© Med Sci Monit, 2006; 12(6): BR208-214
PMID: 16733478

Received: 2005.07.21
Accepted: 2006.03.06
Published: 2006.06.01

A tabular approach to the sequence-to-structure relation in proteins (tetrapeptide representation) for *de novo* protein design

Authors' Contribution:
A Study Design
B Data Collection
C Statistical Analysis
D Data Interpretation
E Manuscript Preparation
F Literature Search
G Funds Collection

Jan Meus[1][G], Michał Brylinski[1,2][B], Monika Piwowar[1][C], Piotr Piwowar[3][B], Zdzisław Wiśniowski[1][E], Justyna Stefaniak[4][C], Leszek Konieczny[5][D], Grzegorz Surówka[6][B], Irena Roterman[1,6][AD]

[1] Department of Bioinformatics and Telemedicine, Collegium Medicum, Jagiellonian University, Cracow, Poland
[2] Faculty of Chemistry, Jagiellonian University, Cracow, Poland
[3] Department of Measurement and Instrumentation, AGH – University of Science and Technology, Cracow, Poland
[4] Institute of Mathematics, Jagiellonian University, Cracow, Poland
[5] Institute of Medical Biochemistry, Collegium Medicum, Jagiellonian University, Cracow, Poland
[6] Department of Information Technologies, Institute of Physics, Jagiellonian University, Cracow, Poland

Source of support: This study was financially supported by a Collegium Medicum grant (501/P/133/L)

Summary

Background: Experimental observations classify the protein-folding process as a multi-step event. The backbone conformation has been experimentally recognized as responsible for the early-stage structural forms of a polypeptide. The sequence-to-structure and structure-to-sequence relation is critical for predicting protein structure. A contingency table representing this relation for tetrapeptides in their early-stage is presented. Their correlation seems to be essential in protein-folding simulation.

Material/Methods: The polypeptide chains of all the proteins in the Protein Data Bank were transformed into their early-stage structural forms. The tetrapeptide was selected as the structural unit. Tetrapetide sequences and structures were expressed by letter codes. The transformation of a contingency table of any size (here: 160,000×2401) to a 2×2 table performed for each non-zero cell of the original table allowed calculation of the ρ-coefficient measuring the strength of the relation.

Results: High values of the ρ-coefficient extracted sequences of strong structural determinability and structures of high sequence selectivity. The web-site program to calculate the ρ-coefficient ranking list was constructed to enable applying this method to any problem of contingency table analysis.

Conclusions: The results revealed sequence-to-structure (and vice versa) correlation in early-stage folding. Surprisingly, the irregular structural forms of loops and bends appeared to be highly determined. Comparison of these results with another method based on information entropy revealed high accordance. The method oriented on interpretation of a large contingency table seems very useful especially for large-scale microarray analysis, a very popular technique in the post-genomic era.

key words: **protein structure • contingency table • ρ-coefficient**

Full-text PDF: http://www.medscimonit.com/fulltxt.php?IDMAN=7833

Word count: 3209
Tables: 3
Figures: 2
References: 44

Author's address: Irena Roterman, Department of Bioinformatics and Telemedicine, Collegium Medicum, Jagiellonian University, Kopernika 17, 31-501 Cracow, Poland

Background

The dependence of 3-D structure on the amino-acid sequence in a polypeptide chain is a basic dogma in biological science. The classic table [1–4] linked the predisposition of a particular amino acid to a particular structural form [5]. Since that time, the data base of protein structures, the Protein Data Bank (PDB) [6], has expanded, reaching more than 20,000 examples. This provides a good basis for a new approach to this problem. The comparative modeling category appeared as the most represented in the CASP5 competition (29 CM targets, and 51 including the fold-recognition category CM+FR) [7]. CASP5 also revealed that comparative modeling is a powerful tool delivering the top best predictions [8–13]. A new approach to understanding the sequence-to-structure relation, based on coupling effects leading to self-consistency of conditional probability, contrasted with the traditional notion of amino-acid composition influencing structure [14]. Examples of seven-amino-acid-long fragments of identical amino-acid sequences representing completely different structural forms can be found in the PDB [15]. These examples found in native structures additionally show that side chain-side chain interaction influences the backbone conformation. The search for common patterns in the early-stage folding conformation of the backbone may help to explain the differences in final, native structural forms.

