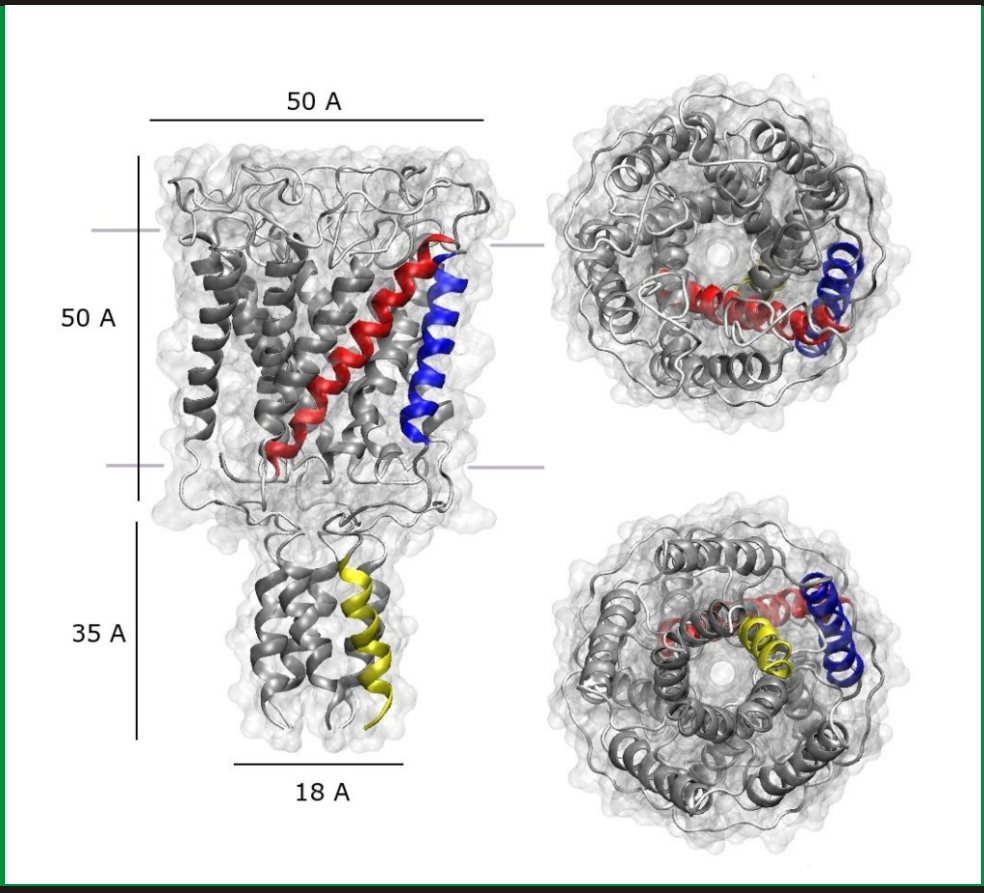
RECENT ADVANCES IN STRUCTURAL BIOINFORMATICS

2007

Editor
Alexandre G. de Brevern



RESEARCH SIGNPOST

RECENT ADVANCES IN STRUCTURAL BIOINFORMATICS 2007 Editor: Alexandre G. de Brevern

Research Signpost
37/661 (2), Fort P.O., Trivandrum-695 023, Kerala, India

Recent Adv. In Structural Bioinformatics, 2007: ISBN: 978-81-308-0208-4
Editor: Alexandre G. de Brevern

5 Early-stage protein folding – *In silico* model

Roterman I[1], Brylinski M[1,2], Konieczny L[3] and Jurkowski W[1,2]
[1]Department of Bioinformatics and Telemedicine, Collegium Medicum – Jagiellonian University, Krakow, Kopernika 17, Poland; [2]Faculty of Chemistry Jagiellonian University, Krakow, Ingardena 3, Poland; [3]Institute of Medical Biochemistry, Collegium Medicum - Jagiellonian University, Krakow Kopernika 7, Poland

Abstract

The protein folding is treated as a multi-step process. An early-stage folding model (in silico) based on the backbone conformation is presented. The limited conformational sub-space for this stage of the process is also shown. The presented model has promising applications to some protein structure-related problems such as the structural similarity search, structure classification, sequence-to-structure and structure-to-sequence relations and protein structure prediction.

Correspondence/Reprint request: Dr. Roterman I, Department of Bioinformatics and Telemedicine, Collegium Medicum – Jagiellonian University, Krakow, Kopernika 17, Poland. E-mail: myroterm@cyf-kr.edu.pl

Introduction

Protein structure, determined by amino acid sequence and specific for the biological function of particular protein remains a secrete of the nature. The tools for correct prediction of protein structure have not yet been constructed, despite of the thirty-year long history of this discipline [1,2].

So far, the protein folding has been recognized as the multi-step process with several intermediates between early-stage and native structural forms; what has been verified experimentally [3-17]. The notion of molten globule state has been introduced to describe the result of partial unfolding of the native structural form of proteins [18-20].

The existence of intermediate steps has been discussed, although the theoretical approaches to protein structure prediction have not yielded models for early-stage folding. Such a model will be discussed in this chapter, albeit only for simulation of early-stage folding *in silico*.

Geometrical background

The commonly accepted opinion is that the early-stage folding of a polypeptide depends mostly on the backbone conformation [21-23]. The backbone can be defined as the chain of peptide bond planes whose mutual orientations are expressed by the Phi and Psi angles determining the particular structure. The complete set of possible Phi (from -180 to 180 deg) and Psi (from -180 to 180 deg) angles creates the conformational space. Knowledge of the complete set of Phi, Psi angles for particular polypeptide chain as it appears in particular protein unequivocally determines the unique three-dimensional structure of protein under consideration. A disadvantage of the use of these angles for structure to identify structure is that they are difficult to visualize without graphic programs. It is difficult to interpret the values of Phi and Psi angles, except for a few regular structures such as helixes, some β-structural forms and turns.

Alternative parameters to describe the backbone structure may be introduced [24-26]. These parameters are easy to visualize, but they cannot identify structure unequivocally. These geometrical parameters describing the pentapeptide structure are:

1. the V-angle – dihedral angle between two sequential peptide bond planes (represented by C=O bonds)
2. the R-radius of curvature describing the form of curvature for pentapeptide.

The following assumptions are adopted before the presented parameters can be defined:

1. all structures observed in proteins are of helical form. The difference between α-helix and β-structure is, that the α-helix represents a well

defined radius of curvature, while β- (or extended) structure is represented by a radius of curvature of infinitely large size. All other structures are in-between forms. A consequence of this assumption is that all structures can fluently change their forms in a continuous way without the discrete categorizations. The value of V-angle is a simple consequence of particular Phi and Psi angle, although the reverse is not possible: knowledge of V-angle is not sufficient to define the Phi, Psi angles.

2. The values of the parameters (particularly R-radius) can be calculated for polypeptide chains uniformly oriented in space. It is the orientation of the Z-axis, determined by the averaged positions of C and O atoms of C=O groups in pentapeptide. The next step is to localize the position of Cα atom projections on the XY plane. Localization of these points makes it possible to estimate the R-radius value.
3. The next assumption is, that the V-angle representing pentapeptide is the angle between second and third peptide bond planes (represented by C=O groups for example). The averaged value of second and fourth peptide bond plane orientations versus the third can also be used, although orientation of second and the fourth C=O may differ significantly in real proteins).

The pentapeptide has been selected as a unit of well defined structural form (helix, β-turn, β-structure etc).

The complete Ramachandran map (conformational space) has been represented by a grid point system (1, 5 and 10 degs grid step) to verify the reliability of geometrical representation of polypeptide chain structure. The pentapeptide structure representing a particular grid point has been created.

The V-angle and R-radius is calculated for each grid point structure. To avoid very high values, the logarithmic scale is applied to express R value.

The distributions of V and lnR values covering the Ramachandran map (the conformational space) are shown in Fig.1. A and Fig. 1.B respectively. The symmetry of these maps reveals that these two parameters do not distinguish chirality.

The relation between V and lnR and their distribution over all Ramachandran map is shown in Fig.2. The structures localized in upper left part of the map are described by high V and high lnR values. The area near helical region is described by low V and low lnR.

Only low-energy part of conformational space (Ramachandran map) is acceptable for polypeptide chains (Fig.3.A). The relation between V and lnR for structures limited to low-energy parts of Ramachandran map suggests a parabolic function as shown in Fig.3.B and eq.1.

$$\ln(R) = A * V^2 - B * V + C \qquad \text{[eq. 1]}$$

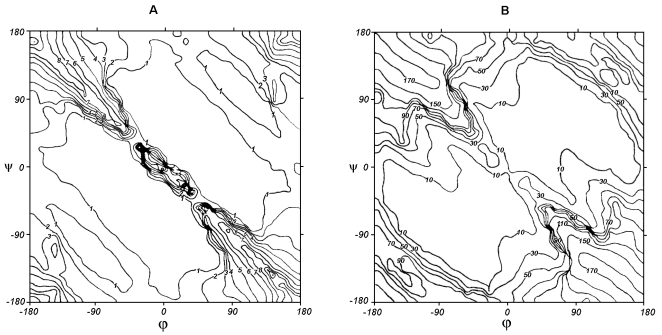


Figure 1. Distribution of geometrical parameters all over the Ramachandran map: A – radius of curvature (in log scale), B – V-angle between two sequential peptide bond planes.

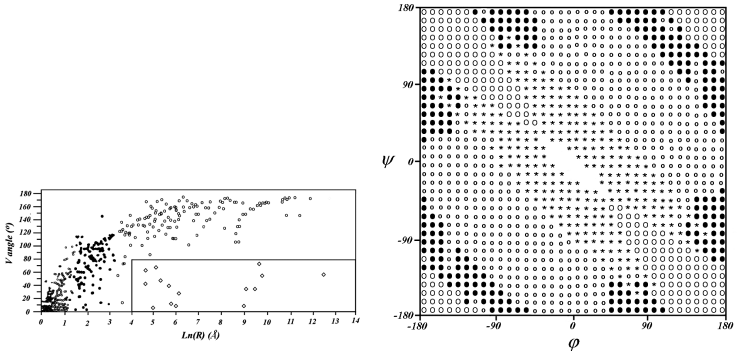


Figure 2. The characteristics of parameters: ln(R) and V-angle: A – the relation between ln(R) and V-angle B-the relation-dependent distribution of ln(R) versus V-angle. The symbols differentiate the characteristics of particular area of conformational space represented by Ramachandran map.

The opposite question can be asked: Which structures do obey exactly the relation expressed by eq 1. and Where are these structures localized on the Ramachandran map?

