Protein Engineering vol.9 no.1 pp.5–14, 1996

Does a backwardly read protein sequence have a unique native state?

Krzysztof A.Olszewski[1,2], Andrzej Kolinski[1,2] and Jeffrey Skolnick[1,3]

[1]Department of Molecular Biology, Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037, USA and [2]Department of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland

[3]To whom correspondence should be addressed

Amino acid sequences of native proteins are generally not palindromic. Nevertheless, the protein molecule obtained as a result of reading the sequence backwards, i.e. a retro-protein, obviously has the same amino acid composition and the same hydrophobicity profile as the native sequence. The important questions which arise in the context of retro-proteins are: does a retro-protein fold to a well defined native-like structure as natural proteins do and, if the answer is positive, does a retro-protein fold to a structure similar to the native conformation of the original protein? In this work, the fold of retro-protein A, originated from the retro-sequence of the B domain of *Staphylococcal* protein A, was studied. As a result of lattice model simulations, it is conjectured that the retro-protein A also forms a three-helix bundle structure in solution. It is also predicted that the topology of the retro-protein A three-helix bundle is that of the native protein A, rather than that corresponding to the mirror image of native protein A. Secondary structure elements in the retro-protein do not exactly match their counterparts in the original protein structure; however, the amino acid side chain contact pattern of the hydrophobic core is partly conserved.

Keywords: lattice representation of proteins/Monte Carlo method/retro-protein A/*Staphylococcal* protein A/three-helix bundle

Introduction

The biological functions of proteins are based on their unique three-dimensional structure. Since the Anfinsen refolding experiments (Anfinsen, 1973), it is believed that the native structure of proteins is uniquely determined by their amino acid sequence. However, the problem of determining a protein's three-dimensional structure from the sequence itself remains unsolved, despite years of intensive research (see Vásquez *et al.*, 1994). Among various attempts to solve the protein folding problem, those which utilize the reduced representation of the protein molecule seem to emerge as methods that allow study of the stability and folding of small proteins at acceptable computational cost (Skolnick and Kolinski, 1989; Chan and Dill, 1993; Godzik *et al.*, 1993a; Liwo *et al.*, 1993; Shakhnovich and Gutin, 1993; Hao and Scheraga, 1994; Kolinski and Skolnick, 1994a,b; Šali *et al.*, 1994; Socci and Onuchic, 1994; Park and Levitt, 1995). In this work, using a high coordination lattice model of proteins, we studied the behavior of a novel retro-protein. We attempted to establish possible links that exist between the retro-protein structure and the structure of the original protein.

Naturally occurring proteins are built from L-amino acids, consecutively connected by amide bonds in order to produce a long backbone of amide bonds with amino acid side groups, which are attached to the chiral C^{α} carbon atoms. The amide (peptide) bond that connects consecutive C^{α} carbons is almost planar (Corey and Pauling, 1953), since the rotation around the bond between NH and CO groups is hindered. Therefore, the conformational variety of protein structures originates from changes in the relative orientation of the consecutive peptide bond plates customarily described by dihedral angles Φ and Ψ (Scheraga, 1968). Thus, the conformation of the protein backbone can be roughly described by specifying only the location of the C^{α} carbons (Oldfield and Hubbard, 1994). Let us define the retro-sequence as the backwardly read sequence of the original protein. One can attempt to rebuild the retro-sequence on the existing C^{α} backbone. This leads to a structure similar to the original protein but with completely translocated side-chains with respect to their original positions (Figure 1). The alternative approach is to rebuild the putative retro-protein structure using a retro-sequence but starting from the C-terminus of the original protein backbone instead of the N-terminus. The resulting structure is compatible with the original protein, but the direction of the protein backbone is opposite. Moreover, the C^{α}–C^{β} bonds now point in different directions, since each C^{α} carbon (except those of glycines) is chiral, e.g. the side chains in helices will lie in the opposite direction (Figure 2). This leads to a potentially different pattern of side chain contacts. If D-amino acids were used in the rebuilding procedure instead of L-amino acids, the overlap of side chains would be greater; however, the backbone direction still remains opposite the native protein. Note, therefore, that the homology between the native protein sequence and its retro-sequence is generally very low. The result of backward reading of the protein sequence (retro-transition) will be further referred to as a retro-protein. The result of changing the absolute chirality of amino acids (chiral transition) performed on a native protein containing L-amino acids (L-protein) will be referred to as a D-protein. The backward reading of the sequence does not change the chirality of amino acids constituting a protein; therefore, it cannot produce a protein composed of D-amino acids, i.e. the retro-transition and the chiral transition are independent of each other.

Since the chiral transition produces a perfect mirror image of the L-protein, it is safe to assume that the D-protein acquires the perfect mirror image structure of the L-protein upon folding. Indeed, D-HIV protease has been synthesized and shown to acquire a perfect mirror image fold of the naturally occurring HIV protease (Milton *et al.*, 1992). Recently, another example of a D-protein, the Leu5 variant of trypsin inhibitor, has also been shown by Nielsen *et al.* (1994) to acquire the mirror-image form. Also, various cyclic and linear oligopeptide hormones have been synthesized that are related to each other

© Oxford University Press

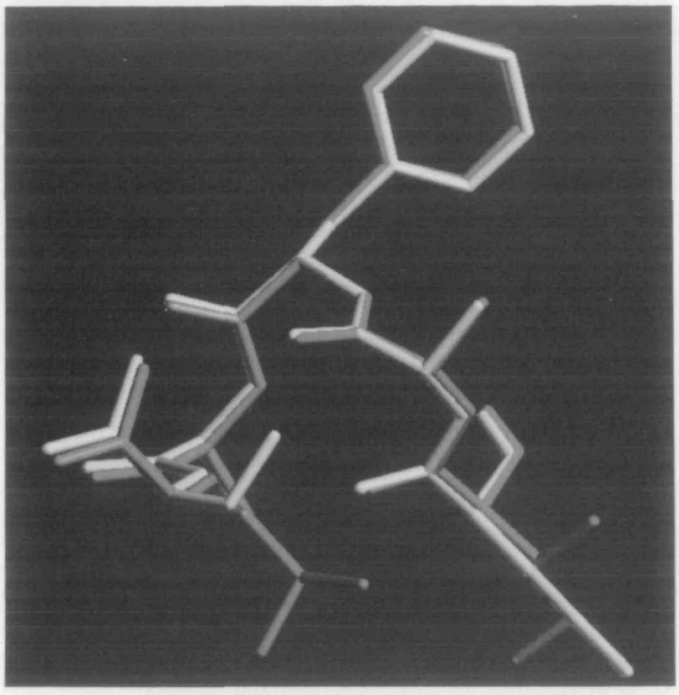

Fig. 1. Two helical fragments of the Ser–Ala–Phe–Ala–Ile peptide (white) and its retro-version (grey) arranged to maximize overlap between consecutive C^{α}s. Note that the backbone direction does not change and side chains point in the same direction. However, both terminal amino acids are exchanged (the middle sequence is palindromic, and therefore it is invariant with respect to the retro-transition).

