Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 251 (2008) 275–285
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Conservative secondary structure motifs already present in early-stage folding (*in silico*) as found in serpines family

Michal Brylinski[a,b], Leszek Konieczny[c], Andrzej Kononowicz[a], Irena Roterman[a,d,*]

[a] *Department of Bioinformatics and Telemedicine, Collegium Medicum—Jagiellonian University, Kopernika 17, 31 501 Cracow, Poland*
[b] *Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30 060 Cracow, Poland*
[c] *Institute of Biochemistry, Collegium Medicum—Jagiellonian University, Kopernika 7, 31 501 Cracow, Poland*
[d] *Faculty of Physics, Jagiellonian University, Reymonta 4, 30 060 Crakow, Poland*

Received 2 August 2007; received in revised form 31 October 2007; accepted 31 October 2007
Available online 7 November 2007

Abstract

The well-known procedure implemented in ClustalW oriented on the sequence comparison was applied to structure comparison. The consensus sequence as well as consensus structure has been defined for proteins belonging to serpine family. The structure of early stage intermediate was the object for similarity search. The high values of $W_{sequence}$ appeared to be accordant with high values of $W_{structure}$ making possible structure comparison using common criteria for sequence and structure comparison.

Since the early stage structural form has been created according to limited conformational sub-space which does not include the β-structure (this structure is mediated by C7eq structural form), is particularly important to see, that the C7eq structural form may be treated as the seed for β-structure present in the final native structure of protein.

The applicability of ClustalW procedure to structure comparison makes these two comparisons unified.

Keywords: Structural similarity; Consensus structure; Consensus sequence; Comparative analysis; Structural alphabet

1. Introduction

The increasing size of sequence and structure data bases demands methods that uniformly evaluate similarity for homology-based structure prediction. In the post-genomic era, when complete knowledge of the full proteome of an organism is available, a theoretical method allowing prediction of protein structure, especially one based on homology, is urgently required (Fischer, 1999). This is why the problem of similarity search is closely related to all methods predicting protein structure and protein function on the basis of the known amino acid sequence (Fetrow et al., 2001; Irving et al., 2001; Kolinski et al., 2001). The relationships between stability and function have been examined by several researchers. Interaction with ligands or other molecules can switch proteins from one functional state to another by selective stabilization of some conformational states. The binding sites of some proteins are characterized by the presence of regions of high and low structural stability (Todd et al., 1998; Freire, 1999). The tools for multiple sequence alignment help to solve the problem of homology search, while with 3-D structure comparison is difficult to identify more than a pair of proteins (Sauder et al., 2000; Leibowitz et al., 2001). A wide range of methods for identification of domains in proteins were compared and presented in Marchler-Bauer et al. (2002). The alignments from sequence and structure comparisons are matched to distinguish fragments with presumably common characteristics, and thus to select motifs with similar biological functions. The structures and functions of many important proteins in biomedicine and drug discovery are derived by using the sequence alignment technique as a first step, what was shown in the comprehensive review paper (Chou, 2004).

*Corresponding author at: Department of Bioinformatics and Telemedicine, Collegium Medicum—Jagiellonian University, Kopernika 17, 31 501 Cracow, Poland. Tel./fax: +48 12 619 96 93.
E-mail address: myroterm@cyf-kr.edu.pl (I. Roterman).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.10.041

The distribution of rigidity and flexibility along the polypeptide chain in proteins appeared not to correlate with the localization of biological function, understood as ligand binding ability in proteins (Luque and Freire, 2000).

This paper presents a model for the search for conservative sequential and structural motifs the latter in intermediate structural forms: early-stage folding and partial unfolding, called "step-back" in this work. The model was introduced on the basis of geometric characteristics of the polypeptide backbone conformation (mutual relation of sequential peptide bond planes) (Roterman, 1995a, b) and information theory (balancing of the amount of information carried by an amino acid and the amount of information necessary to predict the conformation of that amino acid) (Jurkowski et al., 2004b). Structures created according to the ellipse path-limited conformational subspace on the Ramachandran map were previously tested as starting structural forms for energy minimization procedures. Positive results were found for ribonuclease (Jurkowski et al., 2004b), BPTI (Brylinski et al., 2004b), lysozyme (Jurkowski et al., 2004a) and human hemoglobin α and β chains (Brylinski et al., 2004c). A simple energy minimization procedure caused significant approach to the native structural forms of those proteins, although the accordance with native structure was found to be unsatisfactory. The purpose of this paper is to show whether sequence and structure consensus fragments in proteins belonging to serpines family can be recognized already in early-stage folding conformations (*in silico*). The limited conformational sub-space for early-stage folding structural forms appeared to be of ellipse shape. This is how the early-stage conformational sub-space will also be called in this paper.

The Ramachandran map represents the complete conformational space, which (according to the model) is available at late-stage folding after the structure is created based on the limited conformational sub-space. The structures representing conformations limited to the ellipse path can be reached (*in silico*) in two ways: (1) step-back (when the native structure of a particular protein is available), by transformation of the observed Φ, Ψ angles to the ellipse-belonging dihedral angles (obtained according to the shortest-distance criterion) called Φ_{sb}, Ψ_{sb} in this work and (2) early-stage (when only the amino acid sequence is known), on the basis of a sequence-to-structure contingency table representing the probability values of a particular sequence of tetrapeptide to adopt particular structure (the tetrapeptide was selected as the shortest polypeptide representing a well-defined structural form). This procedure was described in detail in Brylinski et al. (2004a, 2005). Since both structures (step-back and early-stage) representing intermediates in the protein folding simulation, are created on the basis of the ellipse-path-limited conformational sub-space (albeit with different starting points—Fig. 1), questions arise: how similar are they, and what is their relation to the native structural form of the protein? The problem of the possibility of consensus fragments recognition for both structure and sequence in early-stage folding (*in silico*). Since both sequence and structures can be represented using one-letter codes, the ClustalW program was used to extract fragments of sequences and structures (early-stage and step-back) of consensus character.

The clearest way to present the model is the diagram shown in Fig. 1 (the similar one—showing another protein as an example is given also in Brylinski et al., 2005). The step-back procedure (the crystal structure is known) begins with calculation of the Φ, Ψ angles, the values of which are changed to the nearest values of the dihedral angles belonging to the ellipse path (shortest-distance criterion, shown on the Ramachandran map). When the Φ_{sb}, and Ψ_{sb} angles are known, the structure representing partial unfolding can be created (shown in scheme). The transformation of all proteins present in January 2003 release of PDB reveals probability profiles for all amino acids. One of them (alanine) is shown in the scheme. The probability profile shall be interpreted as follows: a "*t*" parameter (parameter of ellipse equation) value equals to zero represents the point $\Phi = 90$ and $\Psi = -90^\circ$ on the Ramachandran map. An increase of "*t*" represents the clockwise movement along the ellipse. When all amino acids are put together, seven probability maxima can be distinguished. These seven probability maxima can be identified by letter codes as shown on the scheme. Identification of the ellipse fragment to which particular Φ, Ψ angles belong after their transformation to the ellipse allows the structure to be represented in the form of a symbol string as shown on the scheme. When all known proteins are classified according to the presented procedure, a contingency table expressing sequence-to-structure can be created. Such a contingency table was presented in Brylinski et al. (2005) and used to estimate the structure predictability of a particular amino acid sequence (Brylinski et al., 2004a).