The answers are given in Fig.3.C. The points creating the ellipse-path on Ramachandran map are especially interesting (Fig.3.D) (eq.2). This path links all structurally important areas on Ramachandran map. It links right-handed helix with C_{7eq} energy minimum (for dipeptides) and then with left-handed

helix (Fig.3.E). This suggests the possible path for fluent structural changes leading one ordered structural form to another.
Ellipse parametric equation is as follows:

$$\Phi = -A\cos(t) - B\sin(t)$$
$$\Psi = A\cos(t) - B\sin(t) \qquad \text{[eq. 2]}$$

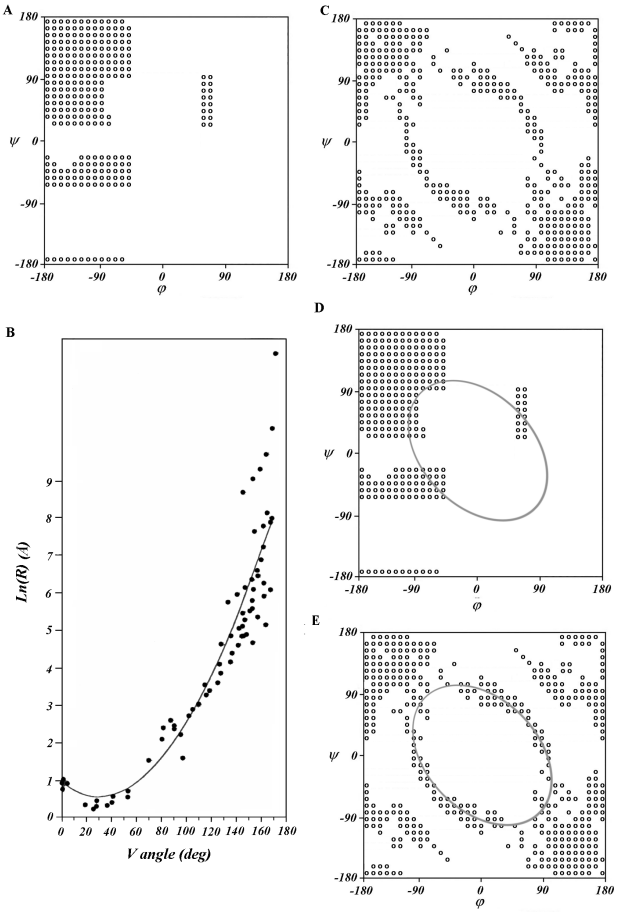


Figure 3. Ellipse-path determination: A – low energy area in conformational space, B – ln(R) as function of V-angle for grid points shown in A. C – grid points satisfying the relation expressed by eq.1. D – ellipse-path as approximation to the grid points presented in C, E – low-energy area linked by ellipse path.

Where t-angular rise of clock-wise movement along the ellipse. A and B calculated according to approximation procedure.

Empirical support for this path can be found in any Phi, Psi angles distribution showing higher concentration of dots representing Phi, Psi angles in real proteins along the proposed ellipse path.

More support based on the computational examples can be found in the works of Levitt and Daggett [27,28], who simulated helix unfolding. They considered two forms of helixes: one free helix of 13 alanines and the other being part of BPTI protein. The ellipse path is covered by the movement of the dots representing changed Phi, Psi angles on the Ramachandran map, particularly in the case of BPTI unfolding. The differences between the observed paths can be explained in the following way. The unfolding of the isolated helix allowed its end to move freely performing large radial displacements. The free movement of terminal fragments of helix in protein (when both its ends are connected to other parts of protein) only the way through the squeeze is possible. The gradual squeezing of the helix leads to the 2_7 structure and then transforms into an elongated structure close in form to β-like structures.

The structural changes accompanied the two observed paths are shown in Fig.4.A and Fig.4.B.

The information theory-based approach

The aim for all methods oriented toward prediction of protein structure is to match a particular (native) structure to a given amino acid sequence. Is this possible ?

Let us assume that probability of selecting one amino acid from among twenty different amino acids is equal to 1/20. According to Shannon's definition [29], the amount of information necessary to select a particular amino acid – or the amount of information carried by one amino acid – is equal to:

$$I = -\log_2 p \qquad \text{[eq. 3]}$$

$$I = -\log_2 \frac{1}{20} = 4.322(bit) \qquad \text{[eq. 4]}$$

Since the frequency of amino acids in proteins varies, the *I* value should be calculated taking *p* equal to actual frequency of each amino acid, what has been calculated for the complete PDB data base (January 2003 release) (Tab.1.) [30].

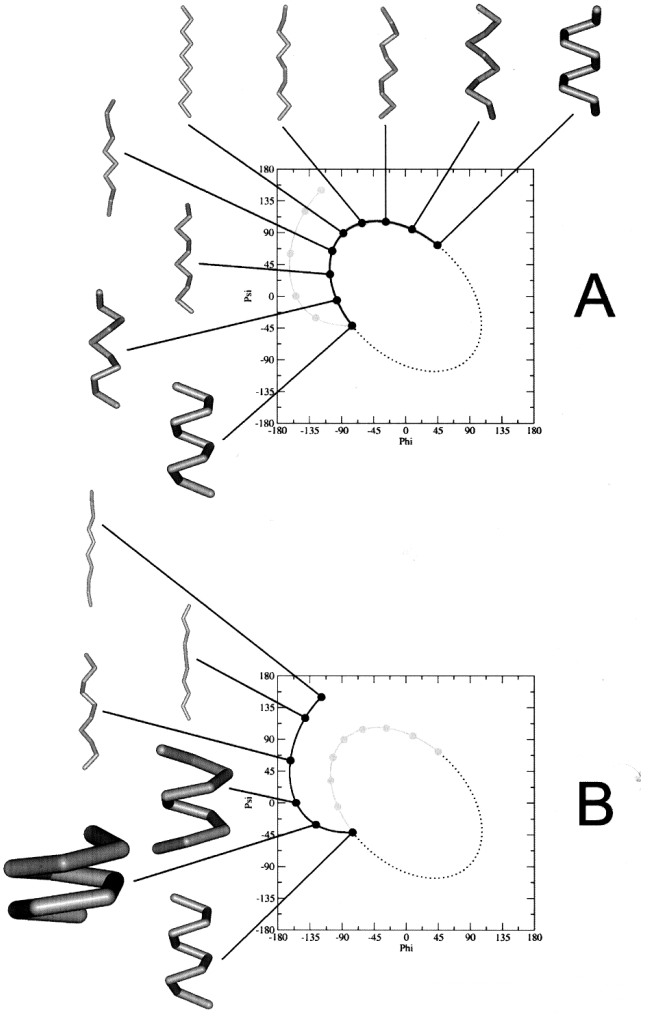


Figure 4. The structure evolution along two paths: A – ellipse-path according to model, B – alternate path suggested on the basis of Daggett and Levitt [27,28]

Continuing the discussion based on the definition of the amount of information, one can assume that sequence prediction needs of information equal to:

$$I = -\log_2 \prod_{i=1}^{N} p_i \qquad \text{[eq. 5]}$$

where: N is a number of amino acids in a sequence and p_i expresses the frequency of particular amino acid in sequence. The eq. 5 is correct when the sequence is treated as independent selection of amino acids (no conditional probability is taken into account).

On the other hand, let us assume that one needs to find one grid point (assume a grid step equal to 1, 5 and 10 deg) representing a particular structure. The amount of information necessary in these cases is equal to:

$$I_1 = -\log_2 \frac{1}{360 * 360} = 16.98(bit) \qquad \text{[eq. 6]}$$

$$I_5 = -\log_2 \frac{1}{72 * 72} = 12.34(bit) \qquad \text{[eq. 7]}$$

$$I_{10} = -\log_2 \frac{1}{36 * 36} = 10.34(bit) \qquad \text{[eq. 8]}$$

These equations are correct on the assumption, that selection of each grid point is equally possible.

Particular grid points are represented by different probability (invert proportion to energy level). The informational entropy (treated as averaged amount of information needed for particular amino acid) shall be calculated as follows:

$$SE = -\sum_{i=1}^{360/n} \sum_{j=1}^{360/n} p_{ij} \log_2 p_{ij} \qquad \text{[eq. 9]}$$

To calculate the entropy according to eq. 9 the distribution of energy all over the whole Ramachandran map with grid step as defined above for each amino acid or distribution of Phi, Psi angles for particular amino acid shall be known (p_{ij}). These calculations (both versions) have been done using ECEPP/3 program and Phi, Psi angle distribution (Fig.5 A and Fig.5.B).

The informational entropy values calculated for each amino acid are shown in Tab. 1.

The analysis of the relation between the amount of information carried by amino acid is significantly too low to supply the expected amount of

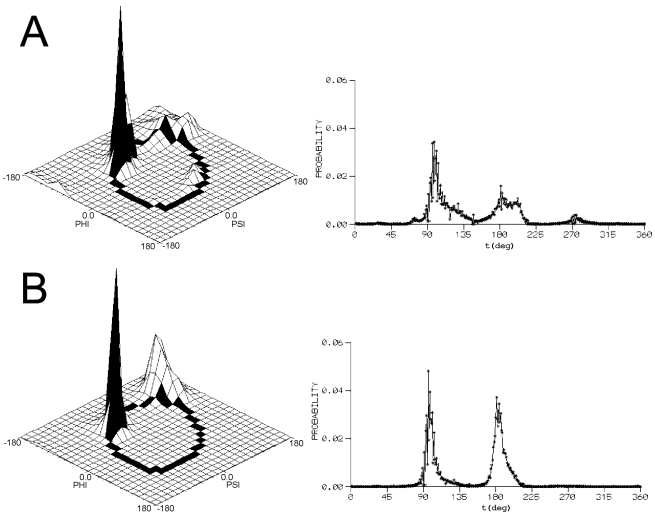


Figure 5. Three-dimensional probability distribution calculated for 5 degs grid step. The black fields distinguish the ellipse path to show the amino acid-dependent relation of Phi, and Psi angle distribution versus the ellipse path for ASN (A) and ILE (B) together with the probability profile as it appears after moving all Phi, Psi angles to the nearest point on the ellipse path.

information needed for structure prediction (grid size dependent). It is obvious, that these two amounts are not balanced.

The conclusion from this analysis is, that the conformational space needs to be limited, at least for the initial steps of protein folding.

When ellipse path is taken as the limited conformational sub-space, the amount of information necessary to predict particular fragment of the ellipse (step size) is balanced in respect to the amount of information carried by amino acids (Tab.1. and Fig.6.).

The opinion that a limited conformational sub-space needs to be introduced has also been expressed in [31,32]. Decreasing degrees of freedom by simplifying the representation of the amino acids (effective atoms, Cα-Cα virtual bonds) did not solve the problem of protein structure prediction [33,34]. The method of force field deformation (decreasing number of local minima) did not solved the problem either [35-40]. Limitation of the conformational space offers an alternative to simplifying the presentation of structure lowering the number of degrees of freedom in energy minimization procedure.