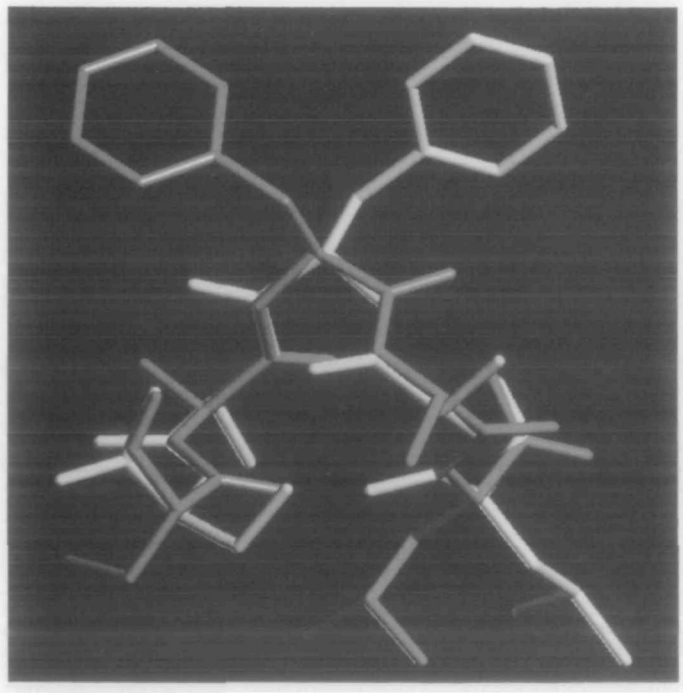

Fig. 2. Two helical fragments of the Ser–Ala–Phe–Ala–Ile peptide (white) and its retro-version (grey) arranged to maximize overlap between C^{α}s corresponding to identical amino acids. Note that the backbone direction is then reversed, and the side chains point in opposite directions.

by the chiral and/or the retro-transition (Goodman and Chorev, 1979). In the case of cyclic peptides, it has been demonstrated that the related retro-peptides maintain biological activity. In the case of linear peptides, however, the D-oligopeptides, retro-oligopeptides and retro-D-oligopeptides were devoid of any biological activity (Goodman and Chorev, 1979). It has been argued that retro-D-oligopeptide derivatives, with altered end groups, might be topologically equivalent to the native oligopeptides. This observation gave rise to speculation that the retro-protein, as a result of the folding process, might adopt the mirror image structure of the native protein (Guptasarma, 1992).

On closer inspection, the hypothesis that the retro-protein will adopt the mirror image structure of the original protein is very unlikely, mainly because right-handed helices would have to be replaced by left-handed helices. Although it is not entirely impossible [according to a Ramachandran map (Ramachandran *et al.*, 1963)] for a left-handed helix to exist, this replacement would require a larger stabilization from the packing interactions to overcome the entropic loss that arises when the molecule is shifted to the much narrower, left-handed helical potential energy well. The lengths of the NH and CO bonds and hydrogen and oxygen radii differ enough to effectively block the frequent occurrence of a left-handed α-helix, while the right-handed helix is commonly observed. Moreover, α-helices in proteins are often capped by residues that can form a hydrogen bond with the NH of the initial residues in the helix and with the C=O of the final residues of the helix (Presta and Rose, 1988; Richardson and Richardson, 1988). If the above hypothesis holds, since the retro-transition changes the direction of helical sequences (and its hydrogen bonds), then the resulting capping residues are not optimally distributed and, in principle, they may not stabilize newly formed helices. Also, turn region sequences, when read backwards, will rarely be in agreement with the turn tendencies observed for real proteins (Wilmot and Thornton, 1988).

However, retro-proteins constitute a very interesting case for the study of protein core packing. The amino acid composition of a retro-protein is the same as the original protein; therefore, all methods based on amino acid composition will predict the same structural class for both of them, cf. Chou (1995). Typically, in proteins, non-polar side chains are tightly packed into the interior to form a solvent-inaccessible hydrophobic core (Richards, 1977). The distribution of non-polar residues (hydrophobic profile) along the protein chain is one of the most conservative determinants of the native structure (Bowie *et al.*, 1990). Obviously, the hydrophobicity profile of the retro-protein remains intact, assuming lack of backbone directionality. On the other hand, among sequences with similar hydrophobic profiles, the possibility of folding is restricted to the subset of sequences for which core packing is sterically allowed (Rose and Wolfden, 1993). However, a retro-protein can approximately recover the packing of the original protein (with a structure adjusted in order to accommodate changes induced by different directions of side chains). The more important the packing interactions are, the greater the tendency of the retro-protein to acquire a fold similar to that of the native protein will be. In this paper, we consider the possibility of folding from the random state of a simple retro-protein using a lattice model of proteins. Also, we discuss the importance of the different contributions to the potential energy that may stabilize the folded structure. In particular, we examine the relative importance of the local secondary structure propagating terms versus hydrophobic core packing interactions in determining the unique topology of the protein.

Protein A constitutes a cell-wall component of *Staphylo-*

coccus aureus that binds to an Fc domain of immunoglobulins. Its extracellular part consists of five highly homologous domains designated E, D, A, B and C, respectively (Table I). The B domain of *Staphylococcal* protein A, complexed with the Fc portion of human polyclonal immunoglobulin G, has been crystallized and the structure of the complex has been solved. The B domain part of the complex consists of two helices, from Gln10 to Leu18 (helix I) and from Glu26 to Asp37 (helix II), which are packed together to form an antiparallel helical hairpin (Deisenhofer, 1981). The three-dimensional solution structure of the B domain has also been determined by NMR spectroscopy (Gouda *et al.*, 1992) in the absence of the complexing immunoglobulin. In water, it forms a stable three-helix bundle motif with helix I (Gln10 to His19) tilted with respect to the antiparallel hairpin formed by helices II (Glu26 to Asp37) and III (Ser42–Ala55). The N-terminal residues up to Glu9 and the C-terminal residues from Gln56 to the terminal lysine do not exhibit ordered structure. The absence of the third helix in the crystal structure of the B domain complex with immunoglobulin is probably induced by crystal contacts (Gouda *et al.*, 1992).

The distribution of secondary structural elements in the solution structure of the B domain of protein A agrees with the capping properties of helical termini (Presta and Rose, 1988; Richardson and Richardson, 1988). The first helix is fairly well capped at the N-terminus by Asn12, and the C-terminus by His19. The second helix N-cap Asn26–Glu25–Glu26 is perfect, while the C-cap is marked only by the Lys36 residue. In the third helix, Asn44, together with Ser42, agrees with the capping properties at the N-terminus, and the C-cap is formed by the Lys50, Lys51 and Asn53. Also, Pro21 and Asn22 that constitute the first turn are highly expected in their positions (Wilmot and Thornton, 1988). The second turn (between helices II and III) is even more exemplary, being built from Asp38–Pro39–Ser40–Gln41 (Wilmot and Thornton, 1988).

Structures of the other domains of protein A have not been reported previously. However, since all domains of protein A are at least 80% homologous to each other and also bind to immunoglobulin, we assume that the overall structure is conserved within the family of domains of protein A. Moreover, in order to remove an Asn–Gly pair from the native sequence of the B domain of protein A, the so-called protein Z, has been proposed, and subsequently expressed as a single point mutation of the B domain involving Gly30 and Ala30 (G30A) (Nilsson *et al.*, 1987). Its NMR structure (Lyons *et al.*, 1993) reveals a three-helix bundle topology for protein Z in solution.

Recently, a lattice model of proteins has been used to redesign the B domain of protein A so that its mutant preserves the three-helix bundle topology, but has a different overall chirality of the global fold (Olszewski *et al.*, 1995). We have shown that, although the native topology of the three-helix bundle is strongly conserved, it is possible to find, by an extensive search of possible mutations, a putative mutant that may exhibit the topological mirror image structure. Moreover, additional studies of protein A mutations have proved that the lattice model used here can differentiate between the two topological alternatives of the three-helix bundle, therefore encouraging its application to the study of the retro-protein folding simulations and packing interaction studies.

The outline of the remainder of this paper is as follows. In the Methods section, we briefly present the lattice model of proteins and the interaction scheme used. The model is essentially the same as that used in our previous studies (Kolinski and Skolnick, 1994a; Olszewski *et al.*, 1995); however, for the reader's convenience, we describe concisely the lattice representation of the protein chain and the various contributions to the force field. Also, the algorithm for all-atom model building is discussed. In the Results section, we discuss in detail the lattice simulations of retro-protein A; this is then followed by the all-atom model building. Additional analysis of secondary structure predictions corroborates our predicted structure of retro-protein A.