A search for common structural motifs was recently performed on a large genomic scale [16]. The structural features of important function-related peptide sequences, found to be invariant across diverse bacterial genera, revealed the preferred conformations in different protein structures. The variability of conformations was registered as due to the flipping of peptide units about the virtual Cα-Cα bonds. These short, invariant polypeptide fragments of structural importance were even proposed to be treated as structural determinants. Only three structural forms, H (helix, R and L), B (β-structure), and U (uncharacterized), were distinguished in the cited work. The codes introduced in the present study distinguish seven structural forms, making the U category more recognizable. The search for the sequence-to-structure (and structure-to-sequence) relation is the main subject of this paper.

The commonly accepted interpretation of polypeptide chain folding treats optimization of the backbone conformation as the driving force for early-stage folding [17]. Side chain-side chain interactions appear later, and optimization of this interaction also influences the backbone conformation. This is why the backbone conformation becomes more or less obliterated (compared with its relaxed form) in the final conformation of the protein (the high-energy phi and psi angles present on the Ramachandran map). The conformation of the polypeptide can be called early-stage as long as the backbone represents its optimal structure.

The backbone conformation can be perfectly well described using the phi and psi angles. However, a simplified model can be introduced to describe the backbone conformation. The V-angle, expressing the dihedral angle between two sequential peptide bond planes (rotation around the Cα-Cα virtual bond), appeared to determine the local radius of curvature of the polypeptide chain. The analysis was done for pentapeptides. The structure of the pentapeptide was created for each grid point on the Ramachandran map (5- and 10-degree steps, all five amino acids represented selected phi and psi angles). The relation between the V-angle and the R-radius of curvature for low-energy structures appeared to have the form of a second degree polynomial. This analysis addressed the relation between the V-angle and the R-radius of curvature. Taking the function expressing this relation, where are the backbone conformations satisfying the relation found? When all the structures (grid points) satisfying the functional relations found were taken into analysis, an ellipse-shaped path appeared on the Ramachandran map. The conformations created according to this limited conformational sub-space are assumed to represent early-stage folding. The basis for this model was described in detail in [18,19].

When all phi and psi angles (as they appear in real proteins in the PDB) are shifted (according to the criterion of the shortest distance) toward the ellipse path, the probability distribution of phi and psi angles belonging to the limited conformational sub-space can be found. A discussion of the omega dihedral angle in omitted in the analysis. It plays an important role in the case of proline. An omega angle other than 180° seems to be the result of a late-stage folding step as it is driven by the side chain-side chain interaction. This is why the influence of *cis* conformation has not been taken into consideration. The probability profile calculated separately for each amino acid revealed some preferences expressed by characteristic probability maxima distributions. The location of the probability maxima on the ellipse path allowed partitioning the ellipse path so that letter codes could be introduced for each probability maximum. Each probability maximum gave a structural code. These seven structural codes (A-G) are used to identify the early-stage structural form of a particular tetrapeptide in the polypeptide chain. The tetrapeptide was selected as the shortest polypeptide chain fragment representing a well-defined structural motif. The structures of four amino acids in the polypeptide can be recognized as ordered structural forms. The helix can be recognized in the 3.6-amino-acid-long polypeptide. The beta-turn can be recognized on the basis of a four-amino-acid-long polypeptide. The beta-structural form can also be recognized for longer polypeptides (the phi and psi of an isolated amino acid or of a dipeptide cannot be classified as an ordered form). This is why the tetrapeptide was selected as the unit for both structure and sequence classification. This coding system enabled the creation of a contingency table expressing the sequence-to-structure (and structure-to-sequence) relation. A table of (potentially) 160,000×2,401 cells was created according to traditional sequence codes and the introduced structural codes and analyzed using the correlation calculus applied to the modified table (described in Materials and Methods). Methods for analyzing the relation between qualitative variables are available [20–25]. The ρ^2 coefficient calculated for the contingency table reduced it (according to the particular procedure) to table a 2×2 table, which seemed to satisfy the expectations of a measure of the relation between the two qualitative variables (sequence-to-structure or structure-to-sequence).

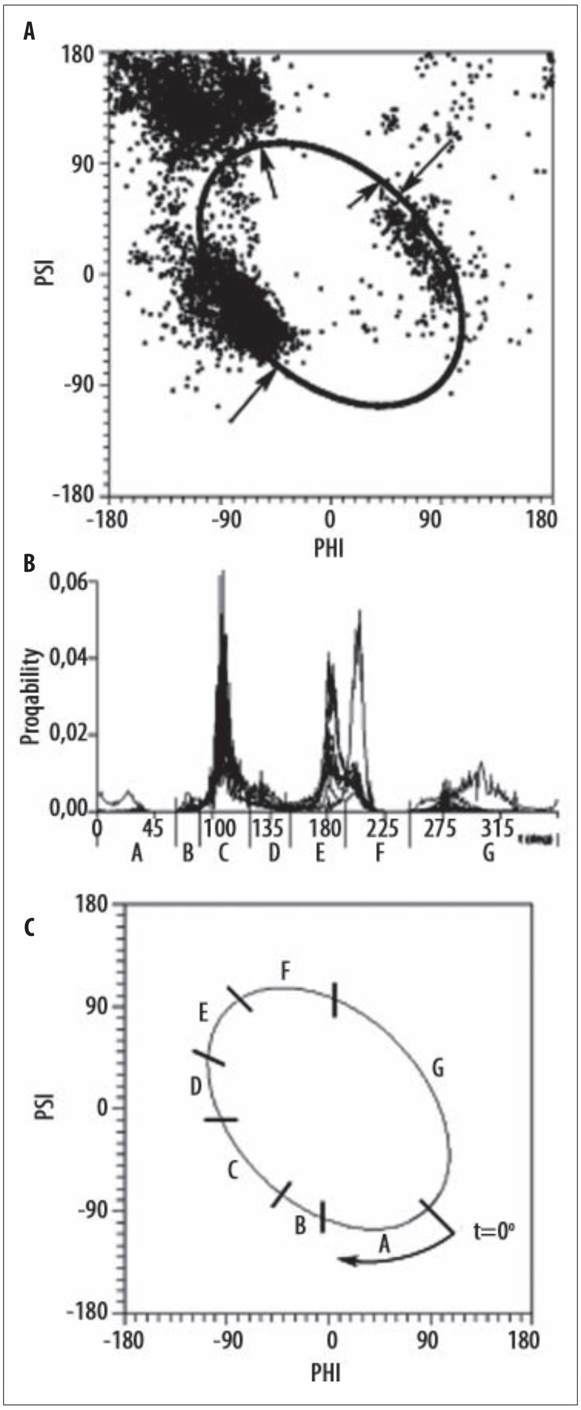

Figure 1. The ellipse path representing the limited conformational sub-space for early-stage protein folding simulation (the same probability profile was used to classify the structural motifs in [40,41]). (**A**) – the ellipse path in relation to the phi, psi angle distribution as it appears in real proteins, (**B**) – shortest-distance criterion for definition of phi, psi angles belonging to the conformational sub-space, (**C**) – The probability of a phi, psi angle distribution as it appears in real proteins after moving its phi, psi angles to the nearest point on the ellipse. The probability maxima were taken to distinguish the structural forms. The t-angle represents the angular variable in the parametric equation of the ellipse. At t=0, phi=90° and psi=–90°, and then increases clockwise along the ellipse (Figure 1B).

MATERIAL AND METHODS

Structure coding system for proteins

The tetrapeptide was taken as the unit to represent both sequence and structure in the polypeptide chain (the shortest polypeptide fragment of possible good recognition of an ordered structural form, such as an alpha-helix, beta-structure, and beta-turn). Each amino acid was represented by a one-letter code of a common coding system for amino acids. Structure was also described by a one-letter code. The basis for this coding system was the early-stage folding model presented in detail in [21,26,18]. Each amino acid was represented by a characteristic and a specific probability distribution on the Ramachandram map. This characteristic dispersion of phi and psi angles all over the Ramachandran map, transformed to the limited conformational sub-space and presented in the form of a profile, allowed for the definition of seven probability maxima. Each maximum, distinguished by a letter code, allows a unified classification of the structural form (Figure 1).

The limitation of conformational space to the ellipse-path conformational sub-space resulted from geometrical analysis of the backbone (the dihedral angle between two sequential peptide bond planes) and information theory analysis. The ellipse path satisfies the condition of balancing the amount of information carried by an amino-acid sequence and the amount of information necessary to predict particular phi and psi angles for the early-stage structural form (Figure 1A) [18]. The shortest-distance criterion applied to transform the native phi and psi angles to the early-stage form (Figure 1B) produces the probability profile shown in Figure 1C. Although the profiles are amino acid dependent, all of them can be characterized by seven probability maxima, as shown in Figure 1C. Each probability maximum can be letter-coded according to the system presented in Figure 1C. Each protein is thus characterized by a one-letter code representing the early-stage structural form of the polypeptide under consideration [40,41]. A window size of four amino acids was applied, similar to the Open Reading Frame in nucleotide sequence analysis, with the difference expressed by the overlapped ORF system applied in our model.

Contingency table construction and analysis

Since there are 160,000 different tetrapeptide sequences possible in the protein universe and 2,401 different tetrapeptide structures after adopting the model presented above, a table of that size was calculated, taking the complete PDB (2003# release) to represent the number of cases representing a particular sequence and its particular structure. The complete (and upgraded) table is available at *http://bioinformatics.cm-uj.krakow.pl/zcoeff/*.

It is impossible to analyze a complete table as large as this. All methods measuring the dependency between variables treat the table as a complete unit. The result of applying any available method may give the general characteristics identifying the presence or absence of a correlation between two variables. Information about the presence of a mutual relation does not say much in our case. A contingency table (Table 1) of any size can be transformed into a 2×2 table

Table 1. The scheme of the contingency table expressing the distribution of sequence-to-structure probability in proteins.

Sequences → Structures ↓	A1	A2	A3	....	A_i	....	A_c
B_1	P_{11}	P_{21}	P_{31}	...	P_{i1}	...	P_{c1}
B_2	P_{12}	P_{22}	P_{32}	...	P_{i2}	...	P_{c2}
B_3	P_{13}	P_{23}	P_{33}	...	P_{i3}	...	P_{c3}
.....	...	...	...	...	...	...	...
B_j	P_{1j}	P_{2j}	P_{3j}	...	$\boxed{P_{ij}}$	...	P_{cj}
......	...	...	...	...	...	...	...
B_r	P_{1r}	...	P_{2i}	...	P_{ir}	...	P_{cr}

Table 2. 2×2 contingency table resulting from reducing the dimensionality of Table 1.

Variables	$A_{i,j}$	A_{ik} k=1,...,c and k=j excluded
B*j,i*	$P(A_i \mid B_j)$	$P(\overline{A}_i \mid B_j)$
B*jn* n=1,..r and n=i excluded	$P(A_i \mid \overline{B}_j)$	$P(\overline{A}_i \mid \overline{B}_j)$

(Table 2) according to the following. Assume that the contingency table represents the observed (empirical) probabilities for *c* different realizations of variable **A** (qualitative) and *r* different realizations of variable **B** (qualitative). For the problem presented in this paper, assume that **A** represents sequences (tetrapeptides) and **B** structures (tetrapeptides) (Table 1). To estimate whether p_{ij} expresses high or low probability, i.e. high or low coupling of a particular sequence (*i*) with a particular structure (*j*), its value is compared with all possibilities for solutions of other **A** (excluding the *i*-th) and other **B** (excluding the *j*-th). Each pair of *i*-th and *j*-th realizations of **A** and **B** can be represented using a 2×2 contingency table (Table 2). The coefficient ρ_{ij} [27] measuring the correlation between two variables can be calculated for each of the reduced contingency tables.

$$\rho^2(A{:}B) = \frac{[P(A \cap B) - P(A)\cdot P(B)]^2}{[\mathrm{P}(A)\cdot \mathrm{P}(B)\cdot \mathrm{P}(\overline{A})\cdot \mathrm{P}(\overline{B})]}$$

The ρ_{ij}^2 coefficient measuring the correlation for table $\mathbf{A}_i\mathbf{B}_j$ expresses the partial strength of mutual dependency. Known ρ values for each individual cell in the ranking list give insight into the relative participation in the global correlation between the two variables under consideration. The higher the ρ^2 value, the higher the relative power of a particular *ij* pair's participation in the general correlation between the two variables under consideration.

Presentation of selected structures

Tetrapeptides found to represent the highest structure-sequence determinability (estimated on the basis of the coefficient) were created as follows:

The phi and psi values found for the probability maxima (according to Figure 1C) were used to create a blocked tetrapeptide (ACE-X_1-X_2-X_3-X_4-MNE) structure (standard parameters according to the ECEPP program). X represents the amino acid under consideration. The energy minimization procedure was applied to each blocked tetrapeptide to remove possible overlapping of side chains, with the ECEPP force field adopted [28,29] with constraints on the phi and psi angles. To emphasize the mutual spatial orientation of the terminal amino acids, short (tetrapeptide) fragments of polyalanine representing the extended form (phi, psi angles equal to 180°) were attached, substituting the terminal (ACE and NME) groups.

Results

Contingency table analysis

The size of the contingency table in real proteins as found in the PDB (2003# release) appeared to be 146,940 columns (tetrapeptide sequences, potentially 160,000) × 2397 rows (tetrapeptide structures, potentially 2401). The values expressing the probability of the presence of a particular form were very low, so log-values are given (complete table available on request). The value of the ρ-coefficient was calculated for each cell of the contingency table. Since the calculation was performed to reveal cases with a strong relation of sequence-to-structure (and structure-to-sequence), the highest ρ values are listed in Table 3. Different fonts were used to distinguish the letter codes for sequence (**bold**) from those for structure (*italics*). Only the top thirty values are shown in Table 3 because the complete list is too large.

Structures of relatively high determinability

The top ten structures, representing relatively high determinability, are shown in Figure 2. The overlapping sequence **NGGDD** was found, although the structure does not represent the continuous form *DBAA* **(NGGD)**/*AGCE* **(GGDD)**. The same contingency table was independently

Table 3. Ranking list of the top twenty ρ-coefficient values attributed to cells representing a particular relation of structure (in italics) to sequence (in bold). The numbers in the right column represent the position on the analogous list of structure/sequence analysis based on information theory. Position 1;2 denotes the first position in the structure-to-sequence analysis, the second the position on the ranking list of sequence--to-structure position.

Z coefficient *E-001	Structure	Sequence	Position number in alternate method
7.45	*AEGD*	**GNES**	7
6.92	*BEBF*	**ERSY**	
6.58	*AFFB*	**GFRN**	10
6.57	*AEBB*	**GPVY**	
6.18	*AEGE*	**GIGH**	
5.78	*ABFF*	**GPHF**	
5.71	*EAAD*	**KGGP**	
5.49	*GAGD*	**FNAG**	5
5.34	*DBAA*	**NGGD**	
5.19	*GCFG*	**GDSG**	
5.16	*DDBG*	**SHHG**	
5.10	*BBDF*	**KTRS**	
5.09	*GDAE*	**ESGH**	1;2
5.03	*DBEB*	**AERS**	
5.02	*BACE*	**GGAE**	
4.95	*CABF*	**AGPH**	
4.95	*AEED*	**GLRL**	
4.91	*DAFA*	**DGPG**	3
4.91	*AEDE*	**GIFR**	
4.88	*AGAE*	**IKYG**	2
4.85	*CFAG*	**NTGG**	
4.79	*EBCB*	**ELPD**	
4.77	*CADD*	**RGRC**	
4.67	*GGED*	**KGHH**	
4.66	*AGCE*	**GGDD**	8
4.57	*DFDA*	**YNPV**	
4.56	*BFBE*	**PEPV**	
4.54	*GFDD*	**GQTN**	
4.51	*FAEA*	**PGFG**	
4.47	*AEGF*	**GCAQ**	6
4.47	*ADFA*	**GTQC**	

analyzed using an information theory approach. High accordance was found between the two kinds of results. The top ten sequence-to-structure and top ten structure-to-sequence highly correlated pairs found using information theory were among the top thirty positions on the ranking list produced by the presented approach. To make comparative analysis possible with the results presented in [41], the results of analysis of the same data base were presented; the upgraded results are available on *http://bioinformatics.cm-uj.krakow.pl/zcoeff/*.

Discussion

Analysis of Figure 2, representing structures found to be highly determined, delivers a very important conclusion. Ordered structural forms and turns are the elements that protein prediction specialists usually look for [30]. Table 3 points out fragments representing different forms of turns without any regular or ordered parts. This is why we treat the contingency table and particularly the coefficient ranking list as sources of information about strategic points for the folding process. This is significant for the creation of a starting structure based on the ellipse-limited conformational sub-space. It is very easy to use the contingency table described in this paper to create these structures. The ellipse-limited conformational sub-space was used and tested for ribonuclease and BPTI to create the starting structures applied to the energy minimization procedure. The good agreement between the structures obtained in this way with the native structures of these proteins allows the contingency table to be used to create a starting structure. The structures obtained according to our analysis are confronted with the structures found on the basis of bacterial genome analysis [31]. The results can be treated as qualitatively consistent, since both methods, applied independently, selected rather irregular, unordered bend structural forms, which seem to influence the further propagation of the chain significantly.

The CASP competition evaluates progress in protein structure prediction every two years [32,33,34]. Despite the long history of this discipline, the results are far from satisfactory or certain [35]. A high-confidence method is very needed. We believe that our analysis delivers data that can be treated as a library for starting structure determination. The model for early-stage folding *in silico* has been verified for ribo-

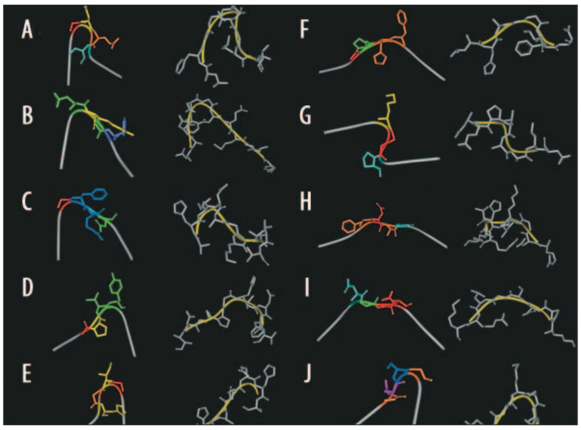

Figure 2. Ten top structures of tetrapeptides with the highest structure-to-sequence determinability according to Table 3 (ten highest values of ρ coefficient). Color notation distinguishes particular structural forms as follows: red – A, green – B, purple – C, light blue – D, yellow – E, blue – F, orange – G. Gray terminal fragments represent the extended form of polyalanine (tetrapeptides) to emphasize the mutual spatial orientation of the terminal fragments. The data for creating these structures are given in Table 3 and Figure 1.

nuclease [18], BPTI [36], lysozyme [37], and the alpha and beta chains of hemoglobin [38]. That is why the complete protein data base was represented on the basis of this model. The approach presented in this paper may be very useful to those involved in *de novo* protein design [39].

The contingency table expressing the sequence-to-structure and structure-to-sequence relation was analyzed using another approach based on information theory [37]. The entropy of information was calculated for each tetrapeptide expressing the sequence and expressing the structural motif defining the early-stage folding conformation. The cells in the contingency table express the probability of a particular sequence-to-structure relation. It was found that some sequences represent a different level of entropy (assumed to measure the uncertainty in predicting the structure for a particular sequence). Those of low entropy can be treated as easy to predict; those of high entropy represent sequence fragments with high difficulty of structure prediction. Analyzing the predictability of a particular amino-acid sequence, the degree of difficulty of structure prediction can be measured. A scale quantitatively expressing this was elaborated and presented in [40]. The advantage of the SPI (Structure Prediction Index) is that the difficulty can be measured in an *a priori* system without knowledge of the structure of the protein under consideration. This SPI index can be calculated for any amino-acid sequence using the interface at *http://bioinformatics.cm-uj.krakow.pl/zcoeff/*. Another approach to analysis of the contingency table shown in this paper was used to analyze the relation of sequence to structure (and vice versa), based on informational entropy calculation [41]. High accordance was found between the results of the two models (see Table 3).

Comparison of the results with those of other tools commonly used for comparative modeling [42,43] is difficult because of the different bases on which the particular models were constructed. The early-stage conformation is the background for our analysis, while the final, native conformation of proteins is used in other methods. The procedure for reducing the contingency table seems to be universal and applicable to any table of large size. This approach should prove useful particularly nowadays, when tools for the analysis of large data bases in microarray form are anxiously awaited [44].

Conclusions

The early-stage model has been verified using BPTI [36], lysozyme [37], and ribonuclease [18]. The verification excluded any forbidden structural forms in these proteins (for example knots). The approach to the native structural forms after application of an energy minimization procedure or molecular dynamics simulation seems to be satisfactory. This observation allowed the generalization of sequence-to-structure and structure-to-sequence in the form of a contingency table. The availability of this contingency makes wider verification possible. The contingency table may be applied to any protein under consideration. The contingency table was applied for the creation of all targets in the CASP6 competition (*http://predictioncenter.org/casp6/Casp6.html*). The aim of participating in CASP6 was to verify the early-stage structures in blind prediction. The best approach appeared quite satisfactory, although only the early-stage model was applied.

Application of the presented model (the sequence-to-structure contingency table) to targets in CASP6 revealed the bias towards helical structures. The conclusion taken from this observation is that the beta-structural forms (letter codes E and F, Figure 1C) shall be taken into account even when they are not the majority in the overlapping system. Four structural sequences are given for each polypeptide sequence. The overlapping system for the "reading frame" (as in the nucleotide reading frame definition) was applied to take into account the influence of short-range residues. There are four possible structure attributions to a tetrapeptide sequence. The most frequent structural letter code decides on the final form of the early-stage definition of structure with the exception of the E and F (beta-structural) forms for the reasons shown above.

The early-stage structure examination is not available for any experimental observations. The only way to verify the model of early-stage folding (in silico) is the simulation of a complete folding process, which will be presented in the near future.

The most important conclusion from the analysis of the contingency table is the observation shown in Figure 2. The sequence for *de novo* creation of proteins, particularly those of expected structure, can be easily predicted on the basis of the presented contingency table. The irregular loop-like and turn-like structural forms shown in Figure 2, selected according to high values of ρ coefficient, seem to be promising.

The presented analysis of the relation between sequence and structure in early-stage folding made possible quantitative measurement of the mutual relation between these two very important variables describing the folding process. The probability values present in particular cells of the table appeared to be useful for a priori estimation of the difficulty in structure prediction in an a priori system, without knowledge of the protein native structure. The loops extracted by high ρ-coefficient values seem to be very important as strategic points for polypeptide chain folding. The regular, ordered structural forms seem to be driven by entropy effects, while bends and loops appeared highly determined. This is important because loops and bends play critical roles in determining the general features of a polypeptide chain.

Acknowledgements

Many thanks to Professor Marek Pawlikowski of the Faculty of Chemistry of the Jagiellonian University for fruitful discussions.

References:

1. Chou PY, Fasman GD: Conformational parameters for amino acids in helical, Beta-sheet and random coil region calculated from proteins. Biochemistry, 1974a; 13: 211–22
2. Chou PY, Fasman GD: Prediction of protein conformation. Biochemistry, 1974b;13: 222–45
3. Chou PY: Prediction of protein structural classes from amino acid composition. In: Prediction of Protein Structure and the Principