Early-stage folding in proteins - *in silico* model.

Michal Brylinski[1,2], Leszek Konieczny[3], Irena Roterman[1*]

[1] Department of Bioinformatics and Telemedicine, Collegium Medicum – Jagiellonian University, Kopernika 17, 31-501 Cracow, Poland
[2] Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Cracow, Poland
[3] Institute of Biochemistry, Collegium Medicum – Jagiellonian University, Kopernika 7, 31-501 Cracow, Poland
* corresponding author: myroterm@cyf-kr.edu.pl

Introduction

The procedure according to which the deposited structures were created is based entirely on the model elaborated by team participants. This is why the assumptions and procedures used shall be presented before the prediction steps can be shown.
A significant number of proteins have been found experimentally to fold via a two-state process in which only the fully unfolded and native states are ever populated **[1,2]**. Our approach assumes two states in the protein folding process: early-stage and late-stage folding. The two-states are closely related with the idea of the reduced conformational space for early-stage protein folding. The complete conformational space is available after reaching the proper conformation determined in a first step of folding. The existence of the first step determines to some extent the search for the final native conformation.
The proteins deposited in CASP6 by our group were oriented on the first step of the assumed mechanism.

Conformational space represented by two geometric parameters: R (radius of curvature) dependent on V-angle (dihedral angle between two sequential peptide bond planes)

The early-stage folding was built on the characteristics of the polypeptide chain treated as a chain of rigid peptide planes, which can create shapes with different radii of curvature depending on the dihedral angles between two sequential peptide bond planes (V-angles) **[3-5]**. The assumption is that all polypeptide conformations can be treated as helicoidal. Beta-structure represents the helical structure characterised by an infinitely large radius of curvature. It turned out that the V-angle can change form 0 deg (parallel mutual orientation) to 180 deg (anti-parallel orientation) allowing creation of a very squeezed helix (V-angle close to zero); then an increase of the V-angle causes an increase of the R-radius of curvature, reaching R infinitely large for the Beta-structure and extended. Instead of the Phi, Psi conformational space, the V-R conformational space can be used for polypeptide structure representation. Using the above defined geometric parameters, the Ramachandran map can be interpreted as shown in Fig. 1. Fig.1A shows the distribution of the R-radius of curvature (on log scale) for polypeptide fragments (pentapeptides) all over the conformational space. The distribution of V-angles is shown in Fig.1B.

