BIOINFORMATICS

Vol. 20 no. 2 2004, pages 199–205
DOI: 10.1093/bioinformatics/btg391

Limited conformational space for early-stage protein folding simulation

M. Bryliński[1,3], W. Jurkowski[1,3], L. Konieczny[2] and I. Roterman[3,]*

[1]Institute of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Kraków, Poland,
[2]Institute of Medical Biochemistry and [3]Department of Bioinformatics and Telemedicine,
Collegium Medicum—Jagiellonian University, Kopernika 17, 31-501 Kraków, Poland

Received on April 14, 2003; revised on July 9, 2003; accepted on July 25, 2003

ABSTRACT

Motivation: The problem of early-stage protein folding is critical for protein structure prediction. The model presented introduces a common definition of protein structures which may be treated as the possible *in silico* early-stage form of the polypeptide chain. Limitation of the conformational space to the ellipse path on the Ramachandran map was tested as a possible sub-space to represent the early-stage structure for simulation of protein folding. The proposed conformational sub-space was developed on the basis of the backbone conformation, with side-chain interactions excluded.

Results: The ellipse-path-limited conformation of BPTI was created using the criterion of shortest distance between Phi, Psi angles in native form of protein and the Phi, Psi angles belonging to the ellipse. No knots were observed in the structure created according to ellipse-path conformational sub-space. The energy minimization procedure applied to ellipse-path derived conformation directed structural changes toward the native form of the protein with SS-bonds system introduced to the procedure.

Availability: Program 'Ellipse' to create the ellipse-path derived structure available on request: myroterm@cyf-kr.edu.pl

Contact: myroterm@cyf-kr.edu.pl

INTRODUCTION

The problem of early-stage protein folding concerns biologists as well as those engaged in the protein folding simulations. In particular, prediction of protein structure in *ab initio* methods requires definition of the starting conformation for the energy minimization procedure. Possible starting configurations (the number of them estimated for a 27 amino acids long polypeptide is about 10^{16}) were suggested in the context of free-energy surface (Dobson, 2001). The starting structures discussed there seem to be arbitrarily selected. The possible pathway (or pathways) leading from random unfolded structures to the native structure, as was suggested, depends on the number of non-bonding contacts (Dobson, 2001; Chan and Dill, 1998; Dill and Chan, 1997).

A mechanism to limit the conformational search to particular regions of the conformational space or to narrow the pathways between unfolded and native states has been suggested (Alonso and Daggett, 1998). Some simplified models have been proposed to solve this problem: the simplified structure of amino acid representation (Liwo *et al*., 2001, 2002) and limitation of the conformational space to four low-energy basins (Fernàndez *et al*., 2001, 1999; Sosnick *et al*., 2002).

Our approach (Roterman, 1995a; Jurkowski *et al*., 2003) introduces a model for a common definition of protein structures which may be treated as the possible *in silico* early-stage form of the polypeptide chain.

The early-stage forms (as may be assumed for folding *in silico*) appeared as the result of a simplified structural model based on only two geometric parameters: the V-angle, which is the dihedral angle between two sequential peptide bond planes; and the R-radius of curvature, which appeared to depend on the V-angle in a parabolic relation (for R on a log scale) (Roterman, 1995a,b). The structures selected from the whole Ramachandran map according to the model parabolic relation ($\ln R$ versus V) (0° for helical form with low R, to 180° for β-structural form with very large R) revealed an area on the map which appeared to be ellipse-shaped. This ellipse-path fragment of the Ramachandran map is assumed to represent the early-stage polypeptide structure. The structures obtained according to this criterion are based on the backbone conformation only (Hayward, 2001), excluding side chain–side chain interaction, since only the mutual orientation of the peptide bond planes was taken in consideration. This assumption is in accordance with suggestions that the backbone structure dominates in early-stage polypeptide structure formation and that side chain–side chain interaction occurs in later stages of polypeptide folding (Baldwin, 2002).