The folding simulation path can be used when only the amino acid sequence is known. The amino acid sequence can be described in structural codes based on the sequence-to-structure contingency table. Because the tetrapeptide was selected as the unit, there are four different possibilities to assign structure to a particular tetrapeptide. The most frequent structural code is selected for the resulting structure string. When structural codes are assigned, the Φ_{es}, Ψ_{es} angles can be attached to each amino acid. The probability profile—the discrete one-differs from the one present in the step-back procedure. The Φ_{es}, Ψ_{es} angles (representing the positions of probability maxima in each ellipse fragment) assigned to each amino acid in a sequence produces the structure shown on the left side of the scheme. This structure is called early-stage folding (*in silico*). Late-stage folding (now in preparation), together with the energy minimization procedure applied to the early-stage structure, is assumed to approach the native structural form.

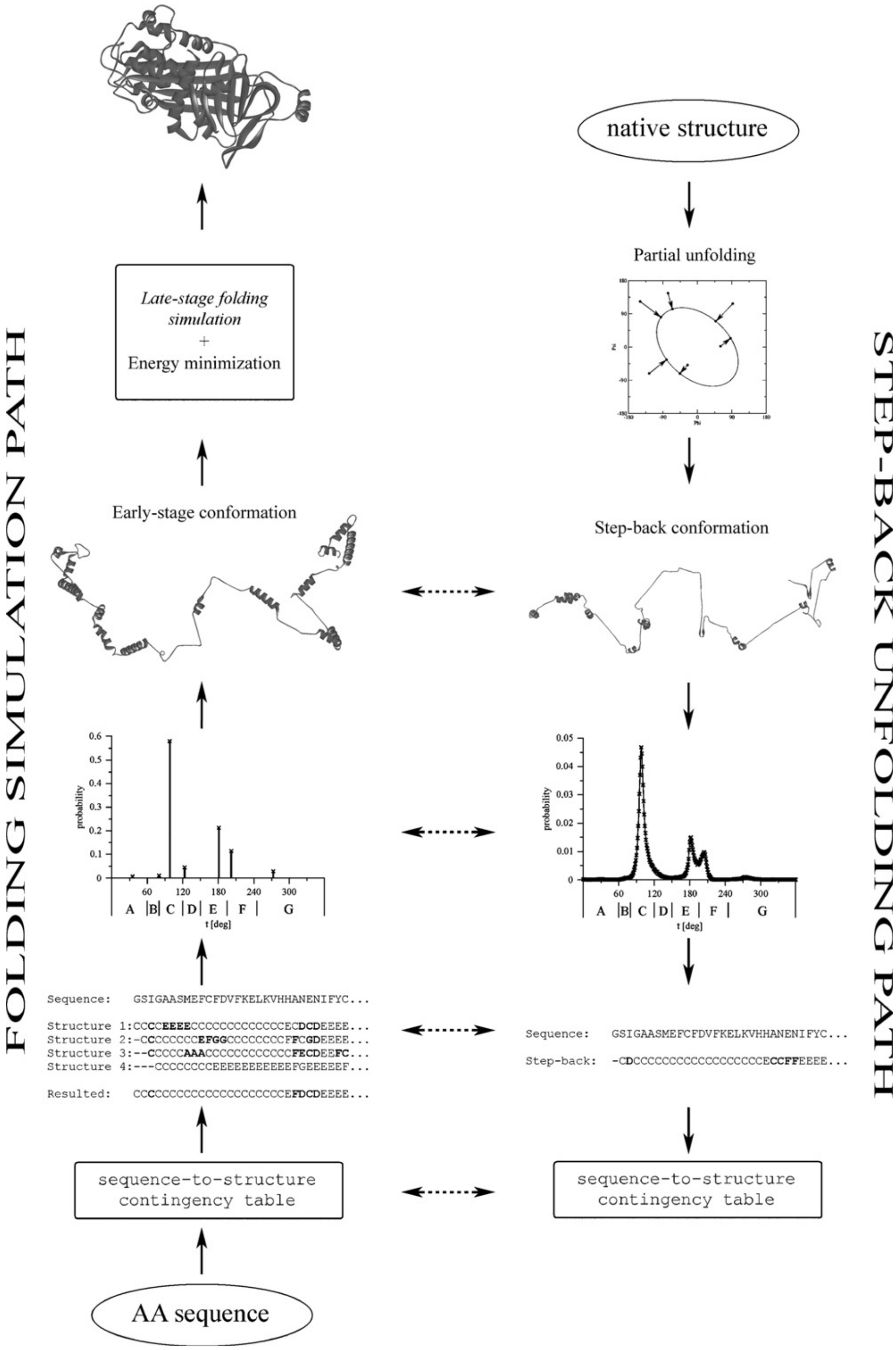

FOLDING SIMULATION PATH
STEP-BACK UNFOLDING PATH
native structure
Partial unfolding
180
90
0
-90
-180
Psi
Phi
Step-back conformation
Late-stage folding simulation
+
Energy minimization
Early-stage conformation
0.6
0.5
0.4
0.3
0.2
0.1
0
probability
0.05
0.04
0.03
0.02
0.01
60
120
180
240
300
A
B
C
D
E
F
G
t [deg]
Sequence: GSIGAASMEFCFDVFKELKVHHANENIFYC...
Structure 1:CCCCEEEECCCCCCCCCCCCCECDCDEEEE...
Structure 2:-CCCCCCCEFGGCCCCCCCCFFCGDEEEE...
Structure 3:--CCCCCAAACCCCCCCCCCCCFECDEEFC...
Structure 4:---CCCCCCCEEEEEEEEEEEFGEEEEEF...
Resulted: CCCCCCCCCCCCCCCCCCCCCEFDCDEEEE...
Sequence: GSIGAASMEFCFDVFKELKVHHANENIFYC...
Step-back: -CDCCCCCCCCCCCCCCCCCECCFFEEEE...
sequence-to-structure contingency table
sequence-to-structure contingency table
AA sequence

Fig. 1. Schematic presentation of the folding/unfolding path according to the limited conformational sub-space. Left: folding simulation path, right: step-back unfolding path. Dotted horizontal arrows denote equivalent stages of both paths. "*Late-stage*" folding procedure not applied (procedure in preparation). Details are given in text.

Step-back and early-stage structures differ in the form of the probability profile; the first is continuous and the second discrete. This paper is aimed at answering the question: to what extent are these two structures similar and to what extent are the structure-based consensus fragments accordant to sequence-based consensus fragments?