The ellipse-path appeared to satisfy two important conditions: it presents a simplified geometrical description of backbone structure, and it balances the

Table 1. The amount of information (bit) carried by particular amino acid (second column) and amount of information (bit) necessary to recognize particular grid point (10 degs step) (third column) followed by amount of information (bit) (fourth column) necessary to recognize particular fragment of ellipse. The two last columns express the deficiency (third column) versus the ten deg grid for Ramachandran map and (fourth column) the deficiency/excess of information to determine ellipse-limited structure (calculated versus amount of information carried by particular amino acid).

AA	AMOUNT OF INFORMATION CARRIED BY AA [bit]	AMOUNT OF INFORMATION NECESSARY TO PREDICT STRUCTURE 10 deg grid [bit]	AMOUNT OF INFORMATION NECESSARY TO PREDICT STRUCTURE Ellipse-limited [bit]	DEFICIT OF INFORMATION TO PREDICT STRUCTURE 10 deg grid [bit]	DEFICIT/EXCESS OF INFORMATION TO PREDICT STRUCTURE Ellipse-limited [bit]
G	3.8	5.74	8.87	-5.07	-1.94
L	3.5	4.44	7.04	-3.54	-0.94
D	4.12	5.02	7.81	-3.69	-0.9
K	3.91	4.76	7.62	-3.71	-0.85
A	3.66	4.46	7	-3.34	-0.8
S	4.1	4.86	7.81	-3.71	-0.76
E	3.83	4.55	7.22	-3.39	-0.72
N	4.55	5.19	8.11	-3.56	-0.64
R	4.25	4.65	7.47	-3.22	-0.4
T	4.2	4.58	7.66	-3.46	-0.38
V	3.89	4.1	7.06	-3.17	-0.21
Q	4.66	4.67	7.36	-2.7	-0.01
I	4.15	4.11	6.91	-2.76	0.04
F	4.71	4.53	7.52	-2.81	0.18
Y	4.94	4.57	7.53	-2.59	0.37
P	4.44	4.06	6.51	-2.07	0.38
H	5.48	4.87	7.92	-2.44	0.61
C	5.54	4.79	7.94	-2.4	0.75
M	5.61	4.48	7.12	-1.51	1.13
W	6.23	4.51	7.38	-1.15	1.72

amount of information. This coincidental agreement has been taken as the basis for introducing the conformational sub-space for early-stage folding.

Step-back structures - Partial unfolding

The reliability of the model was verified with tests using four different protein molecules: lysozyme (PDB –2EQL), BPTI (PDB – 4PTI), ribonuclease (PDB –5RAT) and α and β hemoglobin chains (PDB –3HHB) [30, 41-43]. The crystal structures of these proteins were transformed to the ellipse path-based structures called early-stage structural forms. The procedure applied was as follows:

1. the Phi, Psi angles have been calculated for these proteins
2. the values of Phi, Psi angles has been changed to their representations on the ellipse according to the criterion of shortest distance (Phi_e, Psi_e)
3. the structures were created according to Phi_e, Psi_e

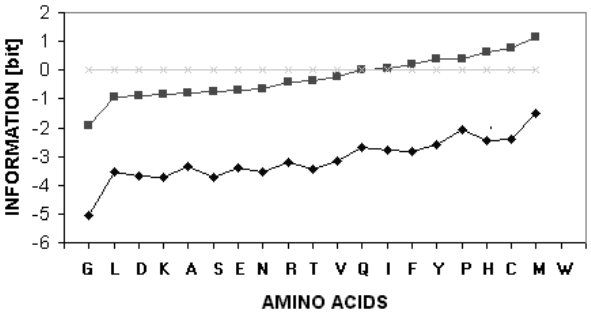


Figure 6. Plot representing the relation between amount of information (bits) carried by individual amino acid (on the basis of its frequency) and amount of information (bit) necessary to select grid point of its Phi, Psi map (10 degs grid step size) (-◆-) and amount of information (bit) necessary to predict particular ellipse fragment (-■-). The horizontal line visualizes the zero level.

The early-stage structural forms of these proteins have been taken as input structures for energy minimization procedure (according to ECEPP/2). Two alternate protocols have been applied: energy minimization procedure without any additional conditions and energy minimization with SS-bonds constraints (when present). The structures found by both approaches are shown in Figs 7., 8. and 9.

The number of non-bonding contacts in the final structures was calculated showing the approach to the number of non-bonding contacts as they appear in native structural forms (Fig.10.A, B, C D for BPTI, Fig.11 for ribonuclease and Fig 12. for lysozyme).

Another criterion to verify the degree of approach to the native structural form was the size profile of the vectors linking the geometrical center of the molecule with sequential Cα atoms (Fig.13 for BPTI, Fig.14 for ribonuclease and Fig.15 for lysozyme). Analysis of these profiles shows this to be a promising method. Some fragments in these profiles were distinguished according to their characteristics, and were color-coded. The RMS-D values for each fragment have been calculated to express quantitatively the differences between particular fragments. The same color scale was used in graphic presentation of the structures in Fig. 7, 8 and 9.

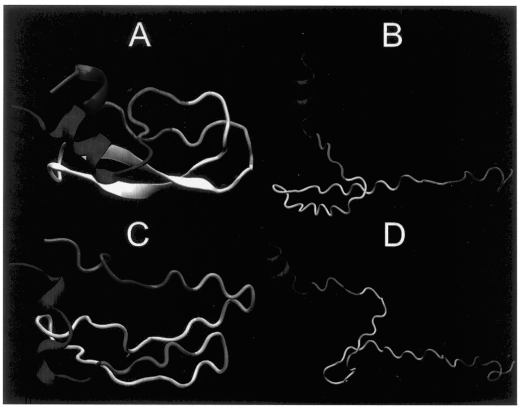


Figure 7. The structure of BPTI: A – native form, B – ellipse-based early-stage form, C – post-energy minimisation with SS-bonds present and expressed as constraints, D – post-energy minimization structure with SS-bonds absent in energy minimization procedure.

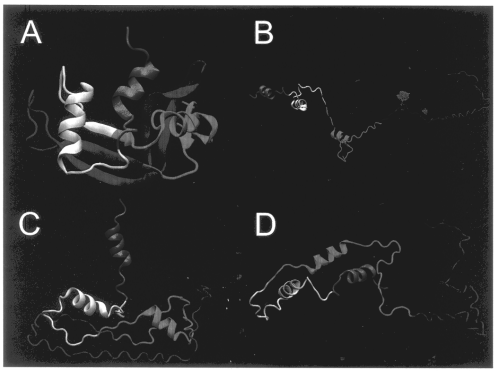


Figure 8. The structure of ribonuclease: A – native form, B – ellipse-based early-stage form, C – post-energy minimization with SS-bonds present and expressed as constraints, D – post-energy minimization structure with SS-bonds absent in energy minimization procedure.

Analysis of the final structures created according to early-stage model shows that the limited conformational sub-space delivers structures able to change and fluently approach the native-like structural forms without any steric collisions.

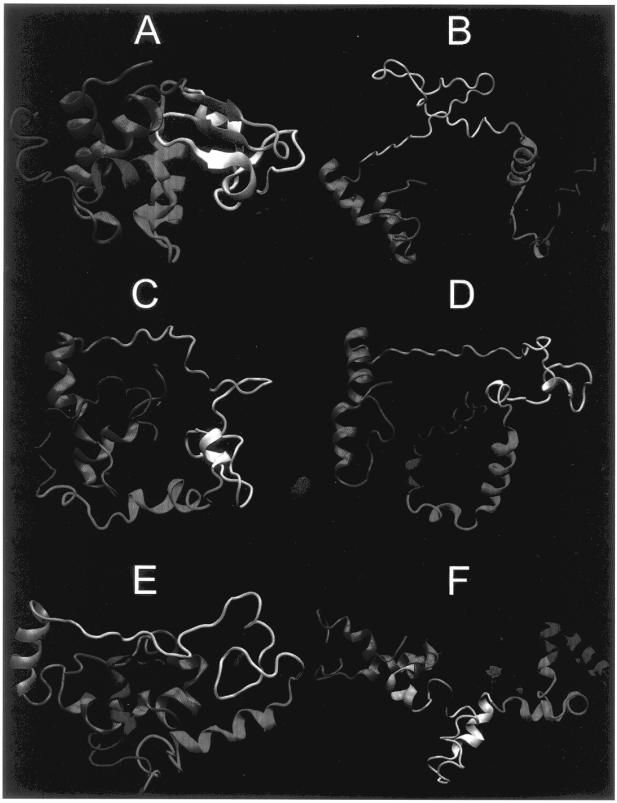


Figure 9. The structure of lysozyme: A – native form, B – ellipse-based early-stage form, C – post-energy minimization with SS-bonds present and expressed as constraints, D – post-energy minimization structure with SS-bonds absent in energy minimization procedure, E – post-dynamics structural form with SS-bonds not declared and F – post-dynamics structure with SS-bonds declared.

Structural alphabet

The early-stage structures expressed by Phi_e, Psi_e angles can be expressed in the form of structural alphabet. This idea has been employed in other approaches in which short polypeptide fragments were categorized into classes representing typical structural motifs, producing a structural alphabet [44-47]; particular structural forms can be attached to particular sequences, allowing prediction of the structure of protein molecules of any size. A similar strategy is adopted in our approach.

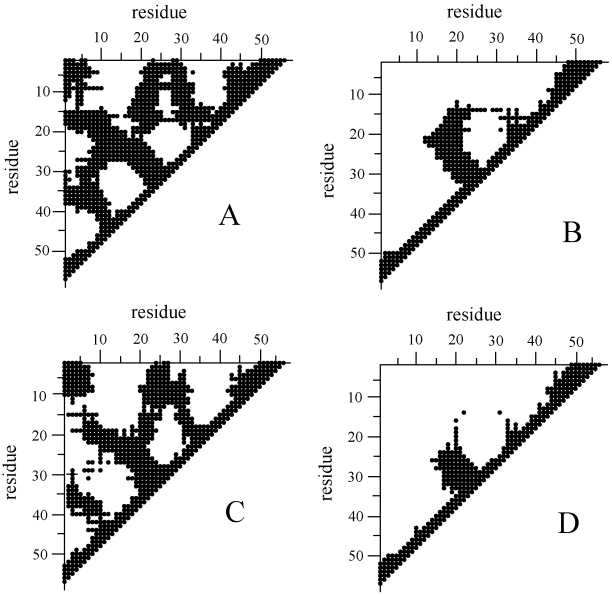


Figure 10. Nonbonding contacts in BPTI: A – native form, B – ellipse-derived structure, C-post-energy minimization with SS-bonds present, D – post-energy minimization without SS-bonds declared.