The ellipse-path links all structurally significant areas (right-handed helix, C7eq energy minimum, left-handed helix) and simultaneously reveals the simplest path for

*To whom correspondence should be addressed.

structural changes allowing α-to-β-structure transformation (Roterman, 1995b).

On the other hand, quantitative analysis of the information stored in the amino acid sequence in relation to the amount of information necessary to predict particular Phi, Psi angles as they appear in real proteins revealed that in the ellipse-path approach these two quantities become equilibrated (Jurkowski *et al*., 2003).

SYSTEMS AND METHODS

Ellipse-path derived structure creation

The BPTI (4PTI according to PDB identification) was taken for analysis.

The Phi, Psi angles were calculated for each amino acid in the polypeptide chain as they appear in native form of protein. The criterion of the shortest distance (Fig. 1) between the observed Phi, Psi and those belonging to the ellipse was taken to find the $\mathrm{Phi_e}$, $\mathrm{Psi_e}$ angles.

The ellipse-derived structure for BPTI was created using the ECEPP/3 program. The Omega dihedral angles were taken as 180° for all amino acids. The side-chain structures were created according to ECEPP/3 standards.

Energy minimization procedure

The energy minimization procedure was performed using the ECEPP/3 program for ellipse-derived structure of BPTI (Scheraga, 1992). The Phi, Psi angles were calculated for post-minimization structures. The unconstrained minimization solver with analytical gradient (Wang *et al*., 2000) was used. The values of absolute and relative function convergence tolerances were set at 1×10^{-3} and 1×10^{-5}, respectively. The energy minimization procedure was carried out both with and without properly defined disulfide bonds. The coordinates and values of the backbone dihedral angles were saved for analysis at 10-step intervals. Energy minimization procedure for BPTI was done on an SGI Origin 2000 in the computing center of TASK in Gdansk.

Two forms of energy minimization procedures were applied to ellipse-path-derived structural forms of BPTI. One of them did not include SS-bonds creation and the second one taking SS-bonds system into account.

The structure comparison

The structures (native, elliptical, post-energy-minimization) were compared using different criteria:

1. The distances between geometric center of molecule and sequential Cα atoms ($\mathrm{D_{center-C\alpha}}$) in the polypeptide chain were calculated. This plot revealed a rough degree of similarity. The polypeptide fragments distinguished according to the profile were also characterized using RMS-D calculation. The RMS-D value for selected fragments were calculated after overlapping the fragments taken from native and ellipse-derived structural form. The RMS-D value was calculated per one amino acid in polypeptide fragment.
2. The number of native non-bonding interactions was calculated for all structural forms assuming cut off distance equal to 12 Å.
3. The box large enough to contain the whole molecule was also calculated for each structural form of the protein molecule.

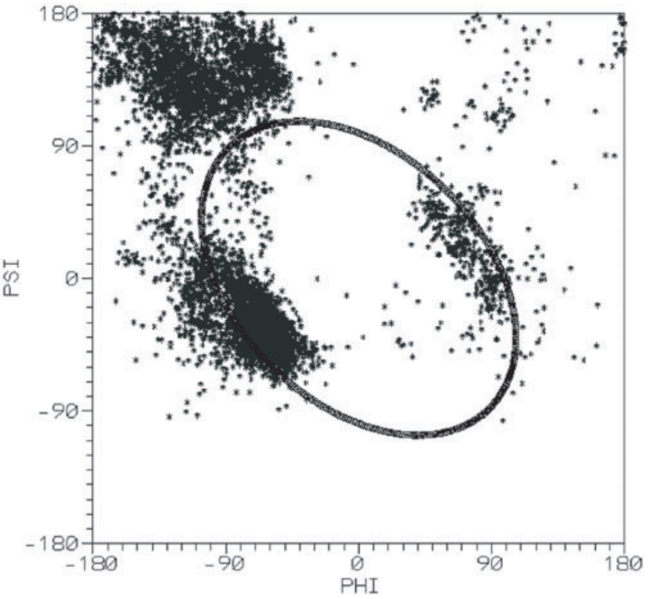

Fig. 1. The ellipse-path limited conformational sub-space in relation to Phi, Psi angles as they appear in real proteins. The shortest distance between particular Phi, Psi angles and ellipse-belonging point represents the way in which the Phi, Psi angles on the ellipse can be found.