2. Materials and methods

2.1. Data collection

The serpine family of proteins was selected for analysis in this work. The BLAST program (Altschul, 1991; Altschul et al., 1997) was used to identify serpine family proteins from the complete PDB set (Berman et al., 2000). This search selected 43 proteins, analyzed in this work: 1A7C, 1AS4, 1ATH, 1ATT, 1ATU, 1AZX, 1B3 K, 1BR8, 1BY7, 1C5G, 1C8O, 1D5S, 1DB2, 1DVM, 1DVN, 1DZG, 1DZH, 1E03, 1E04, 1E05, 1EZX, 1F0C, 1HLE, 1HP7, 1IMV, 1JJO, 1JMJ, 1JMO, 1JRR, 1JTI, 1K9O, 1KCT, 1LJ5, 1OVA, 1PSI, 1QLP, 1QMB, 1QMN, 1SEK, 2ACH, 2ANT, 3CAA and 7API. Each amino acid in the proteins was given a one-letter code expressing the amino acid sequence, and a one-letter code representing the structure.

2.2. Structure assignment for step-back unfolding conformation

Structure assignment (one-letter code) was performed according to the following procedure. The Φ, Ψ angles of all amino acids were calculated for the native forms of all 43 proteins. The ellipse-belonging Φ, Ψ angles were found according to the shortest-distance criterion. Letter codes for particular ellipse fragments were assigned to each amino acid. Seven symbols, distinguishing seven structural motifs according to the scale presented in Fig. 1 are interpreted in Table 1.

2.3. Prediction of early-stage folding conformation

The structure codes were predicted for all amino acids in the proteins, based on the contingency table (Brylinski et al., 2005) and Structure Predictability Index (SPI) (Brylinski et al., 2004a) as shown in Fig. 1 (folding simulation path). Four structural letters can be assigned to each amino acid sequence of each protein. The procedure for finding the best structural letters fitted to a particular amino acid sequence was described in detail in (Brylinski et al., 2004a). The structural classification presented in Table 1 is common for both step-back and early-stage conformations.

The structural alphabet represents different fragments of ellipsoid path characterized by the local maximum of probability after moving all Phi, Psi angles toward the ellipse. Some of structural codes represent well known secondary structures: C–α_R, G–α_L, E and F—different forms of β-structure, although one shall mention, that the β-like structure, when created according to the ellipse path, is represented by the conformations close to C7eq forms.

Table 1
The interpretation of seven different structural classes distinguished on the basis of the limited conformational sub-space for proteins

Code	Interpretation
C	Right-handed α-helix
E, *F*	Two forms of β-strand
G	Left-handed helix
A, *B*, *D*	Different forms of random coil

2.4. Sequence and structure multiple alignment procedure

Sequences as well as structures (one-letter code system) were processed with the ClustalW1.83 program (Higgins and Sharp, 1988; Jeanmougin et al., 1998) (Gap Opening Penalty: 10.00, Gap Extension Penalty: 0.20, Delay divergent sequences: 30%, Protein weight matrix: Gonnet series), in the multialignment system. Only identity status was applied to the structure coding string.

2.5. Sequential and structural weight matrixes

The weight matrix is a quantitative structure motif descriptor assigning a score to any string of a particular length (Bucher, 1999). The weight matrix was adapted to amino acid sequences and both forms of structures (step-back and early-stage) expressed by letter code strings. The procedure of quantitative score value calculation was performed separately for sequence strings and strings representing both structural forms as follows. The set of codes in arbitrary order (Fig. 6a) was processed with the ClustalW program. The result of the multiple alignment procedure (Fig. 6b) was used to find the best alignment of structural codes and to calculate their frequencies (Fig. 6c). The highest frequency of particular structural code is given in bold and underlined on each position. The weight value (Fig. 6d) for each position was calculated according to the formula of log-odds ratios:

$$W = \log_{10}\left(\frac{F}{(N/M)+1}\right), \tag{1}$$

where F is the maximum frequency for a particular position and N is the total number of aligned strings (in our case $N = 43$). M denotes the number of codes. In our calculations, $M = 8$ for structural codes and $M = 21$ for amino acid codes (deletions/insertions included and coded as "-"). Highly conservative fragments were emphasized (thick black line) by averaging the raw data using a five-residue running window frame (Fig. 6c). The same positions in respect to the structural codes are also given in Fig. 6d to show, which fragments appeared to be conservative.

2.6. Consensus sequence and consensus structure

Since the W-value has no absolute scale, a reasonable cutoff value was set for highly conservative (sequence and structure) fragments. The cutoff varied from 10.0 to 18.0 with 0.2 step. A trial sequence was created with amino acids having W-values above a given cutoff. The sequence was then incorporated as an additional "artificial" member into the serpine family, and the multialignment procedure using ClustalW program was performed. The distances from a given sequence and each member of the serpine family were calculated utilizing the phylogenetic tree resulted from multialignment procedure. The mean value of distances (D) was used as the criterion for consensus sequence identification. Exactly the same procedure in the multiple alignment system was applied for sequence and for structure.

2.7. Step-back structure creation

The ellipse-belonging Φ_e and Ψ_e (found according to the shortest-distance criterion during structure assignment) angles become Φ_{sb}, Ψ_{sb} (step-back) angles, which allow creation of the structure called "step-back", representing partial unfolding. The Jackal program (Xiang and Honig, 2001; Jacobson et al., 2002) was used for full atom model creation.

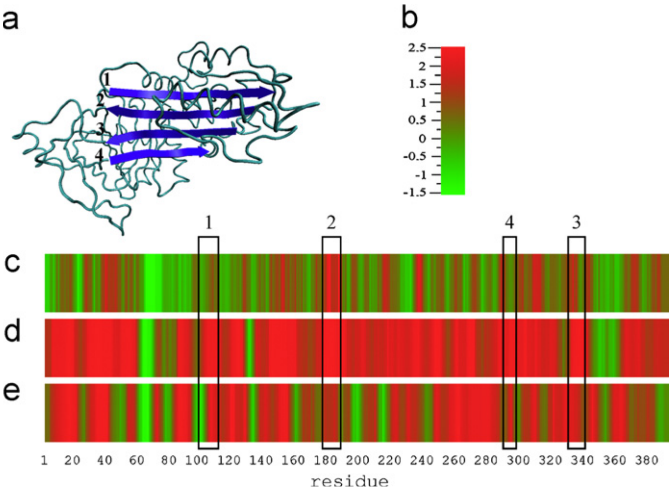


Fig. 2. The presentation of W-value profiles for uncleaved ovalbumin (1OVA): (a) native structure with A β-sheet (responsible for function) colored blue with individual strands numbered, (b) color scale of W-value and (c, d, e) profiles of W-value along the protein calculated for sequence, step-back and early-stage conformation multialignments, respectively. Black bars indicate locations of individual strands presented in (a).

2.8. Early-stage folding structure creation

On the basis of structure codes assigned to all the amino acids in the proteins during early-stage folding prediction (described above), Φ_{es}, Ψ_{es} angles were found according to discrete profiles based on seven probability maxima distinguished for the ellipse path (Fig. 7). Full atom models were created using the Jackal program (Xiang and Honig, 2001; Jacobson et al., 2002).