The procedure for defining the early-stage structural alphabet was as follows:

1. all proteins present in PDB were represented by their Phi, Psi angles
2. the distribution of Phi, Psi angles has been distinguished for each amino acid independently - the three dimensional distribution on the Ramachandran map is characteristic for each amino acid. The three dimensional graphical presentation of Phi, Psi angles distribution for selected amino acids: ASP and ILE are shown in Fig. 5.
3. all Phi, Psi angles were transformed into their Phi_e, Psi_e. The criterion of shortest distance was applied. The black line shows the ellipse path according to eq.2 (Fig. 16.A, and Fig. 16.C)
4. when all twenty probability profiles are put together, the probability distribution of Phi_e, Psi_e along ellipse path appears as shown in Fig. 16. B.

Seven probability maxima can be distinguished in this profile and coded as shown in Fig. 16.B.

The interpretation of the profile is as follows:

1. the t-value expresses the ellipse equation variable
2. the starting point of t=0 is the point described as Phi = 90 deg and Psi = -90 deg and then clock-wise as shown on the Fig. 16.C.
3. the probability maxima interpretation is as follows:
   C – right-handed helix region
   E, F – extended and β-structure-like forms
   G – left-handed helix

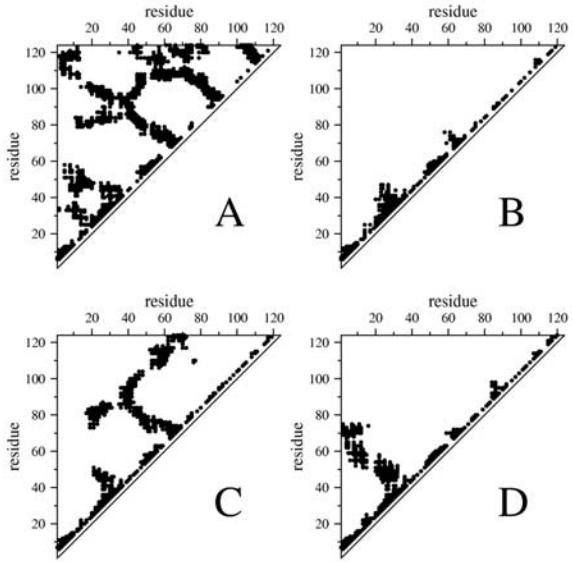


Figure 11. Nonbonding contacts in ribonuclease: A – native form, B – ellipse-derived structure, C-post-energy minimization with SS-bonds present, D – post-energy minimization without SS-bonds declared.

The listed codes identify the probability maxima that represent well defined structural forms. A, B, D codes represent structural forms not identified in any structural analyses and probably are treated together as random coil form. However introduction of the A, B and D structural forms enables better identification of structural forms generally treated as random coil. Structural forms in this group are highly differentiated, and these letter codes may improve their identification.

The maximum D between helix and β-structure is very interesting; it may be interpreted as intermediate between these two regular forms. This probability maximum is occupied mostly by ASN.

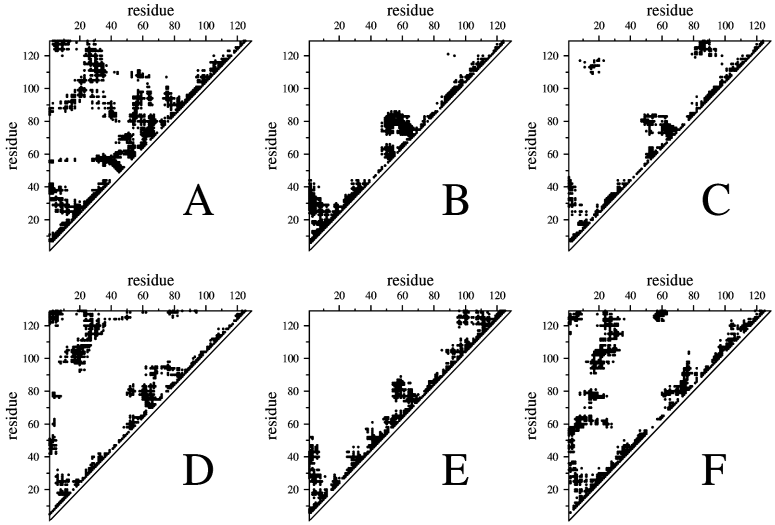


Figure 12. Nonbonding contacts in lysozyme: A – native form, B – ellipse-derived structure, C-post-energy minimization with SS-bonds present, D – post-energy minimization without SS-bonds declared E – post-dynamics with SS-bonds absent and F – post-dynamics form with SS-bonds declared.

The structure of any polypeptide chain in a protein can be represented as a string of codes similar to the one letter coding system used to express amino acid sequence (Fig.17.A and Fig.17.B). Another advantage of this classification system is that all sequence alignment tools can be applied to comparative structural analysis (Fig.17.B and Fig.17.C).

Sequence-to-structure and structure-to-sequence relation

A simple consequence of introducing structural codes is that the sequence-to-structure and structure-to-sequence relation can be presented in the form of a contingency table. Term "structure" means early-stage structure here, assuming that the ellipse-path really represents the limited conformational sub-space determined solely by backbone structure [48].

Theoretically the size of the contingency table describing this problem is of 2 401 (7^4) (structure codes) x 160 000 (20^4) (sequence codes). The table of sequence-to-structure and structure-to-sequence representation was created for the complete PDB data set (January 2003 release). The table as well as its upgraded form is available on-line (www.bioinformatics.cm-uj.krakow.pl).

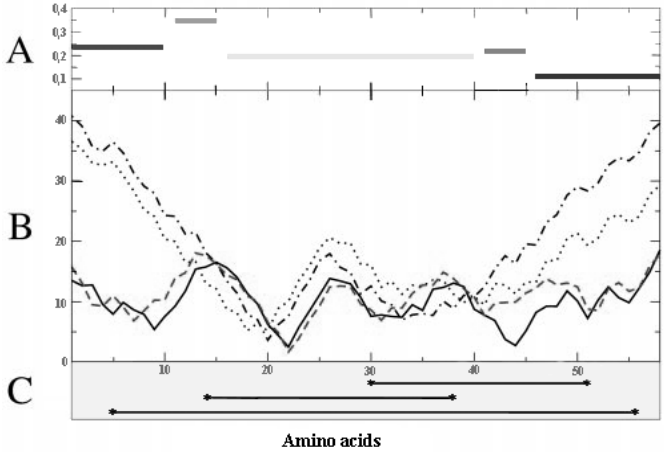


Figure 13. Comparison of structural forms of the BPTI molecule. A – RMS-D (per residue) calculated for structurally differentiated polypeptide fragments defined according to the profile presented in B. The parallel fragments of curves represent correct spatial orientation of the polypeptide, whereas the dissimilar regularity represents the low similarity of the spatial orientation of particular polypeptide fragment. B – profile representing the distribution of distance linking the geometrical center of molecule with sequential Cα atoms. Continuous solid line – native form, dotted line – ellipse-derived structure, dashed line – post-energy-minimization with SS-bonds declared, dotted/dashed line – SS-bonds not declared in energy minimization procedure. C – SS-bonds system in BPTI.

Each cell of the contingency table represents probability of a particular tetrapeptide to represent a particular early-stage structural form (and vice versa).

Finally, the contingency table of size 146 940 × 2 397 was analyzed for the mutual dependence and correlation between a particular tetrapeptide sequence and its structure.

The total number of different tetrapeptides in the January 2003 release of PDB was found to be 1 529 987.

Global analysis of the contingency table revealed that the maximum number of different structures attributed to the same tetrapeptide sequence is 144. This tetrapeptide appeared to be **GSAA**. The maximum number of different sequences was found to be 90 587 for alpha-helix (*CCCC*) and 47 809 for β-structure (*EEEE*). Only four structures were not found in the library: *ABAB*, *ABBD*, *ABFB* and *DBAB*.

How can such a large table be analyzed? A method based on informational entropy was applied to search for pairs (sequence and structure) of strong and weak mutual dependence (influence). Another method employing statistical

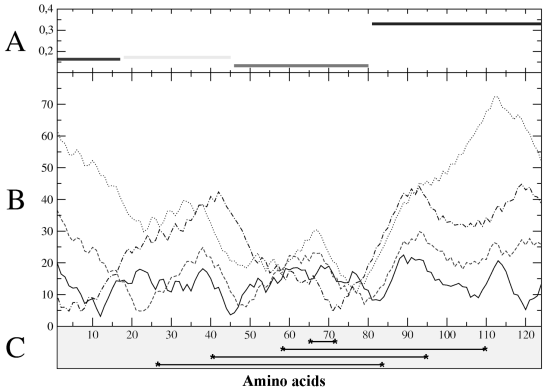


Figure 14. Comparison of structural forms of the robonuclease molecule. A – RMS-D (per residue) calculated for structurally differentiated polypeptide fragments defined according to the profile presented in B. The parallel fragments of curves represent correct spatial orientation of the polypeptide, whereas the dissimilar regularity represents the low similarity of the spatial orientation of particular polypeptide fragment. B – profile representing the distribution of distance linking the geometrical center of molecule with sequential Cα atoms. Continuous solid line – native form, dotted line – ellipse-derived structure, dashed line – post-energy-minimization with SS-bonds declared, dotted/dashed line – SS-bonds not declared in energy minimization procedure. C – SS-bonds system in ribonuclease.

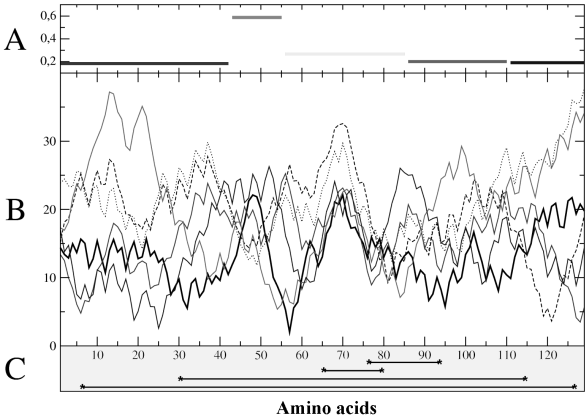


Figure 15. Comparison of structural forms of the lysozyme molecule. A – RMS-D (per residue) calculated for structurally differentiated polypeptide fragments defined according to the profile presented in B. The parallel fragments of curves represent correct spatial orientation of the polypeptide, whereas the dissimilar regularity

Figure 15. Legend continued

represents the low similarity of the spatial orientation of particular polypeptide fragment. B – profile representing the distribution of distance linking the geometrical center of molecule with sequential Cα atoms. Continuous solid line – native form, dotted line – ellipse-derived structure, dashed line – post-energy-minimization with SS-bonds not declared, continuous red line – SS-bonds declared in energy minimization procedure, continuous green line – post-dynamics structure with SS-bonds absent in energy minimization procedure and continuous blue line – post-dynamics and energy minimization procedure with SS-bonds present. . C – SS-bonds system in lysozyme.