Methods

A 90-component, high-coordination lattice model used for the protein backbone representation (Kolinski and Skolnick, 1994a) was constructed by making all possible permutations of the components of the generic vectors (3,1,1), (3,1,0), (3,0,0), (2,2,1) and (2,2,0), with the lattice unit length equal to 1.22 Å. These vectors connect consecutive C^αs along the protein backbone, thus serving as virtual bonds. No backbone atoms other than the C^αs are explicitly used, and only consecutive pairs of vectors that form protein-like angles between virtual bonds (i.e. from 72.5 to 154°) are permitted. The lattice representations of a high-resolution library protein C^α carbon are within 0.7 Å r.m.s. (root mean square deviation of C^α carbons) of their continuous space representations (Godzik *et al.*, 1993b). A library of single ball rotamers is used to represent amino acid side chains. Rotamers are located at the side chain center of mass positions and depend on the local geometry of the protein backbone. The number of allowed rotamers is different for different amino acids and varies from one rotamer (e.g. for alanine) to a maximum of 58 rotamers for arginine, in certain backbone configurations. The local accuracy of this side chain center of mass representation is approximately 1 Å r.m.s. On combining the side chain representation into the model, the overall intrinsic geometrical accuracy of the model decreases to about 2 Å r.m.s. Monte Carlo simulations of protein folding on the above lattice are performed by accepting or rejecting small movements of the protein backbone on the basis of the asymmetric Metropolis criterion (Metropolis *et al.*, 1953). These movements include predefined local two- and three-virtual-bond moves and also long-distance moves designed to enhance the search of conformational space. The latter moves are generated by concerted sequences of overlapping three-bond motions. In addition, random changes of rotamer positions are allowed to facilitate the packing of side chains (Kolinski and Skolnick, 1994a).

The potential energy used consists of five terms, viz.

$$E = w_{\text{prop}}E_{\text{prop}} + w_{\text{hb}}E_{\text{hb}} + w_{\text{one}}E_{\text{one}} + w_{\text{pair}}E_{\text{pair}} + w_{NN}E_{NN} \quad (1)$$

generated by the analysis of the library of high-resolution PDB structures of globular proteins (Kolinski and Skolnick, 1994a) with the weighting factors w for the various energy contributions equal to $w_{\text{prop}} = 1$, $w_{\text{hb}} = 0.5$, $w_{\text{one}} = 1.75$, $w_{\text{pair}} = 2$ and $w_{NN} = 0.25$.

In particular, the local, sequence-dependent term

$$E_{\text{prop}} = \sum_{i=1}^{N-3} E^{\text{prop}}_{i,\, i+2,\, i+3}\,(\tilde{r}_{i,i+1,i+2}, \tilde{r}_{i+1,i+2,i+3}) \quad (2)$$

involves two consecutive C^α_i to C^α_{i+3} and C^α_{i+1} to C^α_{i+4} (the index i represents consecutive C^αs) chiral distances, i.e. $\tilde{r}_{i,\, i+1,i+2}$ and $\tilde{r}_{i+1,i+2,i+3}$, respectively. The chiral

distance is defined as follows: $\tilde{r}_{i,j,k}$ = sign $((\mathbf{b}_i \times \mathbf{b}_j)\cdot\mathbf{b}_k) \| \mathbf{b}_i + \mathbf{b}_j + \mathbf{b}_k \|$, where $\mathbf{b}_i$ is a virtual bond that connects C_i^α to C_{i+1}^α. The subscripts in E_{ijk}^{prop} (Equation 2) signify the amino acid sequence dependence of the term. Since two overlapping C_i^α–C_{i+3}^α chiral distances are involved, E^{prop} propagates protein-like elements of secondary structure along the protein backbone. In addition to this local term, an effective interaction between C_i^α and C_j^α that simulates the formation of a hydrogen bond between backbone atoms of the *i*th and *j*th residues was introduced:

$$E_{hb} = \sum_{i<j}^{N} (E^{H} + E^{HH}\delta_{i\pm1,j\pm1})\delta_{ij} \tag{3}$$

where $\delta_{ij} = 1$ when amino acids i and j form a hydrogen-bonded pair and otherwise $\delta_{ij} = 0$. E^{H} and E^{HH} are equal to -0.5 kT. Amino acids i and j form a hydrogen-bonded pair when, and only when, the following geometrical criteria are satisfied:

$$\begin{aligned} R_{min} \leq \| \mathbf{r}_{ij} \| \leq R_{max} \\ |(\mathbf{b}_{i-1} - \mathbf{b}_i)\cdot\mathbf{r}_{ij}| \leq a_{max} \\ |(\mathbf{b}_{j-1} - \mathbf{b}_j)\cdot\mathbf{r}_{ij}| \leq a_{max} \end{aligned} \tag{4}$$

where $\mathbf{r}_{ij}$ is a vector connecting C_i^α to C_j^α. R_{min} = 4.6 Å, R_{max} = 7.3 Å and a_{max} = 13.4 Å^2. This interaction component is sequence independent and also non-directional, since the models do not specify the location of backbone atoms other than C^α. Proton donors and acceptors are not differentiated. Every amino acid but proline can form up to two hydrogen bonds, whereas proline can participate in only one hydrogen bond. The hydrogen-bonding scheme accounts for the co-operativity of hydrogen bond formation by introducing an effective interaction between adjacent pairs of hydrogen bonds.

A one-body, centrosymmetric burial potential,

$$E_{one} = \sum_{i=1}^{N} E_i(r_i^0/s) \tag{5}$$

reflects the radial distribution of distances r_i^0 from the *i*th amino acid side chain to the center of mass of the protein [where s is the expected radius of gyration, calculated for a closely packed protein (Kolinski and Skolnick, 1994a)]. The purpose of the burial energy is to enforce the compaction of the protein; therefore, the E_{one} component is small for compact states and dominates denatured states. It serves as a driving force in the initial stages of the folding process (Kolinski and Skolnick, 1994a) by narrowing the conformational space search to compact or near-to-compact states.

The final packing of the protein core depends on more detailed packing interactions. They are modeled by the combination of a pairwise, soft-core repulsion augmented by a square-well potential, derived as a potential of mean force from the frequency of close contact occurrences between amino acids:

$$E_{pair} = \sum_{i<j}^{N} E_{ij}(r_{ij}) \tag{6}$$

where i and j indicate interacting amino acids i and j, and equals

$$E_{ij} = \begin{cases} E_{rep}, & \text{for } r_{ij} < R_{ij}^{rep} \\ \varepsilon_{ij}, & \text{for } R_{ij}^{rep} \leq r_{ij} \leq R_{ij} \text{ and } \varepsilon_{ij} \geq 0 \\ f\varepsilon_{ij}, & \text{for } R_{ij}^{rep} \leq r_{ij} \leq R_{ij} \text{ and } \varepsilon_{ij} < 0 \\ 0, & \text{for } r_{ij} \geq R_{ij} \end{cases} \tag{7}$$

The radius of repulsion, R_{ij}, depth ε_{ij} and limits R_{ij} of the square-well width are publicly available by anonymous ftp (Kolinski and Skolnick, 1995). Attractive interactions are modified by a factor

$$f = 1 - \{\cos^2[\angle(\mathbf{u}_i, \mathbf{u}_j)] - \cos^2(20°)\}^2 \tag{8}$$

which is dependent upon the angle between the vector $\mathbf{u}_i = \mathbf{r}_{i+2} - \mathbf{r}_{i-2}$ and the corresponding vector $\mathbf{u}_j = \mathbf{r}_{j+2} - \mathbf{r}_{j-2}$ in order to induce proper supersecondary structure packing. The angle $\angle(\mathbf{u}_i,\mathbf{u}_j)$ represents the relative orientation of the secondary structure in the vicinity of the *i*th residue with respect to the secondary structure surrounding the *j*th residue, and 20° is the most probable packing angle of helices.

The E^{NN} supplemental term was designed to reproduce the occurrence of protein-like side-chain contact maps of globular proteins. An artificial neural network with error back-propagation has been trained to recognize frequently occurring 7 × 7 fragments of side chain contact maps (Milik *et al.*, 1995). For each pair *ij*, if the 7 × 7 fragment of the side chain contact map centered at *ij* is recognized by the neural network as

Table I. Sequences of all *Staphylococcal* protein A three-helix bundle domains and of the retro-protein A (backwardly read B domain sequence).