2.9. Visual analysis of full atom models

Four proteins representing the serpine family were selected for visual analysis: 2ACH—cleaved human alfa-1-antichymotrypsin (Baumann et al., 1991), 1OVA—uncleaved ovalbumin (Stein et al., 1991), 1AZX—human antithrombin (Jin et al., 1997) and 7API—alfa-1-antitrypsin (Engh et al., 1989). In those proteins, equivalent residues were found, corresponding to the consensus sequence and consensus structure identified by the multialignment procedure. Moreover, 10% and 30% of the total sequence and structure of selected proteins characterized by the highest W-value were identified.

3. Results

3.1. Sequence and structure weight matrix calculation

The W calculation was performed for the amino acid sequences and two structural strings (step-back unfolding

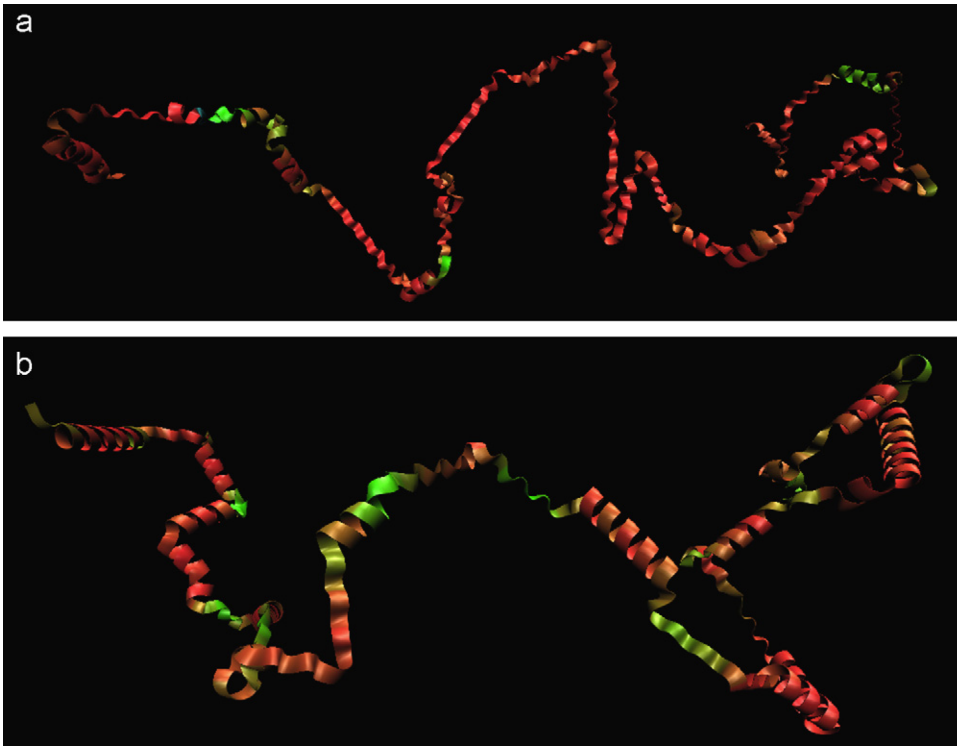


Fig. 3. Full atom models of uncleaved ovalbumin (1OVA): (a) step-back unfolding form and (b) early-stage folding (predicted) form. Both models are colored according to the common color scale expressing W-value (Fig. 2c).

and early-stage folding) of the serpine family (43 proteins). As an example, the profile of the W value for uncleaved ovalbumin (1OVA) is shown in Fig. 2. The high correlation was found between the two structural forms (step-back and early-stage). Particularly interesting are the fragments responsible for A β-sheet (Fig. 2a and b). High W values