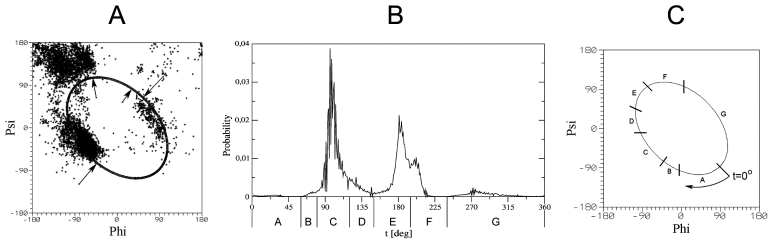


Figure 16. The structural letter codes definition. A – graphical presentation of the Phi, Psi angles movement toward the ellipse path, B – the probability profile (representing all amino acids) along the ellipse path after transformation shown in A. Each probability maximum is represented by letter code. The t-parameter expresses the t-parameter of ellipse equation (eq. 2.). The starting point (t=0) represents the point Phi=90 degs and Psi =-90 degs and then increases according to clock-wise movement along ellipse as shown in C.

tools (correlation coefficient ρ for qualitative variables) was also used. Together these two methods provided a check on each other. They were used to select pairs (sequence and structure) of high mutual determinability, if any.

Information entropy as a measure of sequence-to-structure determinability

High values of probability calculated as above can disclose highly coupled pairs of structure and sequence. A ranking list of the probability values can extract the highly determined relations for both sequence to structure and structure to sequence [48].

Structural determinability can also be measured by calculating informational entropy (SE) calculation.

SE reaches its maximum value for a set of events of equal probability (all p_i equal to each other), that is, each i-th solution is equally probable for the event under consideration, and no solution is preferred. The maximum value of

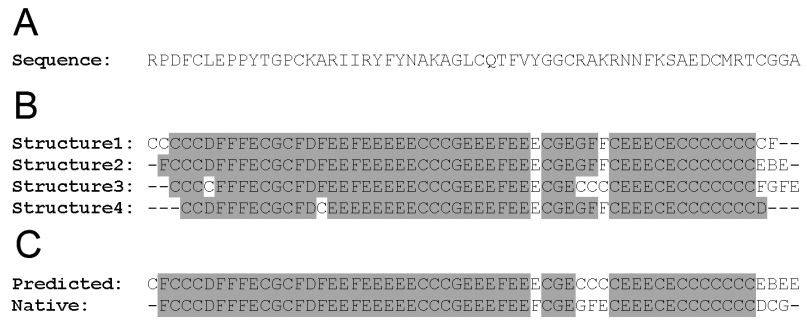


Figure 17. SPI coefficient definition – A – amino acid sequence, B- four forms of structural codes attribution to sequential tetrapeptides (overlapped system) and C – predicted structural codes in comparison to the observed one for protein under consideration. Prediction results for BPTI. Grey – identity. SPI: 95.0, Q3: 91.1, Q7: 87.5, SOV: 89.7.

Table 2. The ten top structures recognized by decrease of entropy *(*Δ*SE* [bit]) in sequence-to-structure (left half of table) and structure-to-sequence relation (right part of table). The bold symbols express the sequence and italics – structural codes.

Sequence	Structure	SE [bit]	SEmax [bit]	**ΔSE [bit]**	*Structure*	**Sequence**	SE [bit]	SEmax [bit]	**ΔSE [bit]**
AAAA	*CCCC*	2.29	6.44	**4.15**	*GCFG*	**DGSG**	4.82	7.99	**3.17**
GDSG	*GCFG*	1.57	5.49	**3.92**	*AEED*	**GLRL**	3.86	6.81	**2.95**
AVRR	*CCCC*	1.04	4.95	**3.91**	*BACE*	**GGAE**	2.20	5.09	**2.89**
LAAA	*CCCC*	1.77	5.61	**3.84**	*EAEG*	**IGIG**	4.79	7.68	**2.89**
EAEL	*CCCC*	1.37	5.21	**3.83**	*AEGE*	**GIGH**	4.74	7.63	**2.89**
LDKA	*CCCC*	1.30	5.09	**3.78**	*BFBE*	**PEPV**	2.28	5.13	**2.85**
DAAV	*CCCC*	0.69	4.46	**3.77**	*AEGD*	**GNES**	2.09	4.91	**2.82**
AKLK	*CCCC*	0.76	4.52	**3.77**	*EBCB*	**ELPD**	3.68	6.38	**2.70**
DSGG	*CFGF*	1.97	5.73	**3.76**	*EBFB*	**FPEP**	2.57	5.17	**2.60**
ELAA	*CCCC*	1.30	5.04	**3.75**	*AFFB*	**GFRN**	2.03	4.58	**2.55**

SE depends on the number of possible solutions for the event. *SE* equal to zero (or to one) represents the determinate case in which only one solution is possible. The higher the difference between SE_{max} and *SE* describing particular space of events, the higher the determinability in the given case. A large difference between SE_{max} and *SE* means that the case is realized by a few solutions and that some of them occur with higher probability, which is interpreted as a case with higher determinability (biased event). Such a measure scale has been applied to elucidate the rows (or columns) of high bias [48] (Tab.l).

Values of *SE*, SE_{max} and ΔSE can be calculated for all rows (structural preferences versus amino acid sequence) and for columns (preference of a sequence for a particular structural form) in the contingency table. The first values extract structures highly determined by the sequence; the second values extract structures highly associated with a particular sequence. The ten top structures and ten top sequences of tetrapeptides (according to ΔSE) are shown in Tab.2.

Since the value of SE_{max} depends on N (number of the non-zero cells in a row or column), the relative ΔSE needs to be calculated. Finally, the values of $(SE_{max} - SE)/ SE_{max}$ are treated as the ranking coefficient for structure-to-sequence (rows) or sequence-to-structure (columns) determinability. The ten top structures and ten top sequences of tetrapeptides are shown in Tab.3. for the relative *SE* values.

Table 3. The ten top structures selected by relative entropy decrease (RelΔSE [bit]). in sequence-to-structure (left half of table) and structure-to-sequence relation (right part of table). The bold symbols express the sequence and italics – structural codes.

Sequence	Structure	SE [bit]	SEmax [bit]	**RelΔSE**	**Structure**	Sequence	SE [bit]	SEmax [bit]	**RelΔSE**
SHCL	*CCCC*	0.043	1.000	**0.957**	*BAFB*	**IDFP**	0.391	1.000	**0.609**
IERM	*CCCC*	0.046	1.000	**0.954**	*AADF*	**GGPL**	1.830	4.322	**0.577**
RMLQ	*CCCC*	0.114	2.322	**0.951**	*AEGD*	**GNES**	2.086	4.907	**0.575**
WVAW	*ECCC*	0.118	2.322	**0.949**	*BACE*	**GGAE**	2.196	5.087	**0.568**
ELHC	*CCCC*	0.083	1.585	**0.948**	*AFFB*	**GFRN**	2.031	4.585	**0.557**
MVFQ	*CCCC*	0.149	2.807	**0.947**	*BFBE*	**PEPV**	2.279	5.129	**0.556**
ANWM	*CCCC*	0.089	1.585	**0.944**	*ADBD*	**ADTE**	1.404	3.000	**0.532**
QLCI	*CCCC*	0.056	1.000	**0.944**	*AEBB*	**GPVY**	1.648	3.459	**0.524**
VWAK	*CCCC*	0.057	1.000	**0.943**	*EBFB*	**FPEP**	2.570	5.170	**0.503**
PNRA	*CCCC*	0.093	1.585	**0.941**	*BEAA*	**GKGS**	1.580	3.170	**0.502**

The ten top structures according to sequence and to structure determinability are also shown in Fig. 18.A.

Another statistics-based method was applied to analyze the contingency table [49]. Some selected structures found according to this method are shown in Fig.18.B.

A surprising finding is that the structure-to-sequence relations of the loop-like rather than the regular forms are highly determined (see Fig. 18.). We again emphasize that the whole discussion regards the classification of early-stage structures.

Structure Predictability Index (SPI)

The contingency table expressing the sequence-to-structure and structure-to-sequence relations on a probability scale can also be used for protein structure prediction, in particular to estimate the degree of difficulty in

structure prediction. Structure predictability based on contingency table is possible to the extent of letter code recognition. Since the probability values for selection of a particular structural code are known from the contingency table, the probability of adoption of a particular structure can be predicted. Thus the degree of difficulty in structure prediction for a particular amino acid sequence can be estimated in an *a priori* system. According to CASP experiences, the easy-to-predict and difficult-to-predict structures have been distinguished [50,51]. So far the degree of difficulty of structure prediction has been measured in an *a posteriori* classification system. Our contingency table expressing the predictability of a particular structure (the probability of a particular sequence to represent a particular structure) allows this procedure to be applied even when the native form is not known.

Since tetrapeptide has been taken as the unit, there are four possibilities of structure-string for the sequence expressing string. (Fig. 17.A.). The example of this procedure is given in Fig.17.B. Tracing the highest probability values

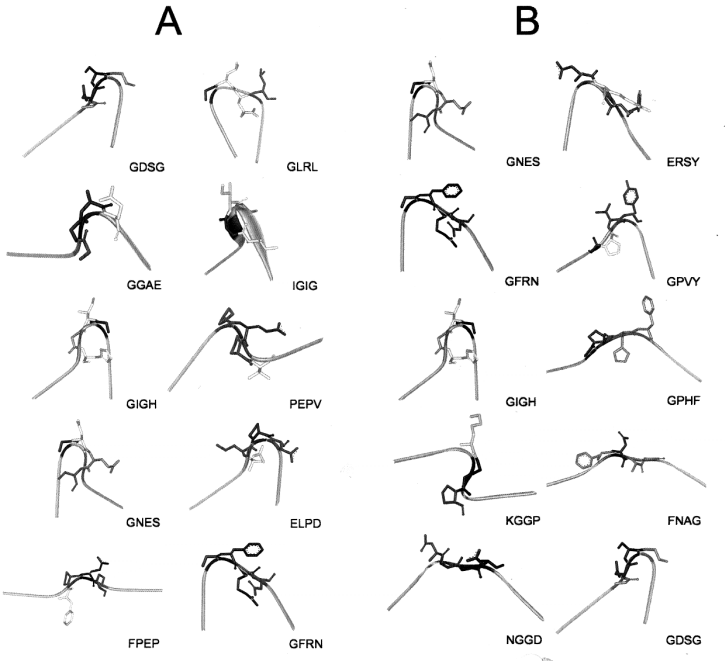


Figure 18. The structural motifs for tetrapeptides found as A - highly determined according to structural entropy based procedure and according to B – ρ^2 statistics-based. Symbols represent sequences of tetrapeptides.

for each structure attributed for particular amino acid, the procedure is able to elucidates the consensus structure shown in Fig.17.C.