The box size was calculated as follows: the longest Cα–Cα distance was taken as the D_Z measure (distance along z-axis), the longest Cα–Cα distance in the xy-plane was taken as the D_Y measure (distance along y-axis), and the difference between the highest and the lowest value of X was taken as the measure for the D_X box edge.

RESULTS

Phi, Psi dihedral angles changes

Presentation of Phi, Psi angle distribution exposing the range of dihedral angle change is shown for all discussed structural forms in Figure 2A for native form of BPTI, Figure 2B for ellipse-derived structure, Figure 2C for post-energy-minimization form with SS-bonds present and in Figure 2D for the post-energy-minimization structure without SS-bonds present. The similarity of Phi, Psi angles distribution can be seen between structures: native and post-energy-minimization with SS-bonds system introduced.

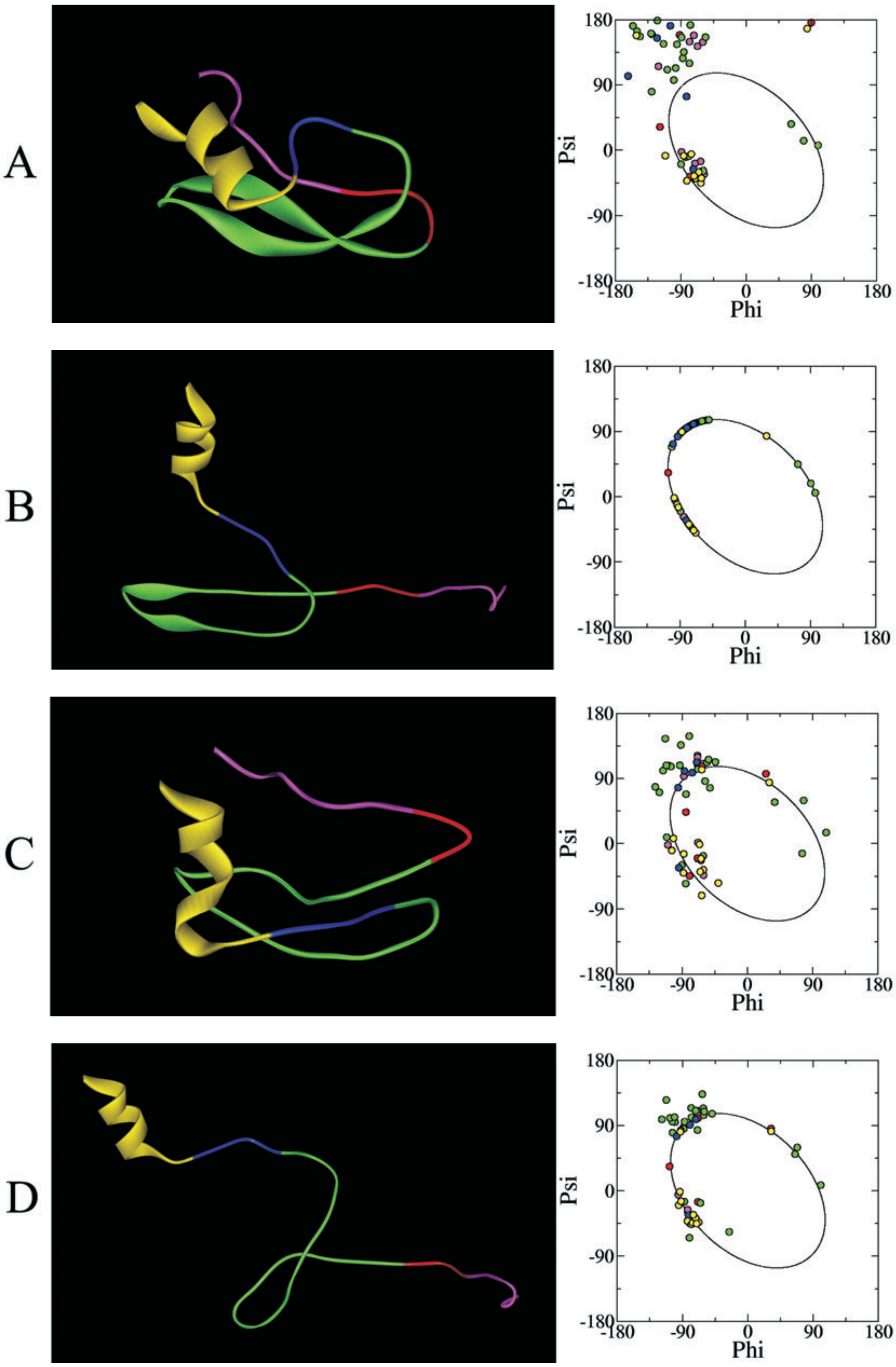

Fig. 2. The structure of BPTI and its Phi, Psi angles distribution versus ellipse-path **(A)** native form, **(B)** ellipse-based structure, **(C)** post-energy-minimization structure of BPTI with SS-bonds present in procedure, **(D)** post-energy-minimization structural form of BPTI with SS-bonds absent in procedure.

Spatial distribution of Cα atoms versus the geometrical center

The profile of the size of vectors linking the geometrical center with sequential Cα atoms gives insight into the three-dimensional relative displacements versus native form of protein (Orengo *et al*., 1999). The structural similarity may be disclosed by overlapping of lines representing two compared structures.

The parallel orientation of profiles represents similar structural forms in both compared molecules oriented in the space in different way. The increase of the vector length in respect to native structure is obvious due to extension of the structure which is always associated with the transformation from native to ellipse-path delimited structure. The profiles for all structural forms of BPTI discussed in this paper are presented in Figure 3.