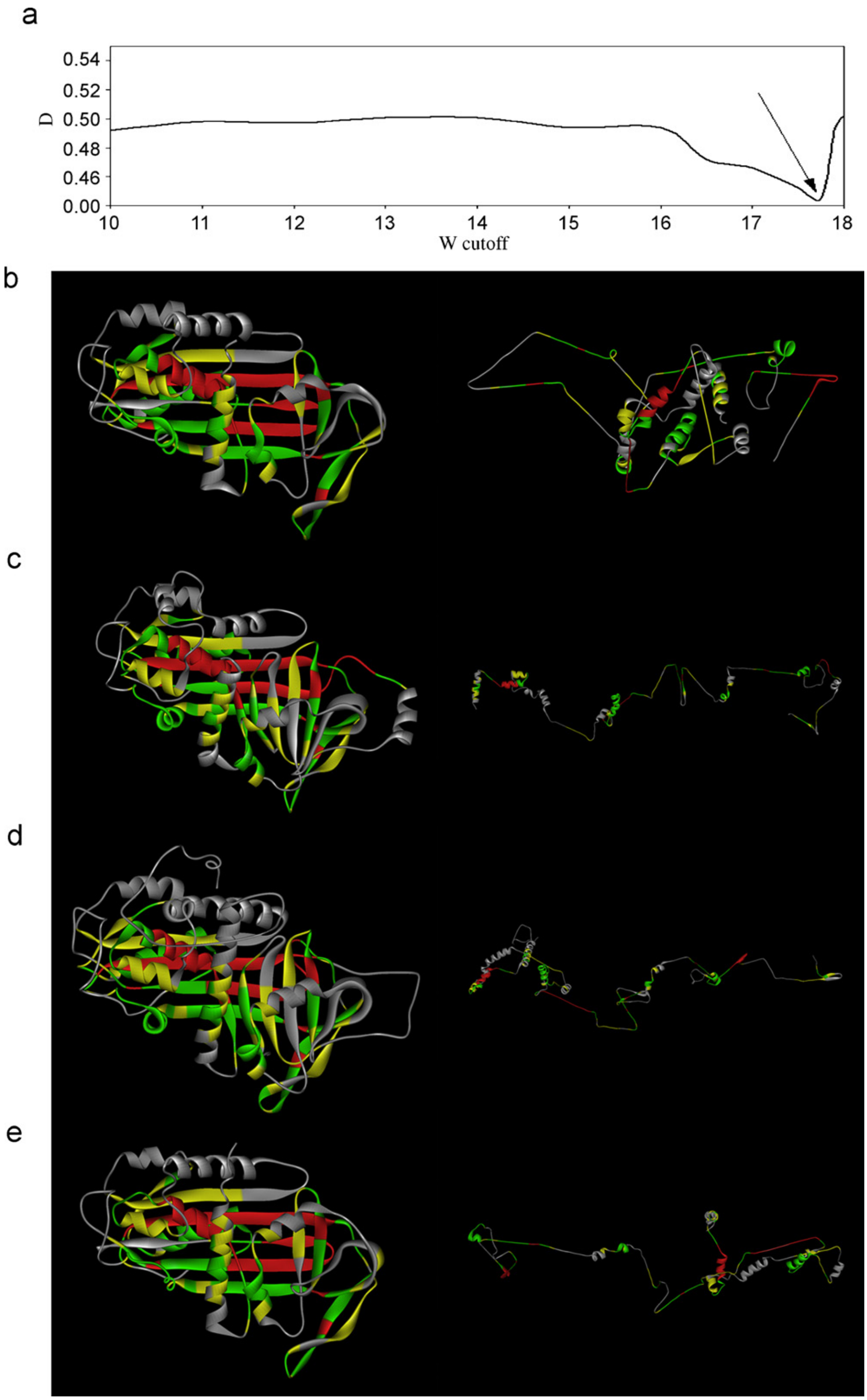


Fig. 4. Consensus sequence: (a) mean value of distances D resulting from the phylogenetic tree for a particular cutoff. Arrow denotes the best multialignment (highly conservative sequence) and (b, c, d, e) native (left column) and step-back (right column) structures of 2ACH, 1AZX, 1OVA and 7API, respectively. Consensus sequence (minimum point in (a)) is colored red. The 10% and 30% of the total sequence characterized by the highest W-value are colored green and yellow, respectively.

for these fragments point to the presence highly accordant similarity in sequence as well as in structure comparison (black bars in Fig. 2d–f). Similarly, the locations of residues representing high structural changeability (low *W* values) are in good agreement in both forms. A color scale was applied for particular *W*-values to show the

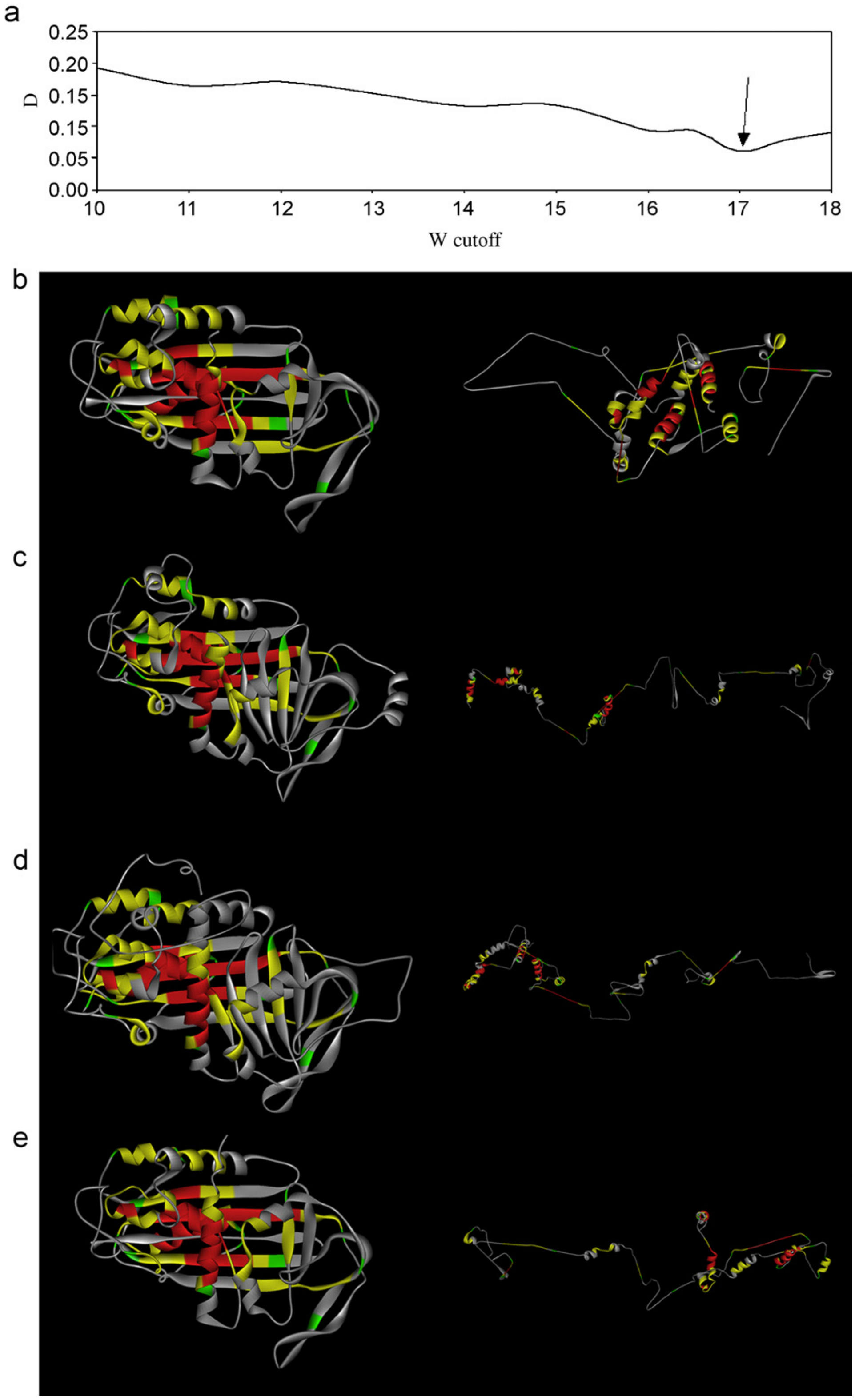


Fig. 5. Consensus structure: (a) mean value of distances *D* resulting from the phylogenetic tree for a particular cutoff. Arrow denotes the best multialignment (highly conservative structure) and (b, c, d, e) native (left column) and step-back (right column) structures of 2ACH, 1AZX, 1OVA, 7API, respectively. Consensus structure (minimum point in (a)) is colored red. The 10% and 30% of the total structure characterized by the highest *W*-value are colored green and yellow, respectively.

localization of fragments with particular degrees of similarity in the full atom models of step-back and early-stage forms (Fig. 3).

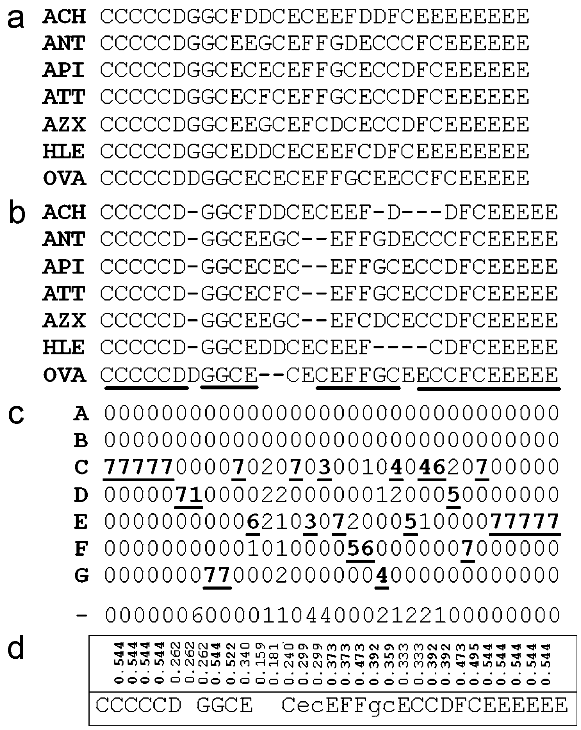
3.2. Consensus sequence and consensus structure