No.	E	D	A	B	C	retro
10	Gln	Gln	Gln	Gln	Gln	Asn
11	Gln	Gln	Gln	Gln	Gln	Leu
12	Asn	Ser	Asn	Asn	Asp	Lys
13	Ala	Ala	Ala	Ala	Ala	Lys
14	Phe	Phe	Phe	Phe	Phe	Ala
15	Tyr	Tyr	Tyr	Tyr	Tyr	Glu
16	Gln	Glu	Glu	Glu	Glu	Ala
17	Val	Ile	Ile	Ile	Ile	Leu
18	Leu	Leu	Leu	Leu	Leu	Leu
19	Asn	Asn	Asn	His	His	Asn
20	Met	Met	Met	Leu	Leu	Ala
21	Pro	Pro	Pro	Pro	Pro	Ser
22	Asn	Asn	Asn	Asn	Asn	Gln
23	Leu	Leu	Leu	Leu	Leu	Ser
24	Asn	Asn	Asn	Asn	Thr	Pro
25	Ala	Glu	Glu	Glu	Glu	Asp
26	Asp	Ala	Glu	Glu	Glu	Asp
27	Gln	Gln	Gln	Gln	Gln	Lys
28	Arg	Arg	Arg	Arg	Arg	Leu
29	Asn	Asn	Asn	Asn	Asn	Ser
30	Gly	Gly	Gly	Gly	Gly	Gln
31	Phe	Phe	Phe	Phe	Phe	Ile
32	Ile	Ile	Ile	Ile	Ile	Phe
33	Gln	Gln	Gln	Gln	Gln	Gly
34	Ser	Ser	Ser	Ser	Ser	Asn
35	Leu	Leu	Leu	Leu	Leu	Arg
36	Lys	Lys	Lys	Lys	Lys	Gln
37	Asp	Asp	Asp	Asp	Asp	Glu
38	Asp	Asp	Asp	Asp	Asp	Glu
39	Pro	Pro	Pro	Pro	Pro	Asn
40	Ser	Ser	Ser	Ser	Ser	Leu
41	Gln	Gln	Gln	Gln	Val	Asn
42	Ser	Ser	Ser	Ser	Ser	Pro
43	Ala	Thr	Ala	Ala	Lys	Leu
44	Asn	Asn	Asn	Asn	Glu	His
45	Val	Val	Leu	Leu	Ile	Leu
46	Leu	Leu	Leu	Leu	Leu	Ile
47	Gly	Gly	Ser	Ala	Ala	Glu
48	Glu	Glu	Glu	Glu	Glu	Tyr
49	Ala	Ala	Ala	Ala	Ala	Phe
50	Gln	Lys	Lys	Lys	Lys	Ala
51	Lys	Lys	Lys	Lys	Lys	Asn
52	Leu	Leu	Leu	Leu	Leu	Gln
53	Asn	Asn	Asn	Asn	Asn	Gln

Only amino acids that correspond to well defined secondary structure for the experimental B domain structure are shown.

being protein-like, then the pair interaction well depth is modified in the following way:

$$\varepsilon_{ij}^{new} = 0.9\varepsilon_{ij} + 0.1\tilde{\varepsilon}_{ij} \quad (9)$$

where

$$\tilde{\varepsilon}_{ij} = \frac{\sum_{7\times 7} \varepsilon_{kl} c_{kl}}{\sum_{7\times 7} c_{kl}} \quad (10)$$

and the summation in Equation 10 is performed over the 7 × 7 fragment of the appropriate contact map; $c_{kl} = 1$ if side chains k and l are in contact and 0 otherwise. Therefore, the neural network term simulates the average effective many-body component of the potential energy responsible for the mutual packing of super secondary structure elements.

Previously, the same lattice model with a similar interaction scheme was successfully applied by Kolinski, Skolnick and co-workers to the simulation of the folding process of small helical proteins (Kolinski and Skolnick, 1994b), coiled coils (Vieth *et al.*, 1994) and crambin (Kolinski and Skolnick, 1994b). However, here the rotamer energy (Kolinski and Skolnick, 1994a) has not been used, and the many-body component of Kolinski and Skolnick (1994a) has been replaced by the neural network, packing regularizing term (Milik *et al.*, 1995). Exactly the same model has been used in the recent analysis of protein A mutations (Olszewski *et al.*, 1995).

The lattice models obtained in the course of the simulations are subsequently transformed into full atom models in order to test the consistency of our results at the atomic resolution level. All atom models are built from lattice C^{α} backbone structures of retro-protein A using the following procedure. First, the complete backbone and C^{β} carbons positions are reconstructed using the method of Milik,M., Kolinski,A. and Skolnick, J. (unpublished), which is based on a statistical analysis of the peptide plate orientation with respect to three consecutive C^{α} virtual bonds. All-atom models were then completed by rebuilding the side chains using the CHARMM package (Brooks *et al.*, 1983). The resulting structures were initially relaxed *in vacuo* using the CHARMM all-atom potential [PARAM19 parameter set with polar hydrogens (Brooks *et al.*, 1983)] and then relaxed in an 8 Å water shell. We used the TIP3P water model (Jorgensen *et al.*, 1983) and periodic boundary conditions to prevent the waters from evaporating. The protein has been almost completely immersed in a box containing about 800 water molecules. For each structure, we performed a few iterations of a relaxation procedure that consisted of ten heating and ten cooling MD simulations. Each heating cycle starts at 50 K and ends at 700 K during the 4 ps cycle; then, the system is cooled to 50 K during the 5 ps

Table II. Average energies for retro-protein A in isothermal simulations corresponding to the native topology $\bar{E}^{nat}$ and to the mirror image topology basin $\bar{E}^{inv}$, respectively ($\Delta\bar{E}^{nat-inv} = \bar{E}^{nat} - \bar{E}^{inv}$)

T	$\bar{E}^{nat}$	$\bar{E}^{inv}$	$\Delta\bar{E}^{nat-inv}$
1.0	−187.4	−182.2	−5.2
0.9	−219.9	−207.5	−12.4
0.8	−252.2	−239.1	−13.1

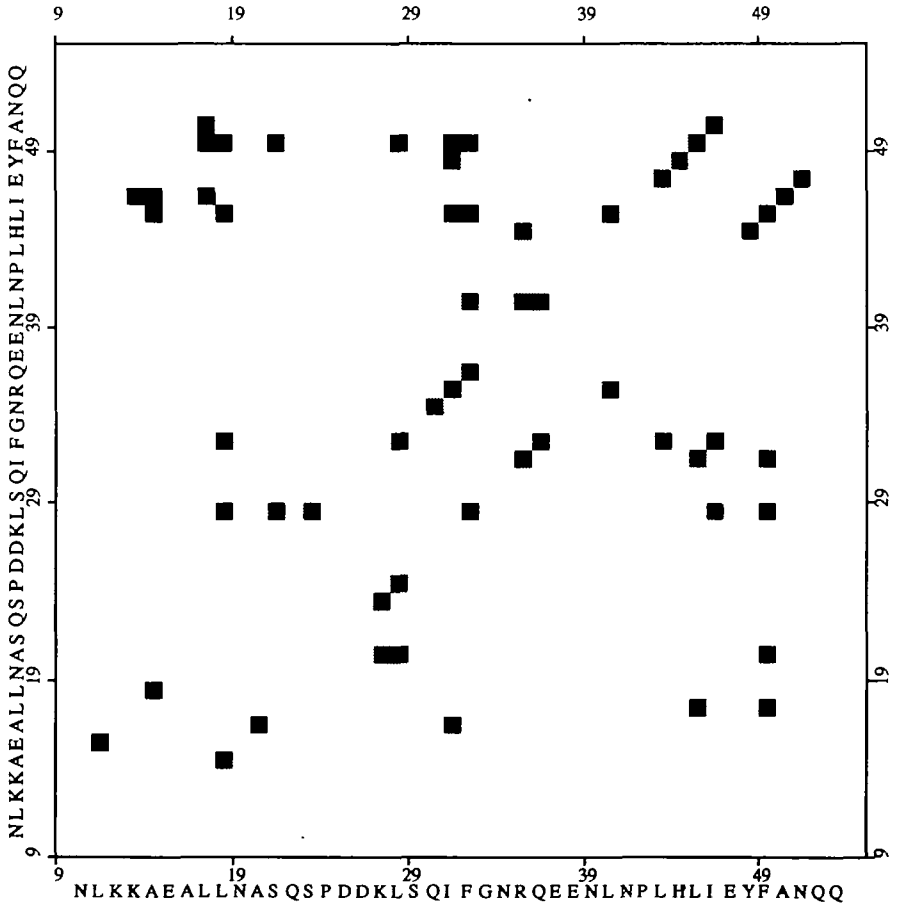

Fig. 3. Average contact map for the retro-protein A in the native topology (above diagonal) and in the mirror image topology (below diagonal).