Three large fragments can be distinguished. The central one containing amino acids 16–38, which is characterized by very similar regularity in $D_{center–C\alpha}$ in all compared structures. In the C-terminal fragment (47–58 amino acids), the distance is rather large in the ellipse-derived structure, although a very similar regularity of maxima and minima distribution is observed in this fragment. The N-terminal part (1–10 amino acids) seems to represent a conformation in the ellipse-derived structure very different from the native one. The short fragments where discrepancies between the native and ellipse-derived $D_{center–C\alpha}$ profiles are observed probably represent turns in the crystal structure.

The RMS-D (see Methods) values expressing the differences between native and ellipse-derived structure were calculated (the mean value for selected polypeptide fragments) for fragments distinguished by different form of $D_{center–C\alpha}$ distributions. RMS-D (Fig. 3) calculated for the distinguished parts reveals differentiation in the degree of similarity of particular fragments. It turned out that the C-terminal fragment (47–58 amino acids) represents lowest RMS-D level. The fragment with the highest RMS-D fragment (11–15 amino acids) is present where a rather significant structural change is necessary to approach the crystal structure.

The energy minimization procedure did not significantly change the ellipse structure. One should mention that no SS-bonds were present in the energy minimization procedure (Fig. 2D).

It is characteristic that the $D_{center–C\alpha}$ values for amino acids participating in SS-bond creation in the ellipse-derived structure are placed in the same distance versus the geometrical center of the molecule (Fig. 2C). It suggests, that only elbow-like action is necessary to direct the N- and C-terminal fragments to the native form.

Visual analysis

The aim of Figure 2 is the visualization of structural changes in BPTI in different conditions. Color notation differentiates particular polypeptide fragments and distinguishes them according to their similarity measured by the $D_{center–C\alpha}$ vectors profiles. The same color notation was used for Phi, Psi angles distribution and RMS-D fragments.

No signs of directing toward native form can be seen in the energy minimization procedure performed without SS-bonds. The quite good approach was reached for the same procedure with SS-bonds defined according to the natural system present in this molecule.