Application of $W_{sequence}$, $W_{structure}$ cutoffs allowed the consensus sequence and consensus structure to be distinguished. The results are presented in Fig. 4 for sequence and Fig. 5 for structure. In both cases a particular cutoff value was found to produce the best multiple alignment (Figs. 4a and 5a). Fragments with a W-value above the limit were treated as consensus fragments for both sequence and structure independently. The structures shown in Figs. 4b–e and 5b–e reveal fragments with particular degrees of similarity, on the basis of the sequence and structure similarity distribution (W value criterion). The images in the left column show these fragments in the final, native structure of the protein; those in the right column distribution of the same fragments in step-back (unfolded) structures of selected proteins. We stress that the similarity criterion (W-value calculated for structural letter codes) was set for early-stage folding (*in silico*) structures. The comparison to native structure was intended to trace the fragments found in early-stage folding (*in silico*) conformations. The aim of this procedure was to find out how far from the native structure is the early-stage folding (*in silico*) conformation of the polypeptide. It can be seen that the presence of high similarity and predisposition for β-structural fragments was already found in the early-stage conformation, which is difficult to recognize by visual analysis. The fragments that were highly conservative in sequence and structure were found exactly for the β-sheet (Figs. 4 and 5). The β-structure related area is situated far on the Ramachandran map compared to the ellipse-shaped limited conformational sub-space. The recognition of the similarity of β-structural fragments in early-stage structural forms is thus particularly important.

4. Discussion and conclusions

The "folding simulation" path applied for BPTI (Fig. 1) was presented earlier (Brylinski et al., 2004b), where the result seemed satisfactory. The structure of ribonuclease appeared less satisfactory after a simple energy minimization procedure was applied to early-stage folding structure (Jurkowski et al., 2004b), although the contact map for this protein appeared quite promising. Application of molecular dynamics simulation to the structures created based on the early-stage folding model of lysozyme did not improve the results, although some new correct inter-residual contacts appeared (Jurkowski et al., 2004a). The early-stage folding structures created according to the limited conformational sub-space introduced in (Roterman, 1995a, b; Jurkowski et al., 2004b) are treated as starting ones for further treatment assumed to represent a latter step of folding, mostly understood as hydrophobic collapse. A new model of latter step is in preparation.

The problem of the relation between sequence and structure in folding (*in silico*) intermediates is the object of this paper. Two structures were compared. One was created as partial unfolding of native structure to the corresponding structure belonging to the limited conformational sub-space. This form was obtained after changing the Φ, Ψ angles (native structure) to angles belonging to the limited conformational sub-space according to the shortest-distance criterion. The second one, called "early-stage" in this work, was predicted according to the highest probability found for each tetrapeptide in a contingency table expressing the sequence-to-structure relation. The comparison of these two structures versus the native structure was focused on the presence of sequence and structure conservation in protein family of serpines. Interesting are also some relations to biological function, which can also be discussed assuming that the biological function of serpines is (to some extent) understood as the ability to incorporate a polypeptide fragment into the



```
a ACH CCCCCDGGCFDDCECEEFDDFCEEEEEEEE
  ANT CCCCCDGGCEEGCEFFGDECCCFCEEEEEE
  API CCCCCDGGCECECEFFGCECCDFCEEEEEE
  ATT CCCCCDGGCECFCEFFGCECCDFCEEEEEE
  AZX CCCCCDGGCEEGCEFCDCECCDFCEEEEEE
  HLE CCCCCDGGCEDDCECEEFCDFCEEEEEEEE
  OVA CCCCCDDGGCECECEFFGCEECCFCEEEEE

b ACH CCCCCD-GGCFDDCECEEF-D---DFCEEEEE
  ANT CCCCCD-GGCEEGC--EFFGDECCCFCEEEEE
  API CCCCCD-GGCECEC--EFFGCECCDFCEEEEE
  ATT CCCCCD-GGCECFC--EFFGCECCDFCEEEEE
  AZX CCCCCD-GGCEEGC--EFCDCECCDFCEEEEE
  HLE CCCCCD-GGCEDDCECEEF----CDFCEEEEE
  OVA CCCCCDDGGCE--CECEFFGCEECCFCEEEEE

c   A 00000000000000000000000000000000
    B 00000000000000000000000000000000
    C 77777000070207030010404620700000
    D 00000710000220000001200050000000
    E 00000000006210307200051000077777
    F 00000000001010000560000007000000
    G 00000007700020000000400000000000

    - 00000060000110440002122100000000

d  0.544 0.544 0.544 0.544 0.262 0.262 0.262 0.544 0.522 0.340 0.159 0.181 0.240 0.299 0.299 0.373 0.373 0.473 0.392 0.359 0.333 0.333 0.392 0.392 0.473 0.495 0.544 0.544 0.544 0.544 0.544
   CCCCCD GGCE  CecEFFgcECCDFCEEEEEEE
```

Fig. 6. Example of the weight score selection procedure for nine proteins: (a) set of codes in arbitrary order, (b) result of multiple alignment (ClustalW) procedure applied to structural codes, (c) structural code frequency table ("-" denotes a gap), the highest frequency is given in bold and underlined and (d) W-values attributed to particular positions in a string (calculated according to Eq. (1) and averaged for three positions) with suggested consensus structural sequence (thick line in (b)) (the capital letters represent the high W score positions, the lower cases represent low frequency due to large number of gaps, where in all cases was the same structural code). The weight score values were calculated according to the Eq. (1) although the numbers were taken according to the example given on the picture (limited number of compared proteins). The results discussed in the paper were calculated for complete data set (43 proteins—as described in "Sequential and structural weight matrixes").

A β-sheet of the serpine molecule (Banzon and Kelly, 1992; Katz and Christianson, 1993; Carrell et al., 1994; Fletterick and McGrath, 1994; Wright and Scarsdale, 1995). The ClustalW program for multiple alignment to sequence and both versions of structure-coded strings was applied to extract the consensus sequence as well as consensus structure of the coded structural motifs in early-stage folding structure. The high accordance of results found for sequence and both forms of structure seems to validate the model of early-stage folding. Also, the one-letter code system for structure identification seems to work well as a tool for similarity recognition even in early-stage folding. The general conclusion is that model works well as a tool for sequence and structure consensus recognition in structures assumed to represent early stage folding (*in silico*) or partial unfolding created according to "early-stage" and "step-back" paths, respectively.

The similarity criterion (W-value calculated for structural letter codes) was set for early-stage folding (*in silico*) structures. The search for relation between the early-stage structural forms distinguished as highly conservative and their structural representation in the native structures revealed that the predisposition for β-structural fragments can be already found in the early-stage conformation. This observation is of particular interest despite the absence of *explicit* β-structural fragments in early-stage limited conformational sub-space. It can be concluded, that the β-structural fragments are already seeded in structural forms of the early-stage folding model.