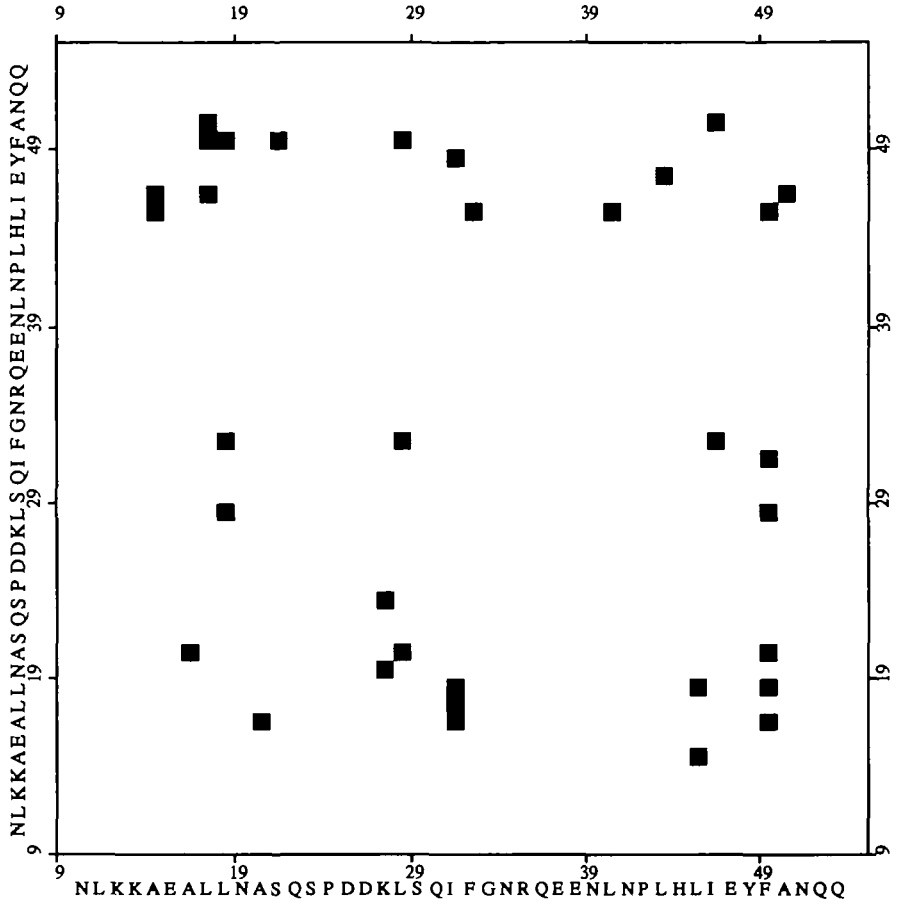

Fig. 4. Average contact retention time map for the retro-protein A in the native topology (above diagonal) and in the mirror image topology (below diagonal).

molecular dynamics run. The averaged structure from the previous iteration is used as a starting point to the next iteration, and the procedure stops when the protein structural changes are within 1.0 Å from the previous iteration structure. During the initial stages of the relaxation procedure, the hydrophobic core of the protein is kept compact by NOE-like constraints between pairs of residues that in the lattice model exhibited long-time contacts. Also, the well defined helical fragments, as identified by the lattice simulations, are initially constrained, in order to allow for the more efficient packaging of the initial structures. All the applied constraints are finally relaxed. In the last iteration, the constraints are not applied, and all side chains are allowed to evolve freely. The final structure is energetically minimized; then, the Kabsh–Sander (Kabsh and Sander, 1983) analysis of the secondary structure in the resulting conformation is performed.

Results and discussion

The retro-protein A sequence has been constructed by the backward reading of the B domain of the protein A sequence. The fragment of the sequence that corresponds to a well defined three-helix bundle motif in the original B domain was studied. The numeration of residues in the retro-protein A retro-sequence changes to the corresponding residues of the native sequence of the protein A, i.e. residue k in the retro-sequence corresponds to residue $54 - k$ in the native sequence (see Table I). The retro-protein A sequence as a whole exhibits low similarity to protein sequences listed in the SWISSPROT (Bairoch and Boeckmann, 1994).

First, to establish the ability of the retro-protein A to acquire a compact, folded conformation, we performed a series of 15 folding experiments, starting from the random fragments of various globular proteins. We performed simulated annealing Monte Carlo over a temperature range from 1.55 to 1.00 (in kT units), and the temperature was lowered linearly during the Monte Carlo run. Out of 15 folding simulations, 12 ended up in three-helix bundle topologies. In nine cases, the topology of the final structure corresponded to the solution structure of the original B domain of protein A (native topology). Three folding simulations produced the mirror image topology of the three-helix bundle. The remaining three can be characterized as a two-helix hairpin with the third helix stretched randomly away from the hairpin. They were obviously incorrectly packed and did not have well defined, long-lasting contacts. A repeated simulated annealing procedure initiated from those structures led in one case to the native topology three-helix bundle, but two runs preserved the initial topology. Isothermal simulations starting from those structures have an average energy ~20–25 kT higher than three helix bundles owing to the large increase in their E_{one} energy (which also indicates lack of compactness). Thus, the folding simulations strongly suggest that the native state of the retro-protein A is a three-helix bundle, but at this point could not definitively differentiate between the two chiral forms of the three-helix bundle topology.

Further clarification of the above result was sought by examining the behavior of the retro-protein A during long

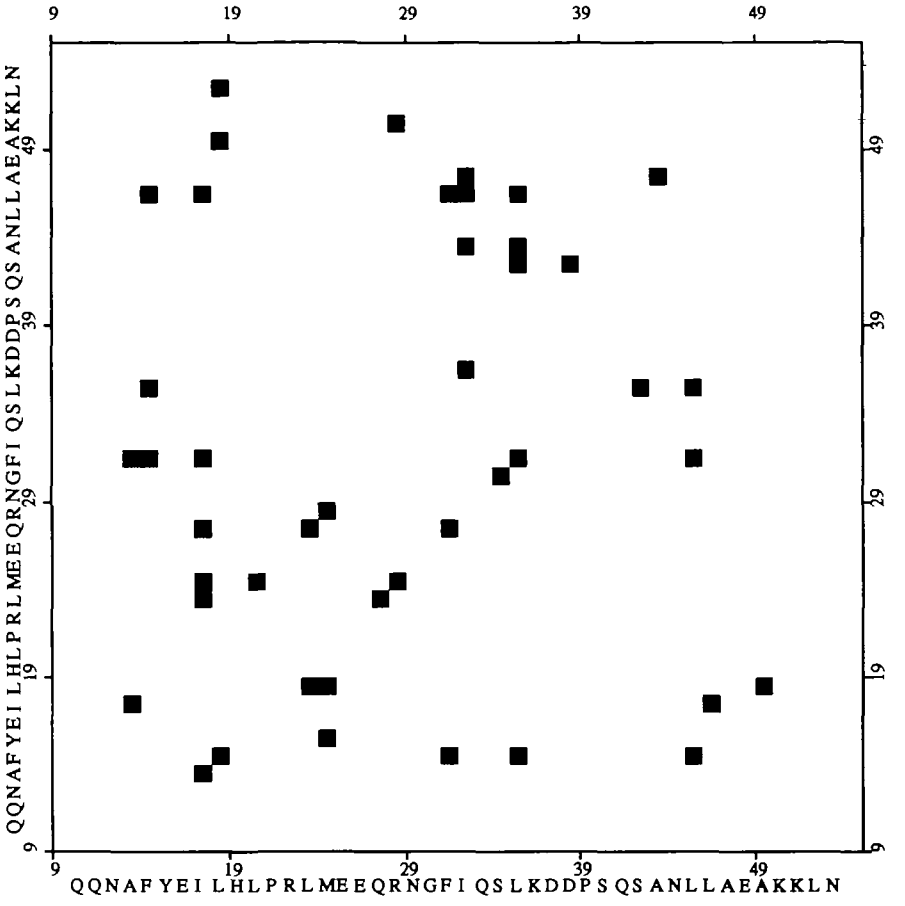

Fig. 5. Average contact map for the native sequence of the B domain of the protein A in the native topology (above diagonal) and in the mirror image topology (below diagonal).

