42	-HHHHHHHHHH-	ARRLKAIYEKKK
				-24.4	0.48	-HHHHHHHHHH-	
11	1MBC	6-17	12	-31.4	0.35	-HHHHHHHHHH-	EWQLVLHVWAKV
				-24.1	0.69	-HHHHHHHHHH-	
12	1MBC	29-40	12	-25.0	1.14	-HHHHHH-GGG-	LIRLFKSHPETL
				-22.7	1.06	-HHHHHH-GGG-	
13	1MBC	99-111	13	-29.8	5.87*	---EETTEE-TT-	IPIKYLEFISEAI
				-28.2	1.97	--HHHHHHHHHH-	
14	1MBC	131-142	12	-33.0	0.21	-HHHHHHHHHH-	MNKALELFRKDI
				-25.4	1.56	-HHHHHHHHHH-	
15	1PGA	43-54	12	-26.3	2.04	-EE-SS-EE---	WTYDDATKTFTV
				-19.8	0.80	-EEETTTTEEE-	
16	1UBQ	3-15	13	-26.6	6.55*	--THHHHHHS--	IFVKTLTGKTITL
				-26.9	0.62	-EEE-TTS-EEE-	
17	2I1B	69-82	14	-23.9	2.68	-----BTTB-----	LSCVLKDDKPTLQL
				-16.5	0.58	--EEE-SSSEEE--	
18	2I1B	103-112	10	-21.9	4.69*	--HHHHHH-	KIEINNKLEF
				-17.0	0.69	----SS----	
19	2MHR	51-62	12	-23.0	2.72	--SS-HHHHH-	TTNHFTHEEAMM
				-21.4	2.22	--TTTHHHHH-	
20	2MHR	67-78	12	-21.5	2.48	-----HHHHH-	YSEVVPHKKMHK
				-18.8	0.45	-TTHHHHHHH-	
21	2MHR	102-113	12	-30.0	1.52	-HHHHHHHHH-	WLVNHIKGTDFK
				-22.8	2.21	-HHHHHHHTS--	
22	2PCY	18-29	12	-26.3	2.99	-----HHHHH-	EPSISPGEKIVF
				-20.0	1.38	-----TT-----	
23	3LZM	24-35	12	-23.6	4.66*	--SS--GGGGG-	YYTIGIGHLLTK
				-16.8	1.81	--EETTEE---	
24	3LZM	99-111	13	-27.1	3.72*	---SSHHHHHH-	LINMVFQMGETGV
				-23.9	2.45	-HHHHHHHHHH-	
25	3SNS	16-29	14	-30.3	4.62*	-HHHHHHS-S---	TVKLMYKGQPMTFR
				-31.4	0.78	----EETTEE----	
26	4PTI	22-33	12	-28.3	5.61*	-THHHT--S--	FYNAKAGLCQTF
				-19.0	0.71	-EETTTTEE---	
27	5CYT	88-101	14	-35.1	0.20	-HHHHHHHHHHH-	KGERQDLVAYLKSA
				-30.7	1.22	-HHHHHHHHHHH-	
28	7RSA	2-13	12	-30.2	1.85	-HHHHHHHHH-	ETAAAKFERQHM
				-21.9	2.90	--TTTTTTTT--	

The Protein column gives the PDB (Bernstein *et al.*, 1977) code for the structure from which the experimental conformation is taken; Range, the set of residues (PDB numbering), and Len the number of residues. For each fragment, the first line gives the free energy of the final conformation, the RMSD to the experimental structure, and the secondary structure assignment for each residue. The second line gives the free energy and RMS deviation of the minimized experimental structure, and the experimental secondary structure. The RMS deviations are calculated on all C^{α} atoms of the fragment, excluding the blocking *N*-acetyl and C-aminomethyl blocking groups. *, Indicates structures that converged to and RMS deviation larger than 3.0 Å. Secondary structure assignments are those of DSSP (Kabsch & Sander, 1983); H, α-helix; B, β-bridge; E, extended strand; G, 3-helix; I, 5-helix; T, turn; S, bend. The sequence column provides the sequence of each fragment.