The Structure Predictability Index (SPI) coefficient expresses the highest probability found in contingency table for particular tetrapeptide. The normalized value of the SPI coefficient allows comparison of sequences, distinguishing sequences of low and high difficulty of structure prediction [52].

The traditional estimation of prediction accuracy was based on the Q3 coefficient. Three structural forms are distinguished in that approach: helical, extended and random coil. The Q3 coefficient expresses the percentage of correctly predicted amino acids (three categories) versus the total number of amino acids in polypeptide chain. Since seven structural forms (seven probability maxima in probability profile along ellipse have been defined), the Q7 coefficient measuring the correctness of structure prediction can be introduced. Seven structural forms (seven probability maxima in the probability profile along the ellipse) have been defined. The Q7 coefficient expresses the percentage of correctly predicted amino acids in the polypeptide chain under consideration. Our approach uses the classification of structures in early-stage folding. These seven structural forms are as follows: C- right handed helix, E,F – β-structural forms, G-left handed helix; and A, B, D – the forms which can be treated as random coil, although different forms of random coil can be distinguished in our model.

The results of comparative analysis of classical Q3, SOV and newly introduced SPI and Q7 parameters for the complete set of proteins deposited in PDB #2003 revealed high accordance between them [52-55]. Since the SPI coefficient can be calculated for amino acid sequences without knowledge of the final native structure, it gives the opportunity to estimate difficulty in an *a priori* system. The degree of determinability may be particularly useful indicator in predicting protein structure prediction by threading methods [56-59].

Structure based on partial unfolding versus the predicted model

The scheme presented in Fig. 19. gives the global overview of the procedure. Partial unfolding leads to well defined Phi_e, Psi_e angles (continuous model). When an unknown structure is to be predicted, it is possible only to the extent of letter codes (discrete model) meaning that Phi_{em} and Psi_{em} of maximum probability (index m) can be found. The relation between the structural form based on partial unfolding (continuous model) and the prediction (discrete model) is shown. Some differences may be seen between these two structural forms, although the similarity is quite high (the bends are

usually localized properly), moreover, steric hindrances disabling fluent structural changes are absent.

The contingency table was created (partial unfolding path) starting with the native form of proteins (calcium-binding region of alpha-lactalbumin - PDB code 1A4V has been used as the example in this scheme) (Fig.19 A). Transformation to the to ellipse-path-limited conformational sub-space (Fig.19 B) allowed creation of the early-stage structure (Fig. 19.C). This transformation adopted to all known proteins, revealed the probability profile in ellipse path-limited conformational sub-space (Fig.19. D). The coding system for early-stage folding (*in silico*) applied to all known proteins, enabled creation of the contingency table expressing the sequence-to-structure and structure-to-sequence relation.

The structure prediction path starts with a known amino acid sequence (Fig.19.F). The contingency table can be exploited to search for tetrepetides of particular sequences adopting corresponding structural codes (Fig.19.F). The structural codes allow only the fragment of the ellipse to be predicted. The exact value of appropriate Phi, Psi angle is to the those representing the probability maximum (Fig. 19.G). Fig.24.H shows the structure created according to the structural codes applied to the same protein as in the partial unfolding path. The structures shown in Fig.19.F and Fig. 19.C visualize the difference between discrete (Fig. 19. G) and continuous (Fig. 19. D) probability profiles. The energy minimization procedure (Fig. 19.I) applied to the structure created in step represented in Fig.19.H produces the structure shown in Fig. 19.J. This result is obtained by simple energy minimization. There is still no late-stage folding model.

How does the model work ?

The protein structure prediction path shown in Fig.19. has been applied to protein structure prediction for targets in CASP6 competition https://www.predictioncenter.llnl.gov. All targets of N<150 (N-number of amino acids) were used to validate the reliability of the model. Highest probability (highest SPI) was taken as the criterion for early-stage structure prediction. The goal of participation was not to achieve a high score but rather to estimate whether the received structures look promising in terms of structure prediction.

The structural letter codes predicted on the basis of contingency table and SPI calculation for our best approach T0215 are shown in Fig. 20.A. The early-stage structure of this protein is also shown together with the crystal structure (Fig.20.B) of this protein. The unified assessment used by CASP6 organizers localized our result versus others is shown in Fig.20.C.

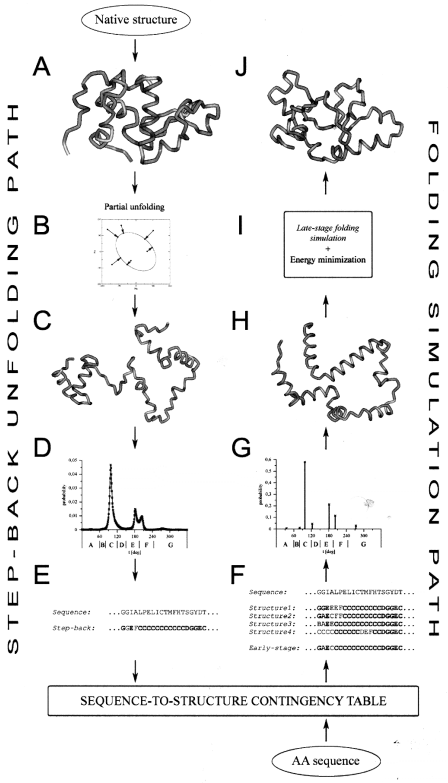


Figure 19. The scheme representing the complete procedure flow in two versions: The step-back procedure (A-E) representing partial unfolding starts when native structure is known. The Phi, Psi angles found for protein under consideration (PDB code 1A4V). The early-stage structure (C) created according to Phi_e, Psi_e angles (obtained according to B). The probability profile obtained on the basis of the complete PDB data base defined the structural codes(D). The amino acid sequence together with the appropriate structural codes (E). All structurally known proteins represented by amino acid sequence (tetrapeptides) and structural codes create the contingency table expressing the sequence-to-structure and structure-to-sequence relation. This contingency table can be used for folding simulation path (F-J). This path can be applied to known amino acid sequence (F). Each structural code represents particular Phi, Psi angles (mediated by t-parameter) (G) although only in discrete form. The early-stage structure created according to discrete structural recognition (H) differs versus the one obtained according to step-back procedure. Energy minimisation procedure (I) applied to structure (H) produces the structure shown in J. The low difference between these two structures (A and J) is the aim of all structure prediction methods. So far only energy minimisation has been applied in step I, although the late-stage folding simulation is under consideration.

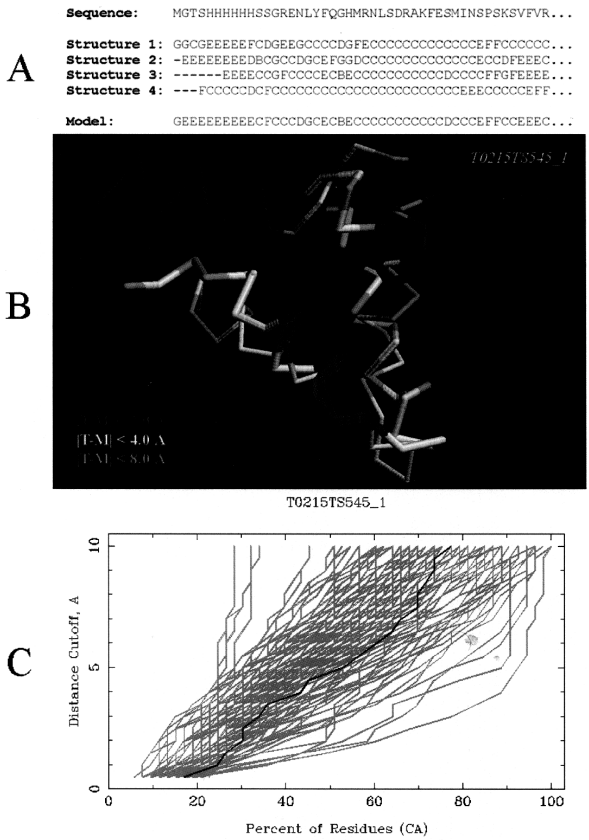


Figure 20. CASP6 result – our best approach (T0215) received according to procedure shown in Fig.19. A – amino acid sequence and four structural forms expressed by structural codes. The lower line – the consensus structural form. B – observed and predicted structural forms of target T0215 as assessed by CASP6 organizers. C – the correctness of prediction. Black line represents our model in comparison to others participants.

Our worst approach was for target T0196. The native structure of this protein represents the β-barrel. The consensus letter codes found for amino acid sequence for this protein (shown in Fig.21.A) suggest rather a helical structure (codes C), although the β-like structural forms (codes E) also occurred in some overlapping. The early-stage structural form created according to the consensus structural letter codes together with the native

structure of this protein are shown in Fig.21.B. The model seems to fail in this case, although the structures shown in Fig.21. created separately according to each line of sequence-to-structure relation demonstrate higher content of β-like structure (Fig. 21. D-G).

Figure 21.

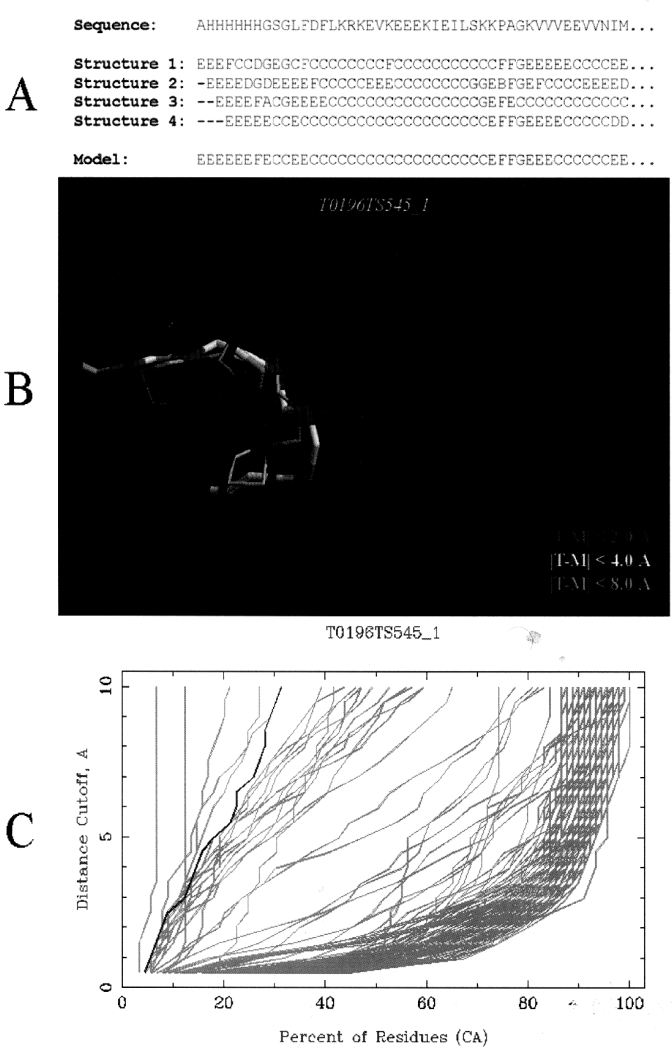


Figure 21.

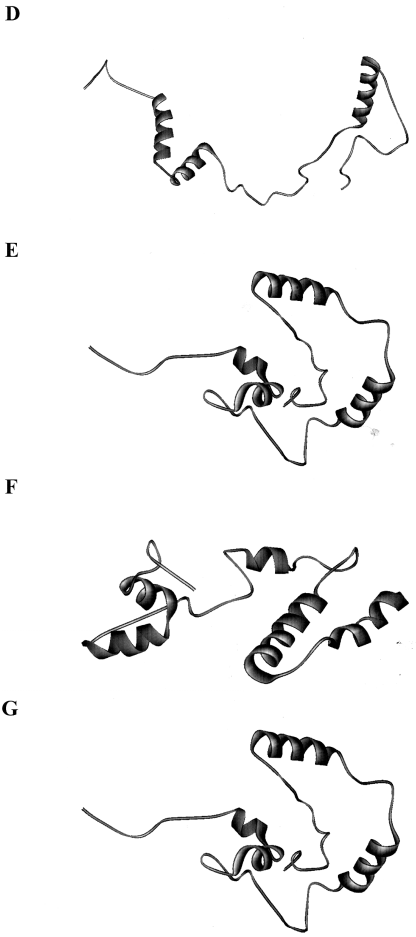


Figure 21. CASP6 result – our worst approach (T0196) received according to procedure shown in Fig.19. A – amino acid sequence together with structural codes alignment. The lower line – structural codes as found for given amino acid sequence. B – results of comparison as assessed by CASP6 organizers. C – the prediction correctness. Black line – our approach. One shall underline, that the low discrepancy structures represent only comparative modelling-based approaches. D, E, F, G – structures created according to four structural codes as shown in A. The thin straight line represents the β-like structural forms not distinguished as β-structure by the program used to draw these pictures.

Comparison of the results of T0196 and T0215 and our other approaches taught us that helical structure may be overestimated in comparison to β-like structural forms in contingency table. This should be taken into account whenever β-like structure appears in the letter codes. Both examples; best and worst were presented [60,61].

Biological function presence in early-stage structural form

As said before, all methods prepared for multiple sequence alignment tools can be applied to structure coding strings. This procedure was applied to reveal structurally conservative fragments in the serpine family (47 proteins found in PDB).

The steps of such procedure are shown in Fig.22. Only a few proteins were selected to show the calculation aimed at selecting structurally conservative fragments. The original sequence of structure related codes in arbitrary order (Fig.22 A) reveal some conservative fragments (Fig. 22.B). Then the frequency can be found for each position occupied by particular code (Fig. 22.C). Next, the maximal frequencies are selected (Fig.22 D). The weight value is calculated in the following step of the procedure (Fig. 22.E). When the window size is introduced (it is equal to 5 amino acids in our case) the resulting high W value (eq. 10) selects fragments of high structural conservation (Fig.22. F). Exactly the same procedure can be applied to sequence multiple alignment enabling comparative analysis of sequence and structure.

$$W = 10 * \log_{10} \frac{f+1}{N/7+1} \qquad \text{[eq. 10]}$$

This is the procedure applied to all 47 serpine family members to examine whether the highly conservative sequence and structural fragments stand in any relation to polypeptide chain fragments related to biological function of these proteins. Biological function has been defined as the presence of A β-sheet in the early stage folding structural form. The A β-sheet is responsible for scavenging different molecules from sera in states after massive proteolysis process (acute state) [62-67].

Biological function has been defined as possibility for A β-sheet to incorporate any other polypeptide chain fragment or some other molecules (Fig.23 A and B).

This problem can be extended to the general question of the presence of biological function-related structural motifs already in early-stage folding.

The result of this analysis is shown in Fig. 23. D, E and F where a color scale (Fig. 23. C) is used to visualize high conservation status.

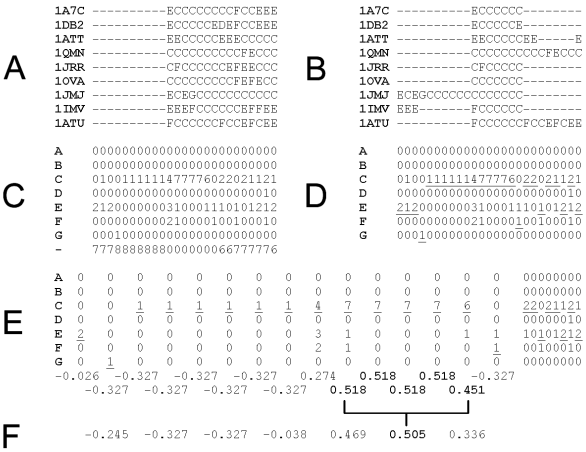


Figure 22. Procedure of W (weight of conservation status) applied to serpine family (fragment shown). A – structural codes arbitrarily ordered, B – results of multiple alignment, C – frequency for particular structural code, D – maximum frequency distinguished, E – the W-values (according to eq.10) calculated for each position and F – for window size of five amino acids. Bold distinguishes the fragment of high conservation status.

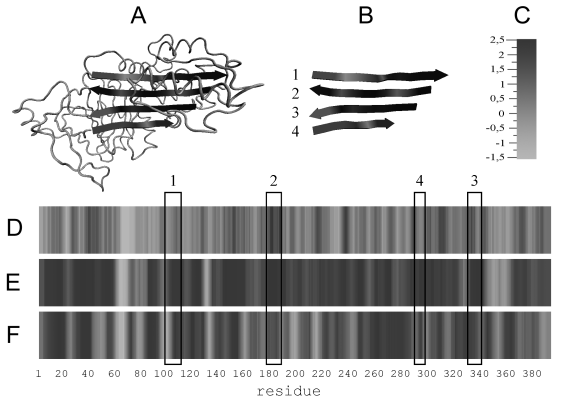


Figure 23. Presentation of W-value (weight of conservation status) for uncleaved ovalbumin (1OVA) as representing the serpine family. A – native structure with A β-sheet (function related) coloured blue, B – isolated A β-sheet with individual strands numbered, C – colour scale of W value, D,E,F – profiles of W-value along the polypeptide chain calculated for sequence, step-back and early-stage conformation multiple alignment respectively. Black bars indicate locations of individual strands in B.

The position of highly conserved sequences and structural forms (estimated on the basis of early-stage folding) seems to represent high agreement, suggesting that at least in this case the biological function-related fragments are already present in early-stage folding stage.

Secondary structure was identified only in a case of serpine family analysis. However, the presence of β-fragments necessary to create the A β-sheet has been positively verified.

More work shall be done to estimate whether other biological function-related elements can be seen in early-stage folding. The work focused on enzymes is currently under way in our group.

Structural similarity search in proteins

The two geometrical parameters introduced in the model, V-angle and R-radius of curvature, can also be applied to the search for structural similarity. Since each protein structure can be represented by the profile of V-angle and R-radius value, why not to compare the profiles representing these geometrical characteristics? The similarity of V-angle, R-radius can be estimated, and simultaneous comparison of the two parameters is also possible [68].

This model for the structural similarity search has been applied to the serpine family of proteins (to represents the large protein molecules of differentiated structural forms). The profiles of V-angle and lnR for two selected proteins (PDB - 1OVA, 1ATT) belonging to the serpine family are shown in Fig. 24 (A, B, C, D). The results of comparison of V and lnR profiles presented in form of a dot-matrix in Fig. 24. E, F. The serpine family compared according to sequence alignment (selected proteins) are shown in Fig. 25.A. The structure comparison based on radius of curvature and V-angle is shown in Fig. 25.B. The structure comparison as received according to DALI program [69] is also shown in Fig.25.C.

Protein folding simulation rather than protein structure prediction

This presentation was focused on the early-stage folding model (*in silico*). There is still no satisfactory model for late-stage folding (*in silico*) (Fig. 19.I). The group is currently focusing on a heuristic model for the next step in folding process simulation, following the early-stage folding model. The CASP experience shows that comparative modeling is delivering better and better results [70]. However, comparative modeling cannot present a mechanism of protein folding explaining why proteins fold the way they do.

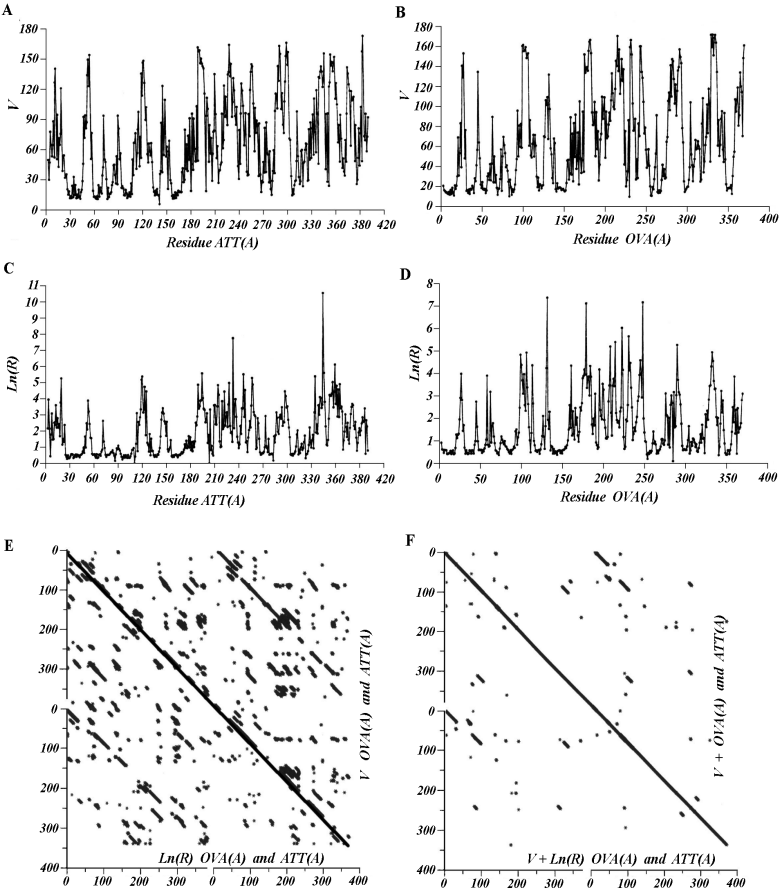


Figure 24. The structural similarity search based on the ln(R) and V-angle distribution along the polypeptide chain in serpine family. A – V-angle distribution in ATT (chain A), B – V-angle distribution in OVA (chain A), C – ln(R) profile in ATT and D – ln(R) profile in OVA. Dot matrixes for OVA and ATT comparison. E – left lower part – ln (R) comparison, upper right part – V-angle comparison. F – dot-matrix for simultaneous comparison of ln(R) and V-angle.

Experiments that address protein folding yield new information every day. Applied to computational models, this information may help to solve the problem of protein folding simulation, taking into account intermediates of folding and their possible relation to biological activity.

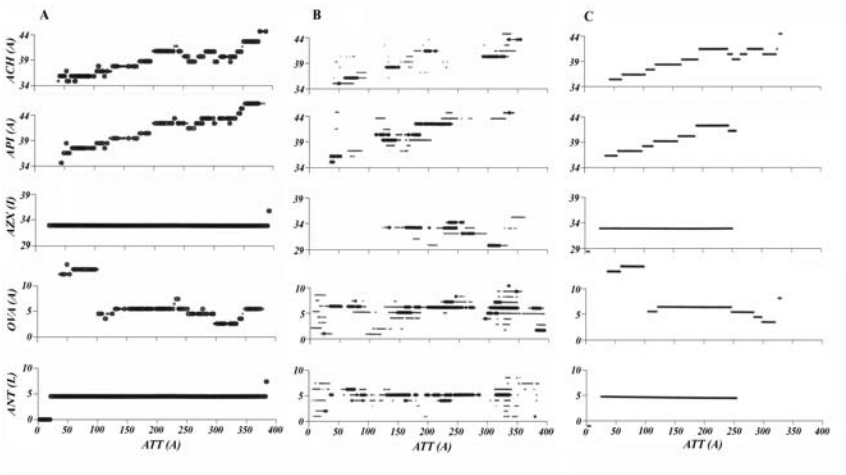


Figure 25. Similarity of selected serpine family members: A – sequence similarity, B – structure similarity based on the model described, C – structural similarity according to DALI program [69]. The x-axis represents the sequence of ATT(chain A) used as a protein versus which the selected proteins are compared. The y-axis represents the relative numbers in a particular protein. The thick lines in B represent the higher similarity score, thin lines the lower similarity score. The symbols in parentheses identify the chains taken into analysis.

References

1. Anfinsen, C.B., and Scheraga, H.A.1975, Adv. Protein Chem., 29, 205.
2. Burgess, W., and Scheraga, H.A.1975, Proc. Natl. Acad. Sci., U.S.A, 72, 1221.
3. Englander,S.W. 2000, An. Rev. Biophys. Biomol. Struct., 29, 213.
4. Booth, P.J., Flitsch, S.L., Stern,L.J., Greenhalgh, D.A., Kim, P.S., and Khorana,H.G. 1995, Nature Struct Biology 2, 139.
5. Brown, S., and Head-Gordon, T. 2004, Prot. Sci. 13, 958.
6. Ozkan, S.B., Dill, K.A., and Bahar, I. 1992, Eur. J. Biochem., 204, 759.
7. Duan, Y., and Kollman, P.A. 1998, Science, 282, 740.
8. Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E., and Razgulyaev, O.I. 1990, FEBS Letters 262, 20.
9. Chauviere, M., Martinage, A., Debarle, M., Sautiere, P. and Chevaillier, P. 2000, Eur. J. Biochem. 267, 2452.
10. Bahar,I., and Jernigan, R.L. 2001 Protein Science, 10,1216.
11. Reader, J.S., Van Nuland, N.A.J., Thompson, G.S., Ferguson, S.J., Dobson, C.M., and Radford, S.E. 1996, FASEB J, 10, 67.
12. Matthew, E.C., Wood, J., Fink AL, and Kliger, D.S. 1998, Biochemistry, 37, 5589.
13. Creighton, T.E., Darby, N.J., and Kemmink, J. 1996 The FASEB J., 10, 110.
14. Kim, P.S., and Baldwin, R.L. 1990 Ann. Rev. Biochem. 59, 631.
15. Clarke, A.R., and Waltho, J.P. 1997, Curr. Op. Biotechnol. 8, 400.

16. Dill, K.A., and Chan, H.S. 1997, Nat Struct Biol 4, 10.
17. Roder, H., and Colon, W. 1997, Curr Opin Struct Biol 7, 15.
18. Ferrer, M., Barany, G., and Woodward, C. 1995, Nature Struct. Biol., 2, 211.
19. Creighton, T.E. 1997, Trends Biochem. Sci. 22, 6.
20. Derfield, C., Smith, R.A.G., and Dobson, C.M. 1994 Stuct. Biol 1, 23.
21. Baldwin, R.L. 2002, Science 295, 1657.
22. Dobson, C.M., and Karplus, M. 1999 Curr. Opin. Struct. Biol. 9, 92.
23. Hayward, S. 2001, Prot. Sci. 10, 2219.
24. Roterman, I., and Konieczny, L. 1995, Computers and Chemistry 19, 247.
25. Roterman, I. 1995, Biochimie, 77, 204.
26. Roterman, I. 1995, J. theoretical Biol. 177, 283.
27. Daggett, V., and Levitt, M. 1992 J. Mol. Biol. 223, 1121.
28. Daggett, V., and Levitt, M. 1993 J. Mol. Biol. 232, 600.
29. Shannon, C.E.A. 1948, Bell Syst. Tech. J. 27, 379.
30. Jurkowski, W,. Bryliński, M., Konieczny, L., Wiśniowski, Z., and Roterman, I. 2004, Proteins: Struct., Func. Bioinformatics, 55, 115.
31. Alonso, D.O.V., and Daggett V. 1998, Prot. Sci. 7, 860.
32. Dobson C.M. 2001, Phil. Trans. R. Soc.Lond. B256,133.
33. Liwo, A., Czaplewski, C., Pillardy, J., and Scheraga, H. 2001, J. Chem Phys. 115, 2323.
34. Liwo, A., Arfukowicz, P., Czaplewski, C., Oldziej, S., Pillardy, J., and Scheraga, H. 2002, Proc. Natl. Acad. Sci. U.S.A. 99, 1937.
35. Pillardy, J., Liwo, A., Groth, M., and Scheraga, H. 1999, J. Phys. Chem. B103, 7353.
36. Wawak, R., Pillardy, J., Liwo, A.,and Scheraga, H. 1998, J. Phys Chem. 102, 2904.
37. Rippol, D., Piela, J., Vasquez, M., and Scheraga, H. 1991, Proteins Struct. Func. Genetics, 10, 188.
38. Piela, J., Kostrowicki, J., and Scheraga, H. 1989, J. Phys Chem. 93, 3339.
39. Piela, J. 2002, Handbook of global optimization, 2, pp 461-488, Ed: P.M. Pardalos, H.E. Romeijn, Kluwer
40. Colubri, A. 2004, J. Biomol. Struc Dynam. 21, 625.
41. Brylinski, M., Jurkowski, W., Konieczny, L., and Roterman, I. 2004. Bioinformatics 20, 199.
42. Brylinski, M., Jurkowski, W., Konieczny, L., and Roterman, I. 2004, TASK Quarterly 8, 413.
43. Jurkowski, W., Brylinski, M., Konieczny, L., and Roterman, I. 2004, J Biomol Struct Dyn 22, 149.
44. de Brevern, A.G., Valadie, H., Hzout, S., and Etchebest, C. 2002, Prot Sci 11, 2871.
45. de Brevern, A.G., Benros, C., Gautier, R., Valadie, H., Hazout, S., and Etchebest, C. 2004 In Silico Biol. 4, 0031.
46. de Brevern, A., Camproux, A.C., Hazout, S., Etchebst, C., and Tuffery, P. 2001, Recent Advances in Prot. Eng. 1, 319.
47. Etchebest, C., Benros,C., Hazout,S., and de Brevern, A. 2005, Proteins Struc Func Bioinfo 59, 810.
48. Brylinski, M., Konieczny, L., Czerwonko, P., Jurkowski, W., and Roterman, I. 2005, J. Biomed. Biotech. 2, 65.

49. Meus, J., Brylinski, M., Piwowar, M., Piwowar, P., Stefaniak, J., Wisniowski, Z., Jurkowski, W., and Roterman, I. 2005 - submitted
50. Zemla, A., Venclovas, C., Reinhardt, A., Fidelis, K., and Hubbard, T.J. 1997 Proteins Supp 1, 140.
51. Cozzetto, D., and Tramontano, A. 2005 Proteins Struct. Func. Bioinform. 58, 151.
52. Brylinski, M., Konieczny, L., and Roterman, I. 2004, In Silico Biol. 5, 0022.
53. Rost, B., and Sander, C. 1993, J. Mol. Biol. 232, 584.
54. Rost, B., Sander, C., and Schneider, R. 1994, J. Mol. Biol. 235, 13.
55. Zemla, A., Venclovas, C., Fidelis, K., and Rost, B. 1999, Proteins 34, 220.
56. Bryant, S.H., and Lawrence, C.E. 1993, Proteins Struct Func Gen 16, 92.
57. Zhang, B., Jaroszewski, L., Rychlewski, L., and Godzik, A. 1997 Folding and Design 2, 307.
58. Rost, B., Schneider, R., and Sander, C. 1997 J. Mol. Biol. 270, 1.
59. Miller, R.T., Jones, D.T., and Thornton, J.M. 1996, FASEB J. 10, 171.
60. Brylinski, M., Konieczny, L., and Roterman, I. 2005 VI European Symposium of the Protein Society 30 Apr – 4 May 2005 Barcelona, Abstr. 05-5.
61. Roterman, I., Brylinski, M., and Konieczny, L. 2005 VI European Symposium of the Protein Society 30 Apr – 4 May 2005 Barcelona, Abstr. 05-6.
62. Banzon, J.A., and Kelly, J.W. 1992, Protein Eng. 5, 113.
63. Carrel, R.W., Whisstock, J., and Lomas, D.A. 1994, Am J Cri Care Med 150, S171.
64. Fletterick, R.J., and Mc Grath, M.E. 1994, Nature Struct Biol 1, 201.
65. Katz, D.S., and Christianson, D.W. 1993, Prot Eng 6, 701.
66. Loebermann, H., Tokuoka, R., Deisenhofer, J., and Huber, R. 1984 J. Mol. Biol. 177 531.
67. Powell, L.M., and Pain, R.H. 1992, J. Mol. Biol. 224, 241.
68. Leluk, J., Konieczny, L., and Roterman, I. 2003 Bioinformatics 19, 117.
69. Guda, C., Scheeff, E.D., Bourne, P.E., and Shindyalow, I.N. 2001, Proceedings of Pacific Symposia on Biocomputing, 6, 251.
70. Cozzetto, D., and Tramontano, A. 2005, Proteins Struct Func Bioinfo. 58, 151.