isothermal stability simulations with the starting structures located in the native topology basin or the mirror image topology basin. In total, we performed 10 simulations (4×10^6 Monte Carlo steps each) for both the native and the mirror image topologies at a temperature 1.0 (kT). For the native topology simulations, all runs were stable (i.e. did not leave the native topology basin) and the average energy for those runs was −187.4. In the mirror image topology case, we noticed a tendency either to unfold or to flip to the native three-helix bundle topology. Nevertheless, the average energy calculated for those runs that stayed in the mirror topology basin was −182.2. The difference between the lowest energies ever found for both topologies was even greater; the minimal energies were −209.6 and −197.9 for the native topology and the mirror image topology, respectively. Since the average energy difference between the two topologies at $T = 1.0$ was not conclusive, further simulations at $T = 0.9$ and 0.8 were performed. The average energy differences between the native topology and the mirror image topology were −12.4 and −13.1 kT for $T = 0.9$ and 0.8, respectively (Table II). Additionally, the r.m.s. deviation of the C^{α}s from the average structure was 1.69 and 2.41 Å for the native topology and mirror image topology, respectively, which suggests the mirror image topology basin is broader and less well defined than the native one. The lower the temperature, the less frequent is the flipping of the mirror image topology to the native topology. This substantiates the assumption that the system can be trapped in the mirror image topology basin. Moreover, based on both energetic considerations and structural uniqueness, the tendency of the retro-protein A to acquire the native topology has been confirmed. However, our results do not preclude the possibility that at higher temperatures the native topology structure is in equilibrium with a molten globule-like, mirror image three-helix bundle.

Average contact maps for the native and the mirror image structures for the retro-protein A are presented in Figure 3. For the native topology, the packing of helices I and II on helix III is clearly seen, and seven long-lived contacts between helices I and III can also be found. Analogous behavior can be noted for the mirror image topology, but the contacts are much less persistent, which can be seen on the map of average contact retention time (Figure 4). For the mirror image topology, the contact retention times indicate that there are much less persistent contacts between helices I and III than in the case of the native topology. In the mirror image topology, contacts are constantly forming and breaking; therefore, it more closely resembles a molten globule than a native state. In contrast, in the native topology, well defined and long-lived contacts between buried residues suggest that it is the native state for the retro-protein A. The average volume of the mirror image topology structure was 5.5% greater than the average volume of the native topology structure, i.e. the mirror image structure is slightly swollen, which is consistent with the suggestion that it has some molten globule character. Moreover, four of the seven contacts between helices I and III involving residues Leu17, Leu18, Ser21, Ile46 and Phe49 are also present

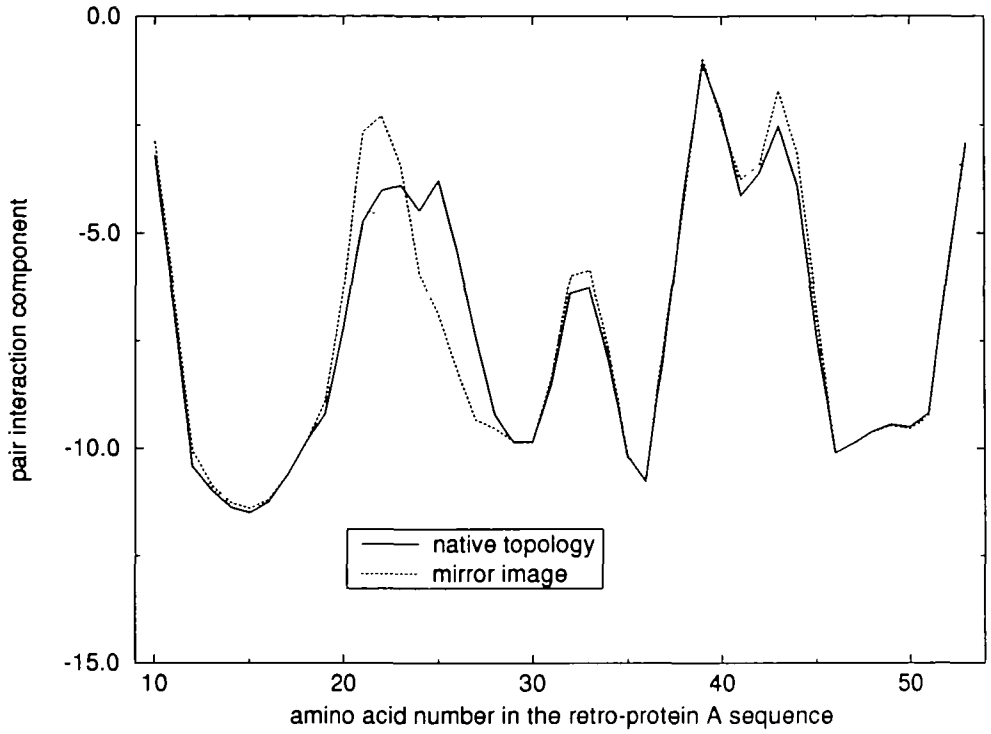

Fig. 6. Decomposition of the average pair interaction energy into one-body terms corresponding to the consecutive amino acids in the retro-sequence of protein A.

in the native structure of the native B domain of protein A (Figure 5). In addition, contacts of Leu45 (helix III) with Ile31 and Phe32 (helix II), and also Phe32 and Leu28 with Leu18 (helix I), are invariant with respect to the retro-transition. In total, eight long-lived contacts (i.e. over half of the total persistent contacts) that contribute to the stabilization of the native three-helix bundle topology are conserved with respect to the retro-transition. However, the decomposition of the pair interaction energy terms into single residue components does not reveal significant differences between the native and the mirror image topologies (Figure 6). The average E^{pair} contribution per side chain at $T = 0.8$ is -7.23 and -7.21 for the native topology and the mirror image topology, respectively. Hence, although the proper hydrophobic core packing of the retro-protein A is a necessary condition to assemble the folded structure, we conclude that it is not sufficient to direct the protein toward the native fold during the folding process.

Although the packing of the native topology is similar to the packing of the original protein A, the changes in the secondary structure of the retro-protein A are significant. The first helix of the retro-protein A is capped by the Asn10 at the N-terminus and by the Gln22 and Ser23 at the C-terminus. The second turn in the protein A, which becomes the first in the retro-protein A, is preserved, since the sequence Ser23–Pro24–Asp25 mimics the turn tendencies from real proteins. Asp26 residue caps the N-terminus of the second helix, which ends with Arg35 and Gln36 as C-terminal residues. On the other hand, the first turn region from the native protein A is no longer a turn in the retro-protein; instead, Asn31 and Pro42 initiate the third helix, which is in agreement with the preferences for N-cap residues. The third helix seems to be well capped at the C-terminus by the asparagine and two glutamines.