The analysis presented in this paper is the continuation of the analysis oriented on similarity search in proteins (Leluk et al., 2003), where the serpines family was the object of structural similarity analysis. The geometric parameters characterizing the early-stage structural forms taken into consideration seem to correspond well with the analysis shown in this paper. The structural alphabet remains in the close relation to the geometric parameters, which appeared to be the basis for early-stage definition in our model (Roterman, 1995b). However, the geometric parameters were calculated for native structure of proteins belonging to the serpines family, the results seem to be accordant.

The highly ordered structural fragments related to biological function were in the focus of the analysis. The multiple alignment applied to structural alphabet seems to work well, although the further analysis oriented on loops

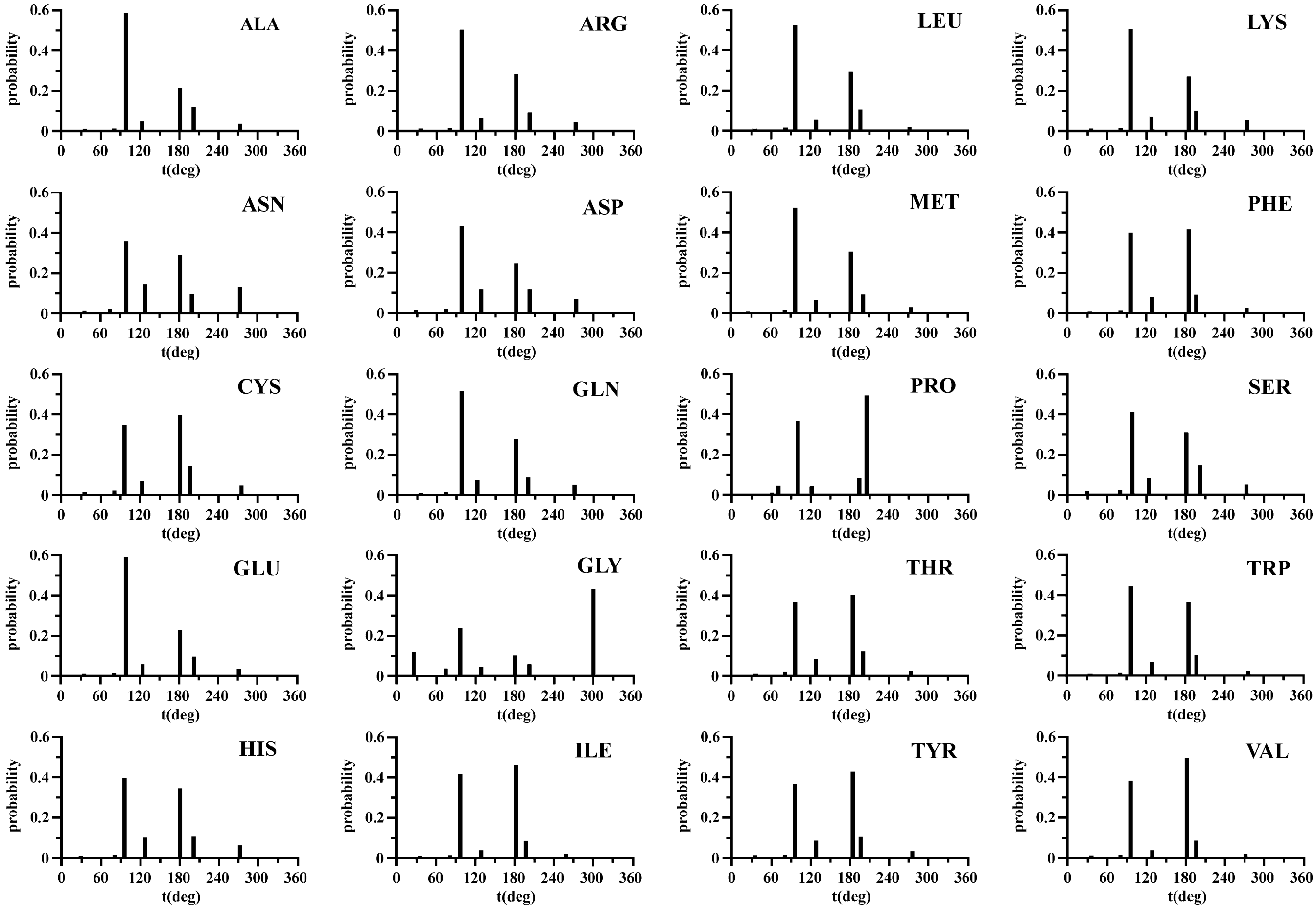


Fig. 7. Discrete profiles of probability in ellipse-shaped limited conformational sub-space for all amino acids.

is necessary. The model applied does not take into account the handedness of the structures. It is well known that the secondary structure motifs in proteins bear remarkable handedness. For instance, all the real beta-sheets in proteins are right-handed twist (Chou and Scheraga, 1982; Chou et al., 1983, 1982), all the β–α–β structures are right-handed cross over (Chou et al., 1989), all the helix-bundles are right-handed twist (Chou et al., 1988), all the β-barrels are right-handed twist (Chou and Carlacci, 1991; Chou et al., 1990), and so forth. The massive analysis of polypeptide fragments linking the fragments of highly ordered secondary regions is the object of the analysis running currently. The high specificity of linkers in β–α–β, helix-bundles and β-barrels is expected in respect to structural codes appearance. The three possibilities are taken into account: 1—the handedness is the natural consequence of particular Φ, Ψ angles of amino acids participating in the loop generation or 2—the specific combination of the structures recognized as *A*, *B* and *D* (according to the structural alphabet notion) may influence the handedness of the whole fragment or 3—the presence of structural form *C* (right-handed helix) for isolated residues in highly unordered fragments may also influence the handedness of the whole "linker". This observation is shown in Fig. 6, where the individual *C* structural form appears accompanied by *G* structural form (according to structural alphabet—represents the left-handed helix). This combination may have significant influence on the handedness of the whole fragment of the linker. This issue (particularly the handedness of "linkers") will be the focus of the prospective analysis oriented on loops' structures analysis in respect to the introduced structural alphabet and comparison to observations available in the literature (Fig. 7).

Acknowledgment

The project was founded in part by grants from Collegium Medicum—Jagiellonian University (P/133/L and WŁ/222/P/L).

References

Altschul, S.F., 1991. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219 (3), 555–565.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Banzon, J.A., Kelly, J.W., 1992. Beta-sheet rearrangements: serpins and beyond. Protein Eng. 5 (2), 113–115.

Baumann, U., Huber, R., Bode, W., Grosse, D., Lesjak, M., et al., 1991. Crystal structure of cleaved human alpha 1-antichymotrypsin at 2.7 Å resolution and its comparison with other serpins. J. Mol. Biol. 218 (3), 595–606.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., et al., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235–242.