Non-bonding interaction

The native non-bonding interactions present in all discussed structural forms are shown in Figure 4. The significant similarity in non-bonding contacts distribution can be seen between native and post-energy-minimization with SS-bonds taken into account during energy minimization procedure. The ellipse-derived structure reveals also the presence of one part of contacts map, which is present in native form of BPTI.

The percentage of native non-bonding interactions present in ellipse-derived structure is equal to 43.66, 60.30 and 40.51% for post-energy-minimization (with and without SS-bonds) structures, respectively. This quantity depends obviously on the molecule under study (Fersht and Daggett, 2002) and is strongly related to the percentage of helical forms in analyzed structure. The ellipse-path goes through the region attributed to helical forms on Ramachandran map, what additionally explains rather high percentage of native-like non-bonding interactions.

Size of molecule change

The most critical problem concerning the relation between ellipse-derived and native structures is the question of how much the size of the molecule is changed, i.e., what degree of compactness of ellipse-derived structures is necessary to reach the native form.

The relative (versus the native form of protein) increase of box size (volume) containing the whole protein molecule (the size of which is calculated according to the procedure given in Methods) appeared to be 3.88 for ellipse-derived, post-energy-minimization with (1.05) and without SS-bonds (3.30), respectively. The significant influence of SS-bonds presence on the box size is obvious.

DISCUSSION

Experiments provide clear evidence for the existence of partially folded intermediates (Dobson and Karplus, 1999; Clarke and Waltho, 1997; Kuwajima *et al*., 1993; Dinner *et al*., 2000).

The stages distinguished as partitioning the protein folding process proposed by Ferguson and Fersht (2003) are as follows: (1) specific or non-specific chain collapse, (2) formation of secondary and tertiary structure, according to the balance of local and non-local, native and non-native interactions, (3) desolvation of the chain as it folds to a lower energy conformation. Although the sequence of events may be dependent on the protein under study (Fersht and Daggett,