A number of the secondary structure prediction methods have been applied to the sequence of the retro-protein A (Levin

```
  Res. number          10        20        30        40        50
                       |         |         |         |         |
B domain sequence      QQNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLN

Gibrat method          HHHHHHHHHHCCCCCHHHHHHEEECCCCCCCHHHHHHHHHHHH
Levin method           CTCEECCEEECTTCCHHCHTHCEHEECCCSCHHHHHHHHHHCC
DPM method             TCTHHHHEEHHCCCCCCHCCCEEHCHCCTCTCTHHHHHHHHHCC
SOPMA method           HHHHEEEEEECCCCCHHHCCHHHECCCCCCCHHHHHHHHHHC
PhD method             CHHHHHHHHCCCHHHHHHHHHHHHCCCHHHHHHHHHCC
lattice model          HHHHHHHHHHTTTTTHHHHHHHHHHHHTTTTHHHHHHHHHHH
NMR structure          HHHHHHHHHHTTTTTHHHHHHHHHHHHTTTTHHHHHHHHHHH

  Res. number          10        20        30        40        50
                       |         |         |         |         |
retro-protein A        NLKKAEALLNASQSPDDKLSQIFGNRQEENLNPLHLIEYFANQQ

Gibrat method          HHHHHHHHHHCCCCCHHEEEECCCHHCCCHHHHHHHHHHH
Levin method           TCCHHHHHHHCCCCCHHHHHHHCCCCTCCCHHHHHHHCTTC
DPM method             CCHHHHHHHHTCTTTCCHCEETCCHCCCCTCHHEEHHHHCCC
SOPMA method           HHHHHHEEHHCCCCCCHHHHHHCCCCCCCCCCEHHHHHCCC
PhD method             CCCHHHHHHHHCCCCCHHHHHHHHHCCCCCCCHHHHHHHCC
lattice model          CHHHHHHHHHHHCTTTHHHHHHHHHHTTTTHHHHHHHHCCC

NMR structure          HHHHHHHHHHTTTTHHHHHHHHHHTTTTTHHHHHHHHH
B domain sequence      NLKKAEALLNASQSPDDKLSQIFGNRQEENLNPLHLIEYFANQQ
                       |         |         |         |         |
  Res. number          50        40        30        20        10
```

Fig. 7. Summary of the secondary structure predictions for the B domain of protein A and for the retro-sequence based on the B domain of the protein A. The methods reported include the Gibrat method (Gibrat *et al.*, 1987), Levin method (Levin *et al.*, 1986), DPM method (Deleage and Roux, 1987), SOPMA method (Geourjon and Deleage, 1994, 1995) and PhD method (Rost and Sander, 1994). The results of lattice Monte Carlo simulation are also reported, and in the case of the B domain those based on the NMR structure are also presented. To facilitate the secondary structure comparison, the B domain sequence together with the NMR structure is repeated backwards at the bottom of the figure, so that the pattern of amino acid side chains exactly matches the retro-B domain.

et al., 1986; Deleage and Roux, 1987; Gibrat *et al.*, 1987; Rost and Sander, 1994; Geourjon and Deleage, 1995). All of them, in general, predict the existence of three helices, although the helical termini locations vary (Figure 7). Nevertheless, fragments Arg12–Ala20, Arg27–Gly33 and Leu34–Ala50 are predicted as helical by nearly all methods. The PhD method predictions are consistent with lattice simulations for the second helix termini and the C-terminus of the first helix. The

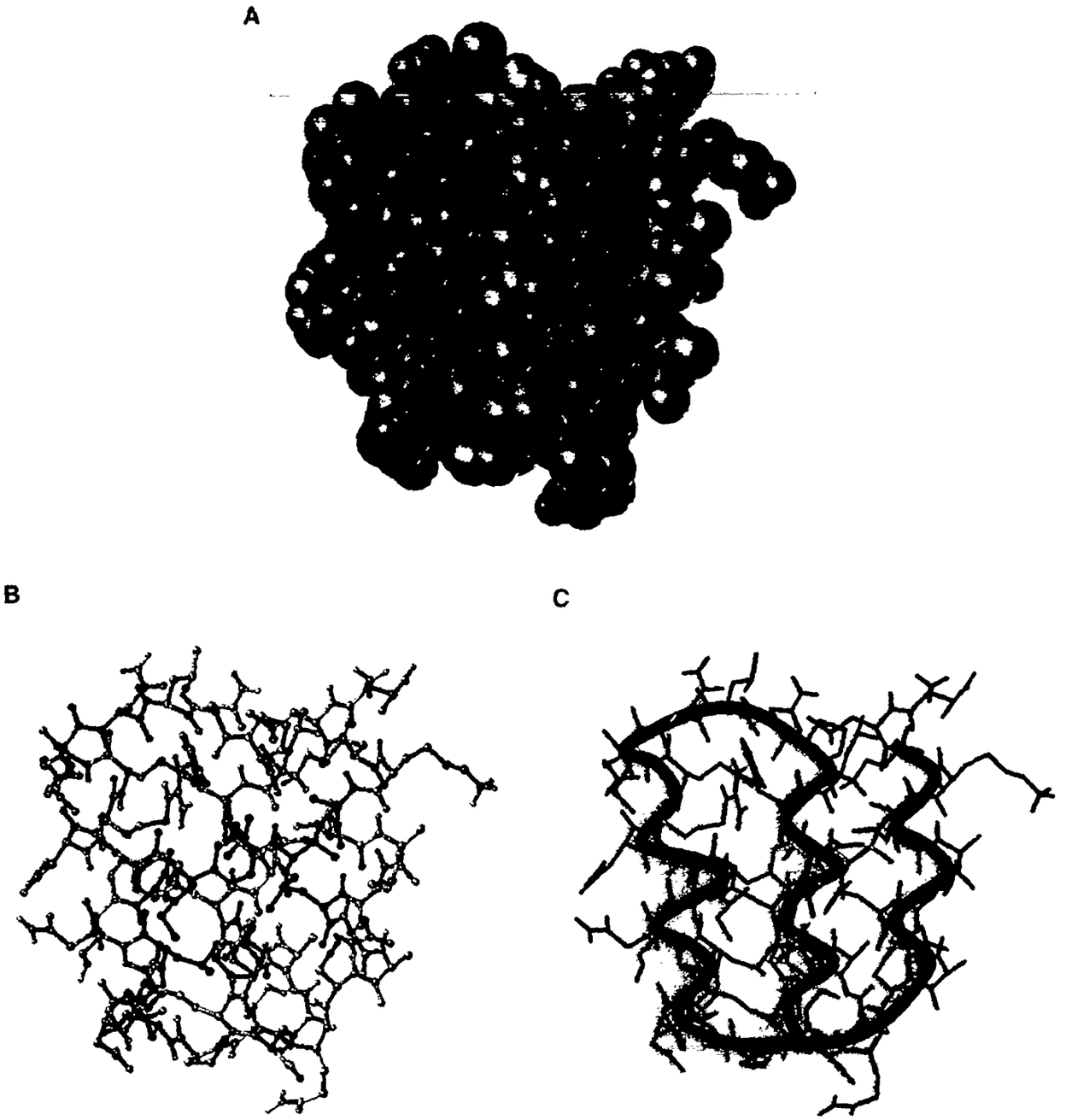

Fig. 8. (a) All-atom model of the retro-B domain of protein A in a space-filling representation. The hydrophobic core (dark grey) is almost completely covered. (b) All-atom model of the retro-B domain of protein A in a ball-and-stick representation. The hydrophobic core is presented in dark grey and the other amino acids are in light grey. (c) All-atom model of the retro-B domain of protein A with a ribbon tube showing the three-helix bundle topology.

lattice prediction for the N-terminus of the third helix is also predicted by Gibrat *et al.* (1987) and Levin *et al.* (1986). Also, the DPM method prediction of the localization of the first turn corresponds to that of the lattice model. The middle helix in the retro-protein A is shorter than the corresponding helix in the native B domain (Figure 7). The first turn in the retro-protein becomes broader than the second turn in the B domain and the opposite tendency can be noticed for the other pair of corresponding turns.