Simulation of independent folding units

Simulations were performed on the full set of 28 IFUs up to 14 residues long that were found to meet the criteria outlined above. Further details of these fragments and references to experimental data are available over the internet (Braxenthaler *et al.*, 1996). Longer fragments were also simulated but those simulations were less successful.

All simulations were performed using the same protocol as described in the previous sections.

The results of the simulations are summarized in Table 1 and Figure 3 shows a superimposition of the final structures from the simulations on the corresponding experimental ones. Final structures are the lowest free energy conformations in the last generation of the GA in each case.

Figure 4 shows the free energy of the minimized experimental structures compared with the free energy of the corresponding simulated structures. In 26 out of 28 cases, the GA finds conformations which are of lower free energy than that of the minimized crystal structures, in single GA simulations.

Agreement with experimental structures

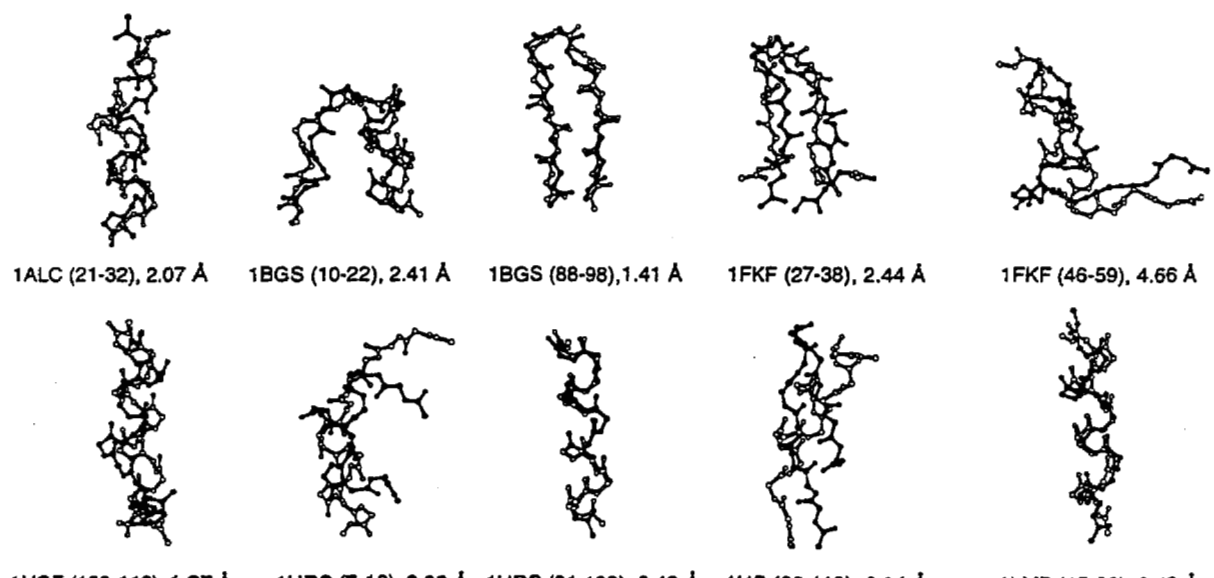
In 18 out of the 28 cases, the lowest free energy structure generated has an RMS deviation of less than 3 Å to the corresponding experimental structure. We consider 3 Å a reasonable criterion for a successful simulation for two reasons: First, inspection of the structures in Figure 3 shows a qualitative agreement for all these cases. Second, the analysis of RMS deviation matrices (Maiorov & Crippen, 1995) suggests this as a threshold for agreement.

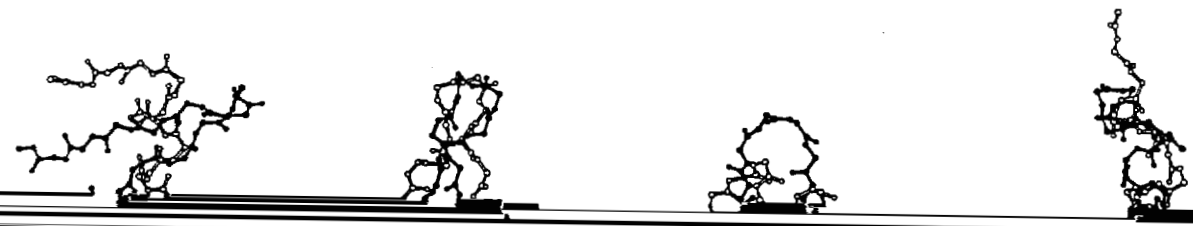
Four factors suspected of contributing to the cases where there is high RMS deviation to experimental structures were investigated; convergence, the role of the hydrophobic effect, sensitivity to main-chain covalent geometry, and the effect of generous allowance of van der Waals overlaps. An analysis of each of these factors is presented in the four next sections.

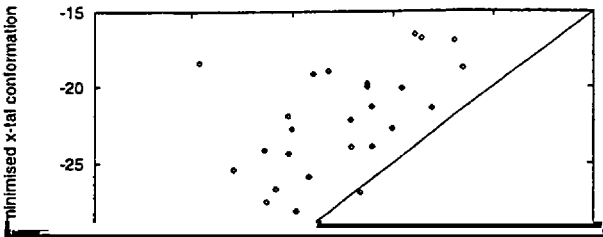
Convergence

Within a single GA run, the populations typically converge to an RMS deviation of less than 1 Å.

For ten of the 28 IFUs, the lowest free energy structures have RMS deviations larger than 3.0 Å to the corresponding experimental structures (marked with an asterisk in Table 1). For each of these fragments an additional five simulations were performed. GA simulations were started with five different MC simulations run with different random initial structures and different random number seeds. The lowest RMS deviation encountered and the corresponding free energy for each of these additional runs are shown in Table 3. In four of the ten cases, lower RMS deviation structures were found, but only one of these was a low-







simulations, with GROMOS (Åqvist *et al.*, 1985), Discover (Dauber-Osguthorpe *et al.*, 1988) and crystal structure geometries were performed on the bovine pancreatic trypsin inhibitor (BPTI) hairpin. GA simulations were performed with the same (random) starting sets of conformations and identical sets of random number seeds for each of the three geometries. First, 20 × 20 000 step MC simu-

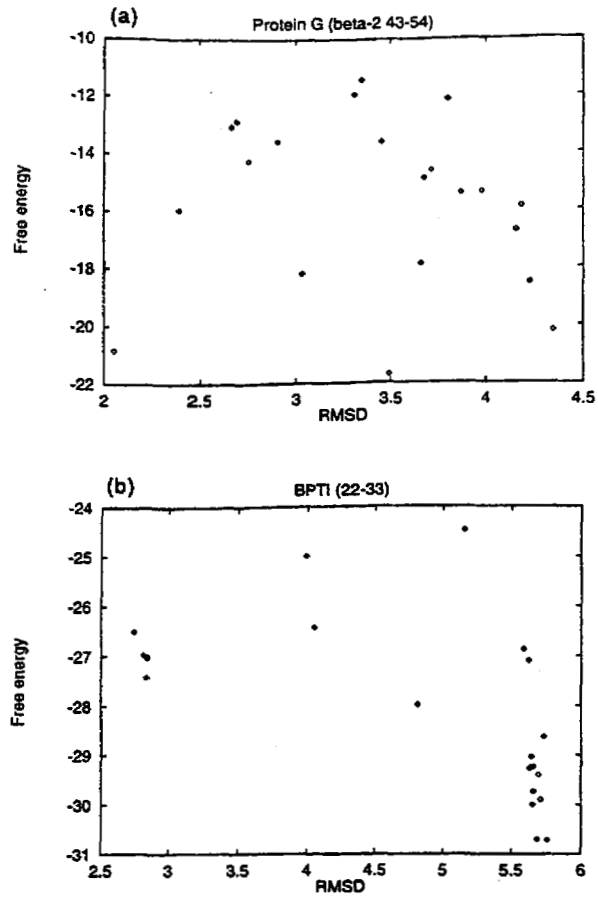
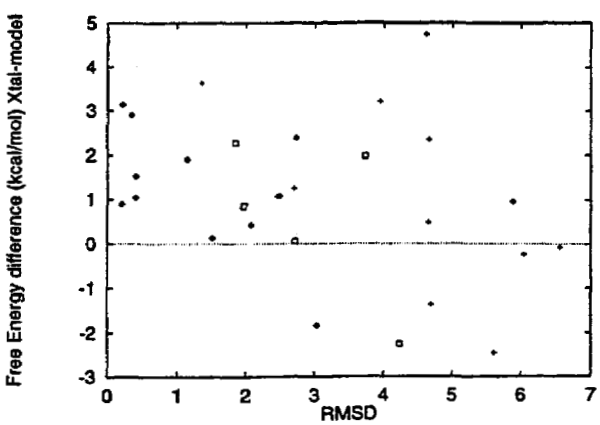