Brylinski, M., Konieczny, L., Roterman, I., 2004a. SPI—structure predictability index for protein sequences. In Silico Biol. 5, 0022.

Brylinski, M., Jurkowski, W., Konieczny, L., Roterman, I., 2004b. Limited conformational space for early-stage protein folding simulation. Bioinformatics 20 (2), 199–205.

Brylinski, M., Jurkowski, W., Konieczny, L., Roterman, I., 2004c. Limitation of conformational space for proteins—early stage folding simulation of human α and β hemoglobin chains. TASK Q. 8 (3), 413–422.

Brylinski, M., Konieczny, L., Czerwonko, P., Jurkowski, W., Roterman, I., 2005. Early-stage folding in proteins (In Silico) sequence-to-structure relation. J. Biomed. Biotechnol. 2 (2), 65–79.

Bucher, P., 1999. Gene feature identification. In: Bishop, M.J. (Ed.), Genetics Databases. Academic Press, San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo, pp. 135–164.

Carrell, R.W., Whisstock, J., Lomas, D.A., 1994. Conformational changes in serpins and the mechanism of alpha 1-antitrypsin deficiency. Am. J. Respir. Crit. Care Med. 150 (6 Pt 2), S171–S175.

Chou, K.C., 2004. Review: structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry 11, 2105–2134.

Chou, K.C., Carlacci, L., 1991. Energetic approach to the folding of alpha/beta barrels. Proteins Struct. Funct. Genet. 9, 280–295.

Chou, K.C., Scheraga, H.A., 1982. Origin of the right-handed twist of beta-sheets of poly-L-valine chains. Proc. Natl Acad. Sci. USA 79, 7047–7051.

Chou, K.C., Pottle, M., Nemethy, G., Ueda, Y., Scheraga, H.A., 1982. Structure of beta-sheets: origin of the right-handed twist and of the increased stability of antiparallel over parallel sheets. J. Mol. Biol. 162, 89–112.

Chou, K.C., Nemethy, G., Scheraga, H.A., 1983. Role of interchain interactions in the stabilization of right-handed twist of β-sheets. J. Mol. Biol. 168, 389–407.

Chou, K.C., Maggiora, G.M., Nemethy, G., Scheraga, H.A., 1988. Energetic approach to 4-alpha-helix packing. Proc. Natl Acad. Sci. USA 85, 4295–4299.

Chou, K.C., Nemethy, G., Pottle, M., Scheraga, H.A., 1989. Energy of stabilization of the right-handed beta-alpha-beta crossover in proteins. J. Mol. Biol. 205, 241–249.

Chou, K.C., Nemethy, G., Scheraga, H.A., 1990. Review: energetics of interactions of regular structural elements in proteins. Acc. Chem. Res. 23, 134–141.

Engh, R., Lobermann, H., Schneider, M., Wiegand, G., Huber, R., et al., 1989. The S variant of human alpha 1-antitrypsin, structure and implications for function and metabolism. Protein Eng. 2 (6), 407–415.

Fetrow, J.S., Siew, N., Di Gennaro, J.A., Martinez-Yamout, M., Dyson, H.J., et al., 2001. Genomic-scale comparison of sequence- and structure-based methods of function prediction: does structure provide additional insight? Protein Sci. 10 (5), 1005–1014.

Fischer, D., 1999. Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge. Protein Eng. 12 (12), 1029–1030.

Fletterick, R.J., McGrath, M.E., 1994. Deconvoluting serpins. Nat. Struct. Biol. 1 (4), 201–203.

Freire, E., 1999. The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. Proc. Natl Acad. Sci. USA 96 (18), 10118–10122.

Higgins, D.G., Sharp, P.M., 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73 (1), 237–244.

Irving, J.A., Whisstock, J.C., Lesk, A.M., 2001. Protein structural alignments and functional genomics. Proteins 42 (3), 378–382.

Jacobson, M.P., Friesner, R.A., Xiang, Z., Honig, B., 2002. On the role of the crystal environment in determining protein side-chain conformations. J. Mol. Biol. 320 (3), 597–608.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment with Clustal X. Trends Biochem. Sci. 23 (10), 403–405.

Jin, L., Abrahams, J.P., Skinner, R., Petitou, M., Pike, R.N., et al., 1997. The anticoagulant activation of antithrombin by heparin. Proc. Natl Acad. Sci. USA 94 (26), 14683–14688.

Jurkowski, W., Brylinski, M., Konieczny, L., Roterman, I., 2004a. Lysozyme folded in silico according to the limited conformational sub-space. J. Biomol. Struct. Dyn. 22 (2), 149–158.

Jurkowski, W., Brylinski, M., Konieczny, L., Wiiniowski, Z., Roterman, I., 2004b. Conformational subspace in simulation of early-stage protein folding. Proteins 55 (1), 115–127.

Katz, D.S., Christianson, D.W., 1993. Modeling the uncleaved serpin antichymotrypsin and its chymotrypsin complex. Protein Eng. 6 (7), 701–709.

Kolinski, A., Betancourt, M.R., Kihara, D., Rotkiewicz, P., Skolnick, J., 2001. Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 44 (2), 133–149.

Leibowitz, N., Fligelman, Z.Y., Nussinov, R., Wolfson, H.J., 2001. Automated multiple structure alignment and detection of a common substructural motif. Proteins 43 (3), 235–245.

Leluk, J., Konieczny, L., Roterman, I., 2003. Search for structural similarity in proteins. Bioinformatics 19 (1), 117–124.

Luque, I., Freire, E., 2000. Structural stability of binding sites: consequences for binding affinity and allosteric effects. Proteins Suppl. 4, 63–71.

Marchler-Bauer, A., Panchenko, A.R., Ariel, N., Bryant, S.H., 2002. Comparison of sequence and structure alignments for protein domains. Proteins 48 (3), 439–446.

Roterman, I., 1995a. The geometrical analysis of peptide backbone structure and its local deformations. Biochimie 77 (3), 204–216.

Roterman, I., 1995b. Modelling the optimal simulation path in the peptide chain folding—studies based on geometry of alanine heptapeptide. J. Theor. Biol. 177 (3), 283–288.

Sauder, J.M., Arthur, J.W., Dunbrack Jr., R.L., 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 40 (1), 6–22.

Stein, P.E., Leslie, A.G., Finch, J.T., Carrell, R.W., 1991. Crystal structure of uncleaved ovalbumin at 1.95 Å resolution. J. Mol. Biol. 221 (3), 941–959.

Todd, M.J., Semo, N., Freire, E., 1998. The structural stability of the HIV-1 protease. J. Mol. Biol. 283 (2), 475–488.

Wright, H.T., Scarsdale, J.N., 1995. Structural basis for serpin inhibitor activity. Proteins 22 (3), 210–225.

Xiang, Z., Honig, B., 2001. Extending the accuracy limits of prediction for side-chain conformations. J. Mol. Biol. 311 (2), 421–430.