For each folding simulation run that rendered the retro-protein A in the native topology of the protein A, an all-atom model building procedure was performed in order to obtain a more detailed view of the retro-protein packing and to preclude the possibility of incorrect packing (e.g. due to the steric overlap that cannot be seen in the lattice model). The protocol for rebuilding all-atom models described in the Methods section was used. During the relaxation procedure, we noticed that the secondary structure became more regular as the

retro-molecule adjusted its hydrophobic core packing. The hydrophobic core of the final structures was well packed and surrounded by solvent-exposed amino acids (Figure 8a and b). A few hydrophobic amino acids are exposed, but this also takes place in the original protein A, since the molecule is too small to accommodate all of its hydrophobic amino acids in the protein core. According to the Kabsh–Sander analysis of the resulting structures (Figure 8c), the first helix usually starts at Leu11 or Lys12 and ends at Ala20 or Ser21, which agrees well with the lattice model and with secondary structure predictions. Asp26 initiates the second helix, but its C-terminus is not well defined. Depending on the starting point for the all-atom model rebuilding, the second helix may propagate up to Glu38 or end at Phe32. The longer the second helix is, the stronger is its tendency to slim and acquire a 3_{10}-helix shape in the last turn. The third helix is always initiated by Pro42 and is usually terminated at Ala50. Thus, overall, the all-atom models are consistent with the lattice model of the retro-protein A.

Conclusions

A three-dimensional structure of the new protein generated by the backward reading of the B domain of *Staphylococcal* protein A has been determined using the protein lattice model approach. The retro-protein A is predicted to acquire a well defined native-like tertiary structure having the three-helix bundle topology. The three-helix bundle topology has two 'chiral isomers', one corresponding to the native structure of the native sequence of protein A and the other to its topological mirror image. The model predicts that the topology adopted in the native sequence of protein A is also preferred by the retro-protein A. This finding is in contrast to previous suggestions that the retro-protein might acquire the mirror image structure of the original protein. The hydrophobic core contacts in the retro-protein A are, to a large extent, conserved. This observation suggests that hydrophobic interactions play an important role in the determination of the topology of the protein A and the retro-protein A. However, the pair interaction contribution to the total energy is not able by itself to distinguish between chiral alternatives of three-helix bundle topology. The secondary structure elements also shift their positions with respect to the structure of the original protein to accommodate the local secondary structure preferences. As a result, the retro-protein A in the native topology of the B domain of protein A has a lower energy than in the mirror image topology. Although our results constitute a fairly strong indication of the conservation of the global fold with respect to the backward reading of the protein sequence, the demonstration of their validity awaits experimental verification.

Acknowledgements

We thank Professor Lucjan Piela for seminal discussions. We gratefully acknowledge NIH grant GM-37408 and the Joseph Drown Foundation for their partial support of this research. A.K. is an International Research Scholar of the Howard Hughes Medical Institute.

References

Anfinsen,C.B. (1973) *Science*, **181**, 223–230.
Bairoch,A. and Boeckmann,B. (1984) *Nucleic Acids Res.*, **22**, 3578–3580.
Bowie,J.U., Reidhaar,O.J.F., Lim,W.A. and Sauer,R.T. (1990) *Science*, **247**, 1306–1310.
Brooks,B.R., Bruccoleri,R., Olafson,B., States,D., Swaminathan,S. and Karplus,M. (1983) *J. Comput. Chem.*, **4**, 187–217.
Chan,H.S. and Dill,K.A. (1993) *J. Chem. Phys.*, **99**, 2116–2127.
Chou,K.C. (1995) *Proteins*, **21**, 319–344.
Corey,R.B. and Pauling,L. (1953) *Proc. R. Soc. Lond.*, **B141**, 10–20.
Deisenhofer,J. (1981) *Biochemistry*, **20**, 2361–2370.
Deleage,G. and Roux,B. (1987) *Protein Engng*, **1**, 239–294.
Geourjon,C. and Deleage,G. (1994) *Protein Engng*, **7**, 157–164.
Geourjon,C. and Deleage,G. (1995) *Comput. Appl. Biosci.*, **9**, 197–199.
Gibrat,J.F., Garnier,J. and Robson,B. (1987) *J. Mol. Biol.*, **198**, 425–444.
Godzik,A., Kolinski,A. and Skolnick,J. (1993a) *J. Comput.-Aided Mol. Des.*, **7**, 397–438.
Godzik,A., Kolinski,A. and Skolnick,J. (1993b) *J. Comput. Chem.*, **14**, 1194–1202.
Goodman,M. and Chorev,M. (1979) *Acc. Chem. Res.*, **12**, 1–14.
Gouda,H., Torigoe,H., Saito,A., Sato,M., Arata,Y. and Schimada,I. (1992) *Biochemistry*, **40**, 9665–9672.
Guptasarma,P. (1992) *FEBS Lett.*, **310**, 205–210.
Hao,M.H. and Scheraga,H.A. (1994) *J. Phys. Chem.*, **98**, 4940.
Jorgensen,W.L., Chandrasekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) *J. Chem. Phys.*, **79**, 926–935.
Kabsh,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Kolinski,A. and Skolnick,J. (1994a) *Proteins*, **18**, 338–352.
Kolinski,A. and Skolnick,J. (1994b) *Proteins*, **18**, 353–366.
Kolinski,A. and Skolnick,J. (1995) available at scripps.edu via anonymous ftp in the /pub/skolnick/mutant directory.
Levin,J.M., Robson,B. and Garnier,J. (1986) *FEBS Lett.*, **205**, 303–308.
Liwo,A., Pincus,M.R., Wawak,R.J., Rackovsky,S. and Scheraga,H.A. (1993) *Protein Sci.*, **2**, 1715–1731.
Lyons,B., Tashiro,M., Cedergren,L. and Montelione,G. (1993) *Biochemistry*, **32**, 7839–7845.
Metropolis,N., Rosenbluth,A., Rosenbluth,M., Teller,A. and Teller,E. (1953) *J. Chem. Phys.*, **21**, 1087–1092.
Milik,M., Kolinski,A. and Skolnick,J. (1995) *Protein Engng*, **8**, 225–236.
Milton,R.C.D., Milton,S.C.F. and Kent,S.B.H. (1992) *Science*, **256**, 1445–1448.
Nielsen,K.J., Alewood,D., Andrews,J., Kent,S.B.H. and Craik,D.J. (1994) *Protein Sci.*, **3**, 291–302.
Nilsson,B., Moks,T., Jansson,B., Abrahamsen,L.A., Elmblad,E.H., Henrichson,C., Jones,T. and Uhlen,M. (1987) *Protein Engng*, **1**, 107–113.
Oldfield,T.J. and Hubbard,R.E. (1994) *Proteins*, **18**, 324–337.
Olszewski,K.A., Kolinski,A. and Skolnick,J. (1995) *Proteins*, in press.
Park,B.H. and Levitt,M. (1985) *J. Mol. Biol.*, **249**, 493–507.
Presta,L.G. and Rose,G.D. (1988) *Science*, **240**, 1632–1641.
Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) *J. Mol. Biol.*, **7**, 95.
Richards,F. (1977) *Annu. Rev. Biophys. Bioengng*, **6**, 151–176.
Richardson,J.S. and Richardson,D.C. (1988) *Science*, **240**, 1648–1652.
Rose,G.D. and Wolfden,R. (1993) *Annu. Rev. Biophys. Biomol. Struct.*, **22**, 381–415.
Rost,B. and Sander,C. (1994) *Proteins*, **19**, 55–72.
Šali,A., Shakhnovich,E.I. and Karplus,M. (1994) *J. Mol. Biol.*, **235**, 1614–1636.
Scheraga,H.A. (1968) *Adv. Phys. Org. Chem.*, **6**, 103–184.
Shakhnovich,E.I. and Gutin,A.M. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 7195–7199.
Skolnick,J. and Kolinski,A. (1989) *Annu. Rev. Phys. Chem.*, **40**, 207–235.
Socci,N.D. and Onuchic,J.N. (1994) *J. Chem. Phys.*, **101**, 1519–1528.
Vásquez,M., Némethy,G. and Scheraga,H.A. (1994) *Chem. Rev.*, **94**, 2183–2239.
Vieth,M., Kolinski,A., Brooks,III,C.L. and Skolnick,J. (1994) *J. Mol. Biol.*, **237**, 361–367.
Wilmot,C.M. and Thornton,J.M. (1988) *J. Mol. Biol.*, **203**, 221–232.

Received September 12, 1995; revised October 30, 1995; accepted October 31, 1995