Figure 5. Convergence of 20 independent GA simulations for two different peptides. a, Protein-G hairpin (43-54); b, BPTI hairpin (22–33). In these cases the GA does not always converge to the same structure.



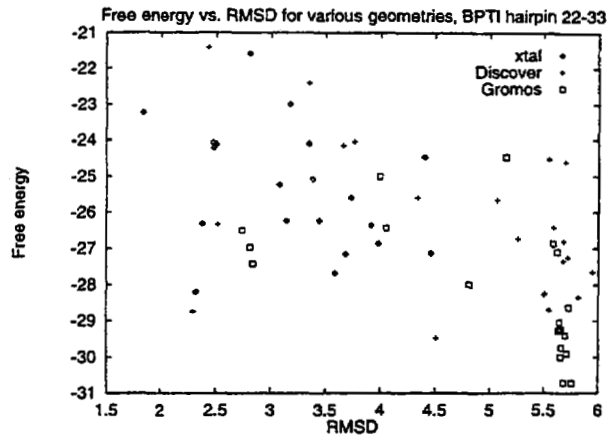


Figure 7. The result of three sets of 20 independent simulations of the BPTI (22–33) hairpin. In addition to the crystal structure (4PTI) geometry two different types of standard main-chain geometry were used, taken from Discover and GROMOS libraries. The simulations with the three geometries result in different distributions of RMS deviations from the experimental structure. Many high RMS deviation structures with low energy are obtained using the GROMOS geometry.

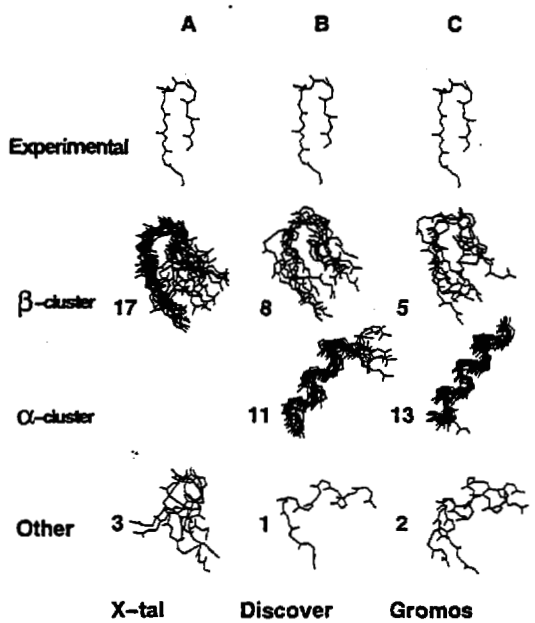


Figure 8. Families of conformations for residues 22 to 33 of BPTI, obtained from simulations with different types of covalent geometry. A, Crystal structure geometry; B, standard (Discover) geometry; C, standard (GROMOS) geometry. For each geometry the generated confor-

Table 3. Lowest RMS deviation and lowest free energy obtained in an additional five runs for each of the ten IFUs, which in the main simulation converged to an RMS deviation larger than 3.0 Å

Protein	Range	Lowest RMS deviation		Lowest free energy	
		Free energy	RMS deviation (C^α)	Free energy	RMS deviation (C^α)
1FKF	49–59	−33.14	3.16	−33.93	4.51
1HRC	7–18	−26.93	2.87	−28.03	3.21
1I1B	99–110	−25.88	3.66	−30.28	6.03
1MBC	99–111	−28.20	4.02	−29.12	6.12
1UBQ	3–15	−24.90	3.62	−31.05	6.91
2I1B	103–112	−24.17	1.53	−27.15	5.53
3LZM	24–35	−25.54	4.43	−27.61	6.12
3LZM	99–111	−24.13	3.29	−31.01	3.78
3SNS	16–29	−24.80	2.74	−30.65	2.93
4PTI	22–33	−26.40	2.71	−30.90	5.72

Only for four of these is a sub-3.0 Å structure obtained in the additional simulations, and only one of these four is a lowest free energy structure. The remaining nine may not be independent folding units, or the generated structures may represent false minima in the free energy surface.

Quality of side-chain conformations

There are clear side-chain conformational preferences as a function of backbone conformation in proteins (Dunbrack & Karplus, 1994). We have

Applying this potential to the 28 simulated independent folding units given in Table 1 and calculating the $\Delta G_{\chi 1}$ for the simulated and experimental conformations shows (Figure 9) that unfavorable side-chain configurations are seen in most of the

than a previously explored MC method (Avbelj & Moult, 1995a), by a factor of about 10^2. The major reason for this is the difficulty of finding acceptable crossover conformations. We have not compared its performance to that of more sophisticated MC move schemes such as that of Eloffson *et al.* (1995), where more extensive local move searches are used.

We have established GA simulation parameters suitable for small peptide fragments (up to 14 residues). It is clear that for longer fragments a larger conformational ensemble is required to provide adequate conformational diversity. In simulations of a 22 residue sequence, e.g. the membrane bind-

viously been used to assess the quality of potentials. This test is probably the most demanding. Regions of serious error in the potential surface may be quite limited so that decoys sparsely sampling the conformational space are unlikely to encounter them. In contrast, a thorough search is likely to follow such wormholes to false minima.

Performance of the potential

As discussed above, the GA is generally effective at finding conformations with free energies at least as low as the corresponding minimized experimental structures. Thus, it is possible to use the RMS

tween such possibilities. Recently, one clear case of distortion of a helical fragment relative to the full context structure, optimizing hydrophobic interactions has been identified (F. Poulsen, personal communication). Interplay with other factors in the force field may also play a role, as discussed in the next section.

The importance of side-chain packing

Figure 9 shows that the majority of the generated structures have average χ_1 angles significantly distorted from the values found in experimental structures. Angle choices in the simulation are primarily based on the distribution of values found for particular residue types in experimental protein structures, without regard to local backbone conformation. Side-chain conformational preferences do vary as a function of residue conformation (Dunbrack & Karplus, 1994). Further, it has been proposed that the steric restraints of side-chain packing are key to the selection of a secondary structure (Creamer & Rose, 1992; Srinivasan & Rose, 1995). In the present simulations, a generous amount of hard sphere overlap is allowed when [redacted]

hydrogen-bond strength. Tests using an ideal helix ($\phi/\psi = -65, -40$) show that a difference in the N–C^{α}–C bond angle of 5° results in a 0.4 kcal/mol per residue difference in hydrogen-bond free energy. This is a significant difference.