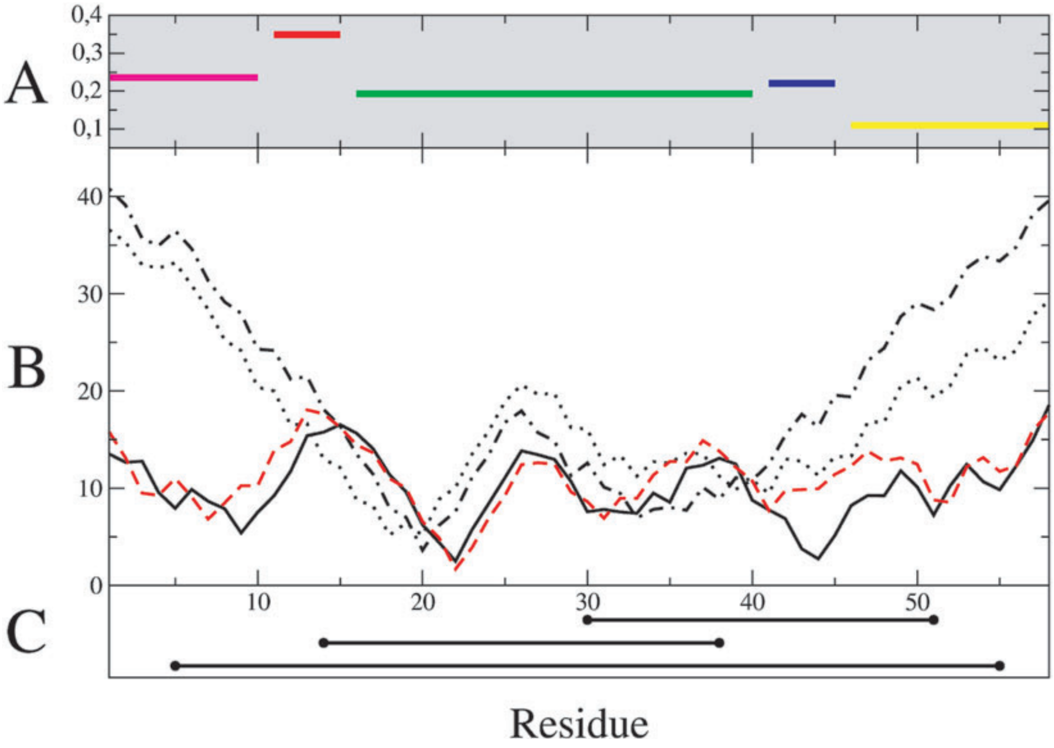

Fig. 3. The comparison of structural forms of BPTI protein molecule. **(A)** RMS-D (per residue) calculated for structurally differentiated polypeptide fragments. The fragments were defined according to the profile presented in **B**. The parallel fragments of curves represent the correct spatial orientation of the polypeptide, while the dissimilar regularity of the curve represents fragments with low similarity of the spatial orientation of the particular polypeptide fragment. **(B)** Profile representing the distribution of distances linking the geometrical center of the molecule with sequential Cα atoms. Continuous line—native form, Dotted line—ellipse-derived structure, Dashed red line—post-energy-minimization structure with SS-bonds present, Dotted/dashed line—post-energy-minimization structure with SS-bonds absent. **(C)** SS-bonds system in BPTI. The color notation introduced in all graphical presentations in this paper in accordance to fragments distinguished in this figure.

2002; Song *et al*., 1998), the chosen experimental conditions and perhaps the time regime of the experiment, the search for a universal folding mechanism may provide sufficient benchmarks for theoretical models, at least for homology proteins. The proposed model is assumed to describe the first two steps in an event sequence oriented on reaching the native structure.

The contact maps for evaluation of protein structure prediction adapted for ubiquitin folding (Sosnick *et al*., 2002) were used in the four-basins model (Fernàndez *et al*., 2001). The discrete model for conformational space limitation determined very well the nucleation point in the folding trajectory expressed by stabilization of a low number of basin changes. The model based on the ellipse path on the Ramachandran map seems to treat the initial limited conformational space in a continuous manner although the search for the final Phi, Psi angles is performed in a similar way in the two models.

The early-stage structures presented in this paper were obtained as a result of a back step with respect to the native structure of the protein. It resembles the frequently used procedure of polypeptide denaturation using molecular dynamics simulation (Fersht and Daggett, 2002; Daggett and Levitt, 1992, 1993; Finkelstein, 1997; Brooks, 1998).

The early stage of folding was also extensively studied using 200-ns fully solvated molecular dynamics simulation (Duan *et al*., 1998). The main characteristic of the intermediate or molten state of the protein in this simulation was the observation of secondary structure present only after a short period of dynamics simulation ($\approx$50 ns) starting from the denatured form. The native helical content reached $>$50%, sometimes even more than 70%. No information is given as to β-structure content. The ellipse-path based structure discussed in this paper keeps the helical structure close to the content of native conformation 90–100% (Duan *et al*., 1998). This is the consequence of the ellipse-path passing through the alpha-helical energy minimum (Kabsch and Sander, 1983). The β-structural form in ellipse-path derived structure is mediated by the structure of the C7eq energy minimum, which is characteristic for isolated amino acids when no inter-chain contacts are present.