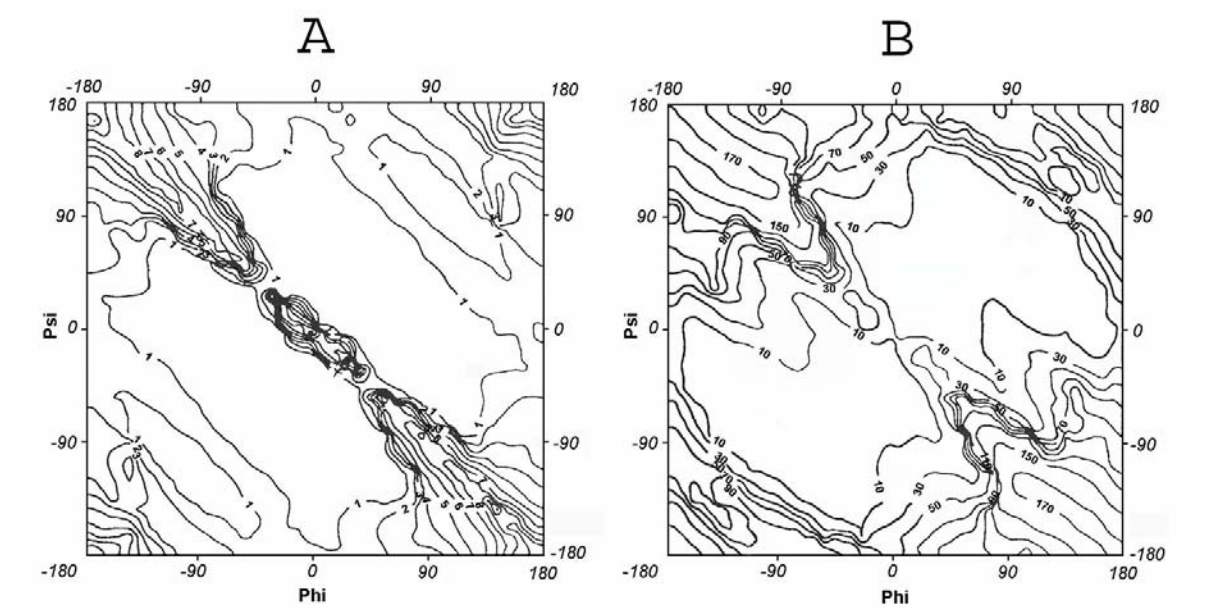

Fig. 1. Distribution of geometrical parameters all over the Ramachandran map.
B – radius of curvature (R) on natural logarithmic scale,
A – dihedral angle (V) between two sequential peptide bond planes.

When only low-energy conformations (Fig.2A) are taken into account, the dependency of R (log scale) on the V-angle can be approximated to a second-degree polynomial function (Fig.2B and eq.1.).

$$\ln(R) = 0.000340V^2 - 0.02009V + 0.848 \qquad (1)$$

On the other hand, the distribution of points that satisfy the presented dependency (Fig.2A) is represented as shown in Fig.2C. A convenient way to search for low-energy conformations seems to be the ellipse-shaped curve shown in Fig.2D. (and eq.2), which links all low-energy conformations very nicely (Fig2E.).

$$\begin{cases} Phi = -A\cos(t) - B\sin(t) \\ Psi = A\cos(t) - B\sin(t) \end{cases} \qquad (2)$$

In consequence, a geometric analysis of the polypeptide chain limited to simple R(V) representations prompts the limited conformational sub-space, which can be treated as the early-stage folding (*in silico*) conformational sub-space. This accords with the generally accepted assumption that the backbone conformation is responsible for the early-stage folding step.