Implications for protein folding

For 18 out of the 28 cases examined, the lowest free energy structure resulting from the GA simulation is less than 3 Å RMS deviation on C^{α} atoms from the experimental structure of the fragment in the full protein environment. That is, for these cases, the preferred structure of the fragment in isolation is found to be similar to that it adopts in the complete structure. Fragments were chosen on the basis of experimental evidence supporting a significant native like population, and the large amount of non-polar burial that occurs on folding. Thus, both the simulation and the experimental data support the idea that the conformation of these fragments is largely context independent: the preferred conformation is not significantly altered by the larger environment of the protein. Note that this does not mean that the fragments will have a [redacted]

stituting a complete protein would exhibit context independence. For the cases where a large set of fragments from a single protein have been examined experimentally (Dyson *et al.*, 1992a,b), it appears to be small, implying a limited number of starting points for the folding process.

Acknowledgements

We thank NIST for use of the IBM SP2 parallel computer, and NIST computing staff for supporting us in its use. This work was partly funded by grants DOC/NIST 60NANBD1594 and NIH GM40134.

References

Abagyan, R. & Totrov, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002.

Allen, M. P. & Tildesley, D. J. (1989). *Computer Simulation of Liquids*, Oxford Science Publications, Clarendon Press, Oxford.

Alonso, D. O. V. & Daggett, V. (1995). Molecular dynamics simulations of protein unfolding and limited refolding: characterization of partially unfolded states of ubiquitin in 60methanol and in water. *J. Mol. Biol.* **247**, 501–520.

Anderson, D. E., Becktel, W. J. & Dahlquist, F. W. (1990). pH-induced denaturation of proteins: a salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, **29**, 2403–2408.

Åqvist, J., Gunsteren, V. W. F., Leijonmarck, M. & Tapia, O. (1985). A molecular dynamics study of the C-terminal fragment of the l7/l12 ribosomal protein: secondary structure motion in a 150 picosecond trajectory. *J. Mol. Biol.* **183**, 461–477.

Avbelj, F. (1992). Use of a potential of mean force to analyse free energy contributions in protein folding. *Biochemistry*, **31**, 6290–6297.

Avbelj, F. & Moult, J. (1995a). Determination of the conformations of folding initiation sites in proteins by computer simulation. *Proteins: Struct. Funct. Genet.* **23**, 129–141.

Avbelj, F. & Moult, J. (1995b). The role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, **34**, 755–764.

Baudet, S. & Janin, J. (1991). Crystal structure of a barnase complex at 1.9 Å resolution. *J. Mol. Biol.* **219**, 123–132.

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

Blanco, F. J. & Serrano, L. (1995). Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *Eur. J. Biochem.* **230**, 634–649.

Brandt, D. T. & Flory, P. J. (1965). The role of dipole interactions in determining polypeptide conformation. *J. Am. Chem. Soc.* **87**, 663–664.

Braxenthaler, M., Pedersen, J. T., Samudrala, R., Lou, R. & Moult, J. (1996). Carb biocomputing web pages, force field and parameters: http:/ /iris4.carb.nist.gov/WWW/moultgroups/potentials.html; pre-processed library of folds: http: / /prostar.carb.nist.gov:8000/PDec/PDecRetr_strucib.html; independent folding units: http: / /iris4.carb.nist.gov/WWW/moultgroup/ifu.html.

Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.

Brooks, C. L. I. & Head-Gordon, T. (1991). Virtual rigid body dynamics. *Biopolymers*, **31**, 77–100.

Brooks, C. L. I., Karplus, M. & Pettitt, M. (1991). G. Nemethy on proteins: a theoretical perspective of dynamics structure and thermodynamics. *Advan. Chem. Phys. Bull. Mathemat. Biol.* **53**, 313.

Brunne, R. M., Berndt, K. D., Guntert, P., Wuthrich, K. & van Gunsteren, W. F. (1995). Structure and internal dynamics of the bovine pancreatic trypsin inhibitor in aqueous solution from long-time molecular dynamics simulations. *Proteins: Struct. Funct. Genet.* **23**, 49–62.