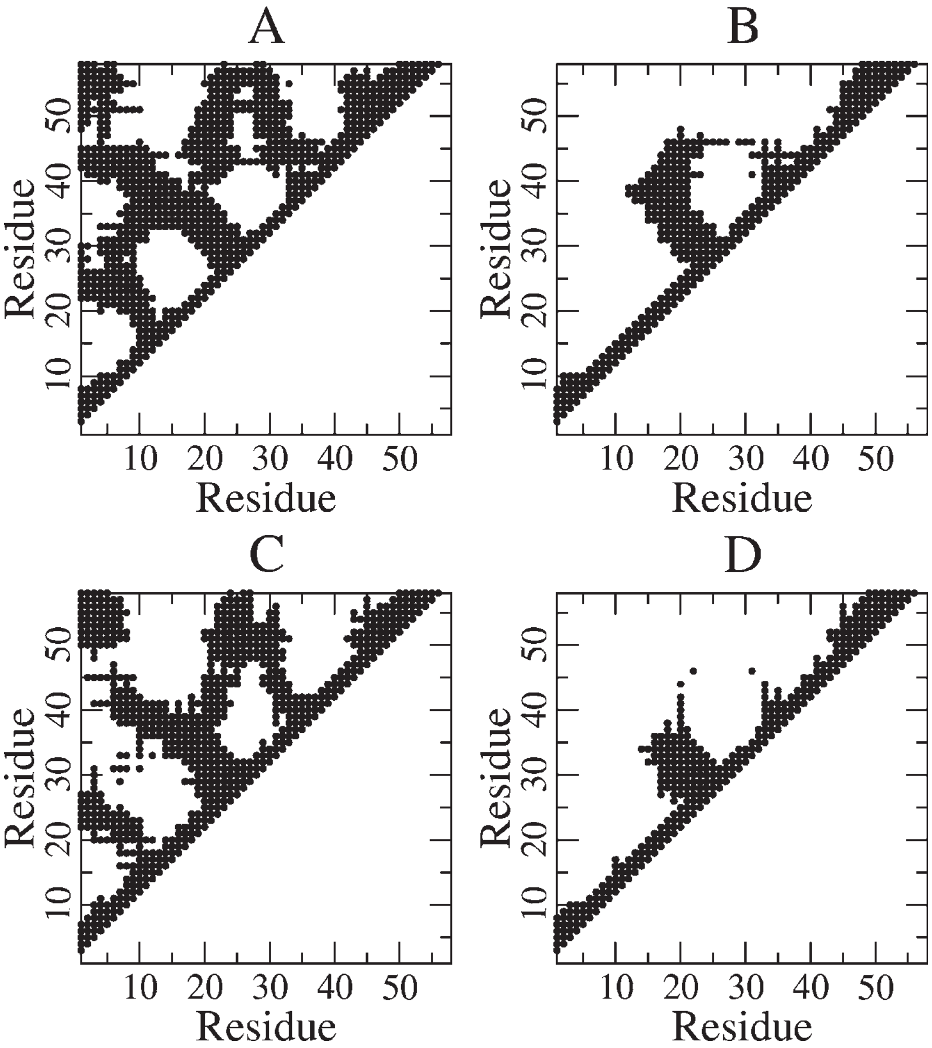

Fig. 4. Non-bonding contacts in BPTI: **(A)** Native form, **(B)** ellipse-derived structure, **(C)** post-energy-minimization with SS-bonds present, **(D)** post-energy-minimization with SS-bonds absent.

Another support for the presented model is the suggestion that low-resolution structural features rather than high-resolution detail are more helpful in addressing the fundamental physics of the folding process (Alm and Baker, 1999). This paper discusses only the initial step of folding, defining the conformation with its secondary structural elements. The next step of the folding model, complementary to the ellipse-path derived structure and describing creation of the hydrophobic center, will be published in a further paper.

REFERENCES

Alm,E. and Baker,D. (1999) Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.*, **9**, 189–196.

Alonso,D.O.V. and Daggett,V. (1998) Molecular dynamics simulations of hydrophobic collapse of ubiquitin. *Protein Sci.*, **7**, 860–874.

Baldwin,R.L. (2002) Making a network of hydrophobic clusters. *Science*, **295**, 1657–1658.

Brooks,C.L. III (1998) Simulations of protein folding and unfolding. *Curr. Opin. Struct. Biol.*, **8**, 222–226.

Chan,H.S. and Dill,K. (1998) Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins Struct. Func. Gen.*, **30**, 2–33.

Clarke,A.R. and Waltho,J.P. (1997) Protein folding pathways and intermediates. *Curr. Opin. Biotechnol.*, **8**, 400–410.

Daggett,V. and Levitt,M. (1992) Molecular dynamics simulation of helix denaturation. *J. Mol. Biol.*, **223**, 1121–1138.

Daggett,V. and Levitt,M. (1993) Protein unfolding pathways explored through molecular dynamics simulation. *J. Mol. Biol.*, **232**, 600–619.