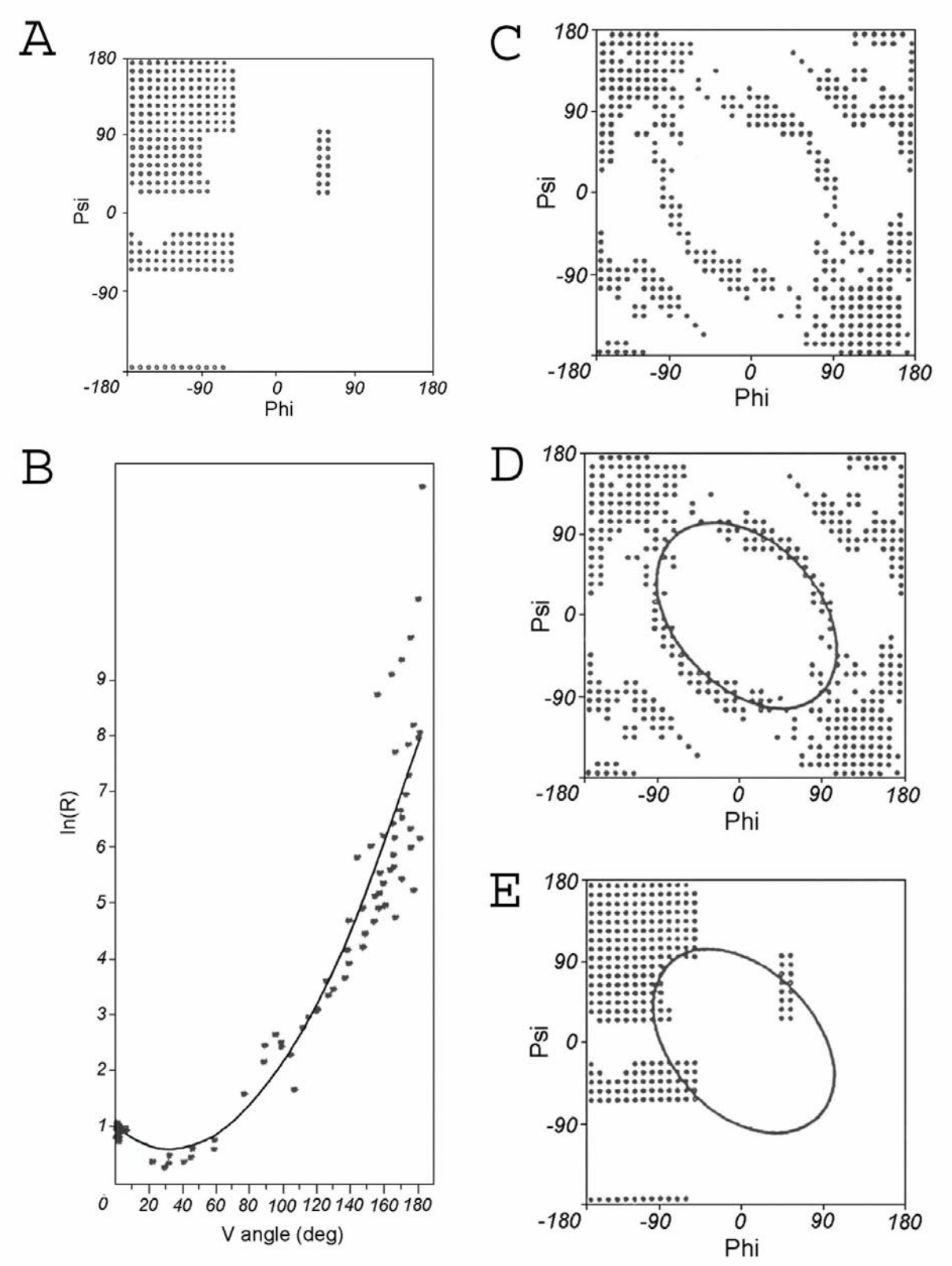

Fig.2. Ellipse path determination.
A – Phi, Psi map with low-energy area distinguished,
B – ln(R) as a function of V angle for grid points shown in A,
C – Ramachandran map with grid points, where the structure satisfies eq. 1,
D – low-energy areas linked by ellipse.

Limited conformational sub-space balances the amounts of information stored in the amino acid sequence and needed for structure prediction

Independent support for the model presented above came from elements of information theory **[7]**. The balance between the amount of information stored in the nucleotide (tri-nucleotide) appeared to be comparable with the amount of information necessary to determine the amino acid from among twenty of them. The same idea was applied for the relation between the amount of information carried by an amino acid and the amount of information necessary to predict a particular conformation determined by Phi, Psi angles (point on the Ramachandran map).

The amount of information [bit] carried by a particular amino acid can be calculated using Shannon's **[6]** equation:

$$I_i(p_i) = -\log_2 p_i \qquad (3)$$

where p_i expresses the probability of the *i*-th amino acid's presence in a sequence (in approximation p=1/20).

The amount of information necessary to predict a particular Phi, Psi angle can be calculated according to the same equation with *p* equal to the probability of the occurrence of these angles (which is much lower than 1/20). The entropy of information measuring the averaged amount of information necessary to predict structure with 1x1 deg precision can be calculated for each amino acid as follows:

$$SE_i = -\sum_{i=1}^{359*359} p_i \log_2 p_i \qquad (4)$$

where p_i represents the probability of occurrence of particular Phi, Psi angles (which can be calculated for PDB). These quantities were calculated (taking into account the different frequencies of particular amino acids and the probability distribution all over the Ramachandran map). Equation 4 applied to the ellipse-limited conformational sub-space leads to equilibration between these two quantities (Tab.1).

It can be proved that the large disproportion between these two events for the whole Ramachandran map makes the problem unsolved. Limitation of the conformational space (Ramachandran map) to the sub-space (for example the ellipse path) balances these two quantities. The amount of information carried by an amino acid (calculated according to its frequency) and the amount of information (entropy of information) necessary to predict the fragment of the ellipse (10 degs) becomes comparable (Tab.1).

Amino acid	Amount of information carried by amino acid	Amount of information necessary to predict the ellipse-belonging structure
	I_i [bit]	$SE_i^{\phi e \psi e}$ [bit]
GLY	3.805	5.740
ASP	4.117	5.016
LEU	3.492	4.437
LYS	3.908	4.764
ALA	3.662	4.462
SER	4.095	4.857
ASN	4.545	5.186
GLU	3.833	4.550
THR	4.196	4.579
ARG	4.249	4.650
VAL	3.886	4.108
GLN	4.663	4.667
ILE	4.151	4.115
PHE	4.713	4.528
TYR	4.941	4.574
PRO	4.442	4.062
HIS	5.477	4.868
CYS	5.544	4.792
MET	5.614	4.484
TRP	6.236	4.512

Tab. 1. The amount of information carried by an individual amino acid in relation to the amount of information necessary to predict the fragment of the ellipse (10 degs).

The relation between the distribution of Phi, Psi angles of GLN as was found in all proteins in PDB (January 2003 release) is shown in Fig.3.

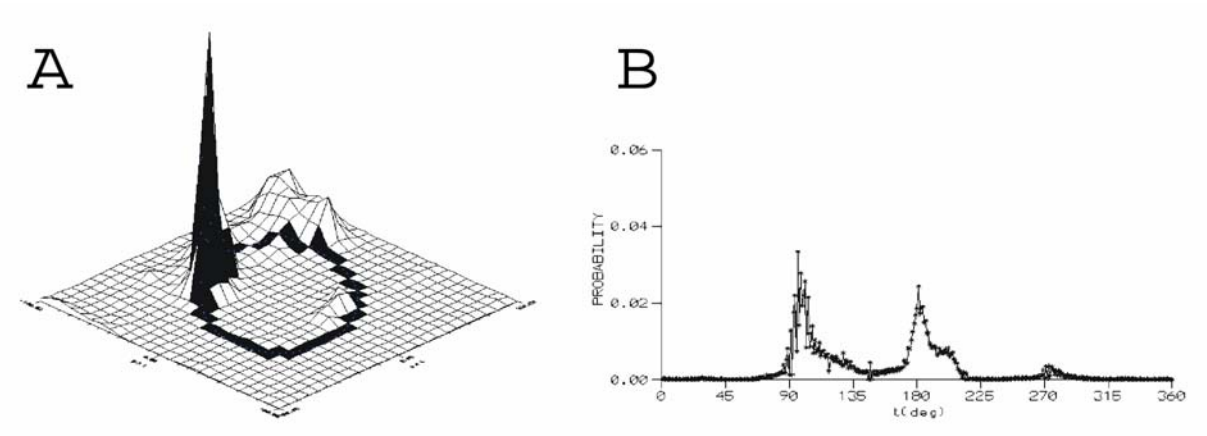

Fig.3. Phi, Psi angle distribution of GLN as found in the complete database of PDB (January 2003 release).
A – all over the Ramachandran map. Black line distinguishes the ellipse path.
B – after moving all Phi, Psi angles toward the ellipse path. The variable called t expresses the variable in the ellipse equation (eq. 2.). Zero value of t represents the point Phi = 90 deg, Psi = -90 deg and then increases clockwise along the ellipse.

The procedure allowing the calculation of probability distribution along the ellipse path is shown graphically in Fig.4A. The Phi, Psi angles present in proteins are transformed to Phi_e and Psi_e according to the criterion of shortest distance. The probability profiles of ten selected amino acids are shown in Fig.4B. The t-variable (eq.2.) from the 0 deg value (90, -90) and then clockwise to 360 degs changes as shown in Fig.4C. Seven probability maxima can be distinguished in the probability profile after limitation of the conformational space to the ellipse-shaped path. These seven probability maxima can be identified as structural motifs (letter codes shown in Fig.4B,C.). Two different notations were introduced to distinguish one-letter codes in– bold for sequence one-letter codes in italics for structure (probability maxima).

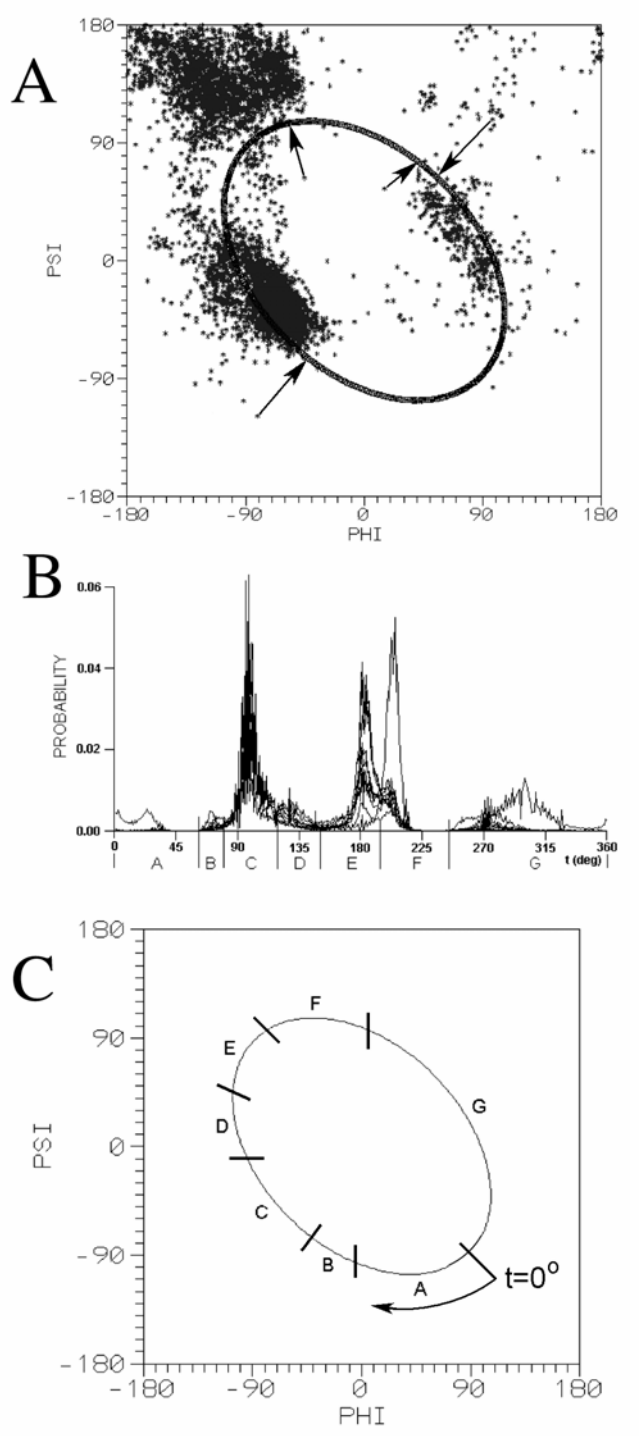

Fig. 4. Letter codes for structure classification.
A – the ellipse path-limited conformational sub-space in relation to Phi, Psi angles as they appear in real proteins. Arrows denote the shortest distance criterion for definition of Phi_e and Psi_e angles belonging to the ellipse for arbitrary selected points.
B – probability maxima as they appear along the ellipse (starting t-point shown in C) and corresponding letter codes for structure identification.
C – limited conformational sub-space with fragments distinguished according to probability maxima shown in B.

Sequence-to-structure contingency table

The tetrapeptide was selected as the shortest polypeptide fragment with a well-defined structural motif (Beta-turn, alpha-helix, Beta-structure). The four-letter string expressing sequence and four-letter string representing structure (ellipse-limited conformational space) were used to analyse the sequence-to-structure relation in early-stage folding. A contingency table of 160 000 possible tetrapeptide sequences (rows - 146 940 found in real proteins) and 2401 possible tetrapeptide structures (columns – 2397 found in real proteins) representing the relation between these two characteristics was created (Tab.2). A four-amino-acids Open Reading Frame (as in nucleotide sequence analysis) in an overlapping system (in contrast to nucleotide sequence analysis) was applied to find low- and high-probability values for each cell in the contingency table. Two different methods were applied to analyse a table of such large size: information theory **[12]** and statistics-based **[13]**.

Sequence ⇒ Structure ⇓	…	**AAAA**	**GDSG**	**AVRR**	**EAEL**	…
…	…	…	…	…	…	…
GCFG	…	p *GCFG* /**AAAA**	p *GCFG* / **GDSG**	p *GCFG* /**AVRR**	p *GCFG* /**EAEL**	…
AEED	…	p *AEED* / **AAAA**	p *AEED* / **GDSG**	p *AEED* /**AVRR**	p *AEED* /**EAEL**	…
BFBE	…	p *BFBE* / **AAAA**	p *BFBE* / **GDSG**	p *BFBE* /**AVRR**	p *BFBE* /**EAEL**	…
AFFB	…	p *AFFB* / **AAAA**	p *AFFB* /**GDSG**	p *AFFB* /**AVRR**	p *AFFB* /**EAEL**	…
BACE	…	p *BACE* / **AAAA**	p *BACE*/ **GDSG**	p *BACE* /**AVRR**	p *BACE* /**EAEL**	…
…	…	…	…	…	…	…

Tab. 2. Fragment of the contingency table representing the sequence-to-structure relation for tetrapeptides using the coding systems in Fig. 4B,C. (sequence – bold, structure-italics).

SPI – Structure Predictability Index

Tetrapeptides determine their structure to different degrees. There were sequences found with high and low determinability. Thus it is possible to measure the degree of predictive power of particular polypeptide fragments. In consequence, the Structure Predictability Index (SPI) is introduced for quantitative estimation of structural stability. The SPI coefficient can be calculated as follows: the sequence of each target protein has been read using a four-amino-acid frame, in four possible ways (overlapped reading). For each read fragment of the target sequence, the ellipse-limited structure from the database was chosen using the criterion of the highest number of polypeptide chains belonging to a particular structure and representing a particular sequence divided by the number of Open Reading Frames (Tab.2: p *XXXX*/**XXXX**). Each amino acid in the resulting structure achieved the state with the highest p *XXXX*/**XXXX** value. In addition, the mean value for all amino acid positions can be calculated (SPI). SPI reaches values from 0.14 to 1.00, where 0.14 means completely random prediction. The SPI scale was standarized to the 0-1 interval.

Procedure applied

The procedure of protein structure prediction was performed according to the following steps:

1. high probability structures for tetrapetide sequence pieces were found in the contingency table using the tetrapeptide as the unit in an Open Reading Frame system (in analogy to nucleotide sequences) in overlapped form (in contrast to nucleotide sequence analysis) (Fig.5B).

2. the structure of highest probability for four structural motifs for each amino acid was selected (Fig.5C).

3. *C*, *E*, *F* ellipse fragments (Fig.4C) were assumed to form ordered local structure (*C* – helix and *E*, *F* – extended), while *A*, *B*, *D*, *G* are responsible for protein disorder. SPI was transformed to reach values from 0.0 (the most probability of disorder) to 1.0 (the most probability to form ordered structure) for each amino acid position. Predictions in the category Order-Disorder were submitted for all targets.

4. 3D all-atom models were created for targets shorter than 150 amino acids, based on 1D predictions.

5. the coding system of structural motifs was transformed to Phi, Psi representations finding the ellipse-belonging Phi_e and Psi_e angles according to the position of the probability maximum in a particular code.

6. the early stage folding structure was created based on the set of Phi_e and Psi_e angles.

7. the energy minimization procedure was performed using the ECEPP/3 program. The constraints for SS-bonds were introduced in the case of the presence of disulphide bonds declared in the target profile. For more than two residues able to form a disulphide bridge, every possible combination of them was taken into consideration. Postminimization structures with the lowest energy according to ECEPP/3 standards were submitted as 3D atomic coordinates (tertiary structure) prediction (Fig.5D).

8. pair-wise contacts were calculated for all 3D-structures using an 8.0 Å cutoff. Probabilities (based on SPI) of particular contacts for targets with previously predicted 3D structure were submitted in the category of Residue-Residue separation distance prediction (Fig.5E).

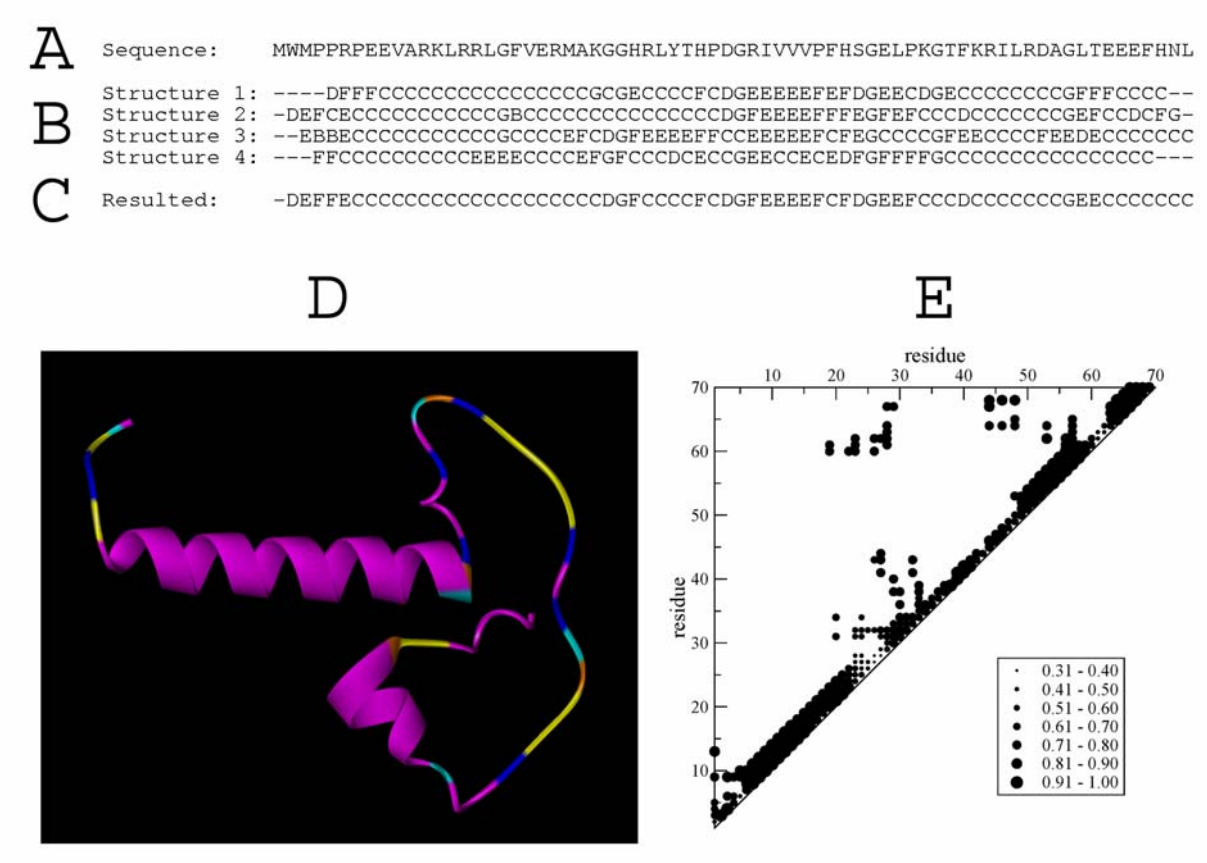

Fig. 5. Early-stage structure prediction for CASP6 target t0281.
A – the sequence of the target,
B – the attributed structure for four different reading frames (coding system as in Fig.4B,C),
C – resulted 1D structure,
D – 3D all-atom models based on 1D prediction. Phi_e and Psi_e angles were found according to the position of the probability maximum in a particular code (Fig.4B). Color notation distinguishes particular structure form as follow: violet – *C*, light blue – *D*, yellow – *E*, blue – *F*, orange – *G*,
E – contacts map calculated for 3D-structure. The size of a circle corresponds to the probability of particular contact.

Conclusions

The goal for the participants group was estimation of the early-stage folding structures prediction. No low RMS-D was expected. The energy minimization procedure applied to the assumed early-stage folding conformations delivered structural forms of targets. Estimation and analysis of comparison with the native structure of the target proteins will inform further development of the method.

The details concerning the geometric basis of the model can be found in **[3-5]** and the information theory-based model in **[7]**. The model was verified by structure prediction of selected proteins: BPTI **[8]**, hemoglobin **[9]**, lysozyme **[10]**, ribonuclease **[7]** and the serpine family **[11]**. The serpine family was also analysed in the context of biological function recognition based on the presented model.

References

1. Jackson SE, Fersht AR. (1991). Folding of chymotrypsin Inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* **30**, 10428-10435.
2. Akmal A, Muñoz V. (2004). The nature of the free energy barriers to two-state folding. *Proteins* **57**, 142-152.
3. Roterman I. (1995). Modelling of optimal simulation path in the peptide chain folding – Studies based on geometry of alanine heptapeptide. *J Theoretical Biol* **177**, 283-288.
4. Roterman I. (1995). The geometrical analysis of polypeptide backbone structure and its deformation. *Biochimie* **77**, 204-216.
5. Roterman I, Konieczny L. (1995). Geometrical analysis of structural changes in immunoglobin domains' transition from native to molten state. *Computers and Chemistry* **19**, 204-216.
6. Shannon CEA. (1948). Mathematical theory of communication. *Bell. Syst. Tech. J.* **27**, 379-423.
7. Jurkowski W, Brylinski M, Konieczny L, Wisniowski Z, Roterman I. (2004). The Conformational Sub-Space in Simulation of Early-Stage Protein Folding. *Proteins* **55**, 115-127.
8. Brylinski M, Jurkowski W, Konieczny L, Roterman I. (2004). Limited conformational space for early stage protein folding simulation. *Bioinformatics* **20**, 199-205.
9. Brylinski M, Jurkowski W, Konieczny L, Roterman I. (2004). Limitation of conformational space for proteins – early stage folding simulation of human α and β hemoglobin chains. *TASK Quaterly* **8**, 413-422.
10. Jurkowski W, Brylinski M, Konieczny L, Roterman I. (2004). Lysozyme folded *in silico* according to the limited conformational sub-space. *J. Biomol. Struct. Dynam.* **22**, 149-158.
11. Leluk J, Konieczny L, Roterman I. (2003) Search for structural similarity in proteins *Bioinformatics* **19**, 117-124.
12. Brylinski M, Konieczny L, Czerwonko P, Jurkowski W, Roterman I. Early-stage folding in proteins (*in silico*) – sequence to structure relation. (submitted).
13. Meus J, Brylinski M, Piwowar M, Piwowar P, Wiśniowski Z, Stefaniak J, Konieczny L, Roterman I. A tabular approach to the sequence-to-structure relation in proteins. (submitted).