Chothia, C. (1984). The principles that determine the structure of proteins. *Annu. Rev. Biochem.* **55**, 537–572.

Creamer, T. P. & Rose, G. D. (1992). Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc. Natl Acad. Sci. USA*, **89**, 5937–5941.

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

Dandekar, T. & Argos, P. (1994). Folding the main-chain of small proteins with the genetic algorithm. *J. Mol. Biol.* **236**, 844–861.

Dauber-Osguthorpe, P., Roberts, V. A., Osguthorpe, D. J., Wolff, J., Genest, M. & Hagler, A. T. (1988). Structure and energetics of ligand binding to proteins: *E. coli* dihydrofolate reductase-trimethoprim, a drug receptor system. *Proteins: Struct. Funct. Genet.* **4**, 31–47.

DeBolt, S. E. & Skolnick, J. (1996). Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. *Protein Eng.* **9**, 637–655.

Dunbrack, R. & Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nature Struct. Biol.* **1**, 334–340.

Dyson, H. J., Merutka, G., Waltho, J. P., Lerner, R. A. & Wright, P. E. (1992a). Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding. I. Myohermerythrin. *J. Mol. Biol.* **226**, 795–817.

Dyson, H. J., Sayre, J. R., Merutka, G., Shin, H. C., Lerner, R. A. & Wright, P. E. (1992b). Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding. I. Plastocyanin. *J. Mol. Biol.* **226**, 819–835.

Elofsson, A., LeGrand, S. & Eisenberg, D. (1995). Local moves, an efficient method for protein folding simulations. *Proteins: Struct. Funct. Genet.* **23**, 73–82.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple models for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **78**, 97–120.

Geist, A., Beguelin, A., Dongarra, J., Jiang, W., Manchek, R. & Sinderam, V. (1993). *PVM 3 User's Guide and Reference Manual* (ORNL/TM-12187) .

Gilbert, G. & Baleja, J. (1995). Membrane-binding peptide from the C2 domain of factor VIII forms an amphiphatic structure as determined by NMR spectroscopy. *Biochemistry*, **34**, 3022–3031.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, San Mateo, CA.

Hagler, A. T. (1985). The Peptides: Analysis, Synthesis, Biology. (Udenfriend, S. & Meienhofer, J., eds), vol. 7, pp. 213–299, Academic Press, Orlando, FL.

Hao, M.-H. & Sheraga, H. (1994). Monte Carlo simulation of a first-order transition for protein folding. *J. Phys. Chem.* **98**, 4940–4948.

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

Holm, L. & Sander, C. (1994). Protein-structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123–138.

Matouschek, A., Serrano, L., Meiering, E. M., Bycroft, M. & Fersht, A. R. (1992). The folding of an enzyme: V. H/^{2}H exchange-nuclear magnetic resonance studies on the folding pathway of barnase: complementarity to and agreement with protein engineering studies. *J. Mol. Biol.* **224**, 837–845.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1091.

Moult, J. & Unger, R. (1991). An analysis of protein folding pathways. *Biochemistry*, **30**, 3816–3824.

Mounge, A., Lathrop, E. J. P., Gunn, J. R., Shenkin, P. S. & Friesner, R. A. (1995). Computer modelling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995–1012.

Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367–392.

Pedersen, J. T. & Moult, J. (1995). *Ab initio* structure pre-

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

Weiner, S., Kollman, P., Case, D., Singh, U., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). A new force field for molecular mechanics simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–783.

Wetlaufer, D. (1973). Nucleation, rapid folding, and globular interchain regions in proteins. *Proc. Natl Acad. Sci. USA*, **70**, 697–701.

Williams, S., Causgrove, T. P., Gilmanshin, R., Fang, K. S., Callender, R. H., Woodruff, W. H. & Dyer, R. (1996). Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry*, **35**, 691–697.

Edited by F. E. Cohen

(Received 7 November 1996; received in revised form 21 February 1997; accepted 21 February 1997)