Dill,K.A. and Chan,H.S. (1997) From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, **4**, 10–19.

Dinner,A.R., Sali,A., Smith,L.J., Dobson,C.M. and Karplus,M. (2000) Understanding protein folding via free energy surfaces from theory and experiment. *Trends Biochem. Sci.*, **25**, 331–339.

Dobson,C.M. and Karplus,M. (1999) The fundamentals of protein folding: bringing together theory and experiment. *Curr. Opin. Struct. Biol.*, **9**, 92–101.

Dobson,C.M. (2001) The structural basis of protein folding and its links with human disease. *Philos. Trans. R. Soc. Lond. B*, **356**, 133–145.

Duan,Y., Wang,L. and Kollman,P.A. (1998) The early stage of folding of villin headpiece sub domain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proc. Natl Acad. Sci. USA*, **95**, 9897–9902.

Ferguson,N. and Fersht,A.R. (2003) Early events in protein folding. *Curr. Opin. Struct. Biol.*, **13**, 75–81.

Fernàndez,A., Kostov,K. and Berry,R.S. (1999) From residue matching patterns to protein folding topographies: General model and bovine pancreatic inhibitor. *Proc. Natl Acad. Sci. USA*, **96**, 12991–12996.

Fernàndez,A., Colubri,A., Aqpigmanesi,G. and Burastero,T. (2001) Coarse semiempirical solution to the protein folding problem. *Physica. A*, **293**, 358–384.

Fersht,A.R. and Daggett,V. (2002) Protein folding and unfolding at atomic resolution. *Cell*, **108**, 573–582.

Finkelstein,A.V. (1997) Can protein unfolding simulate protein folding? *Protein Eng.*, **10**, 843–845.

Hayward,S. (2001) Peptide-plane flipping in proteins. *Protein Sci.*, **10**, 2219–2227.

Jurkowski,W., Bryliński,M., Wiśniowski,Z., Konieczny,L. and Roterman,I. (2003) The conformational sub-space in simulation of early-stage protein folding. *Proteins Struct. Func. Gen.*, in press.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kuwajima,K., Semisotnov,G.V., Finkelstein,A.V., Sugai,S. and Ptitsyn,O.B. (1993) Secondary structure of globular proteins at the early and final stages in protein folding. *FEBS*, **334**, 265–268.

Liwo,A., Czaplewski,C., Pillardy,J. and Scheraga,H.A. (2001) Cummulation-based expression for the multibody terms for the correction between local and electrostatic interaction in the united residue force field. *J. Chem. Phys.*, **115**, 2323–2347

Liwo,A., Arlukowicz,P., Czaplewski,C., Ołdziej,S., Pillardy,J. and Scheraga,H.A. (2002) A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape—Application to the UNRES force field. *Proc. Natl Acad. Sci. USA*, **99**, 1937–1942.

Orengo,C.A., Bray,J.E., Hubbard,T., LoConte,L. and Sillitoe,I. (1999) Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure and contacts prediction. *Proteins Struct. Func. Gen.*, **3**(suppl.), 149–170.

Roterman,I. (1995a) Modeling of optimal simulation path in the peptide chain folding—Studies based on geometry of alanine heptapeptide *J. Theor. Biol.*, **177**, 283–288.

Roterman,I. (1995b) The geometrical analysis of peptide backbone structure and its local deformations. *Biochimie*, **77**, 204–216.

Scheraga,H.A. (1992) Predicting three-dimensional structures of oligopeptides. *Rev. Comput. Chem.*, **3**, 73–142.

Song,J., Bai,P., Luo,L. and Peng,Z. (1998) Contribution of individual residues to formation of the native-like tertiary topology in the α-lactalbumin molted globule. *J. Mol. Biol.*, **280**, 167–174.

Sosnick,T.R., Berry,R.S., Colubri,A. and Fernàndez,A. (2002) Discriminating foldable proteins from nonfolders: when and how do they differ? *Proteins Struct. Func. Gen.*, **49**, 15–23.

Wang,J., Cieplak,P. and Kollman,P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, **21**, 1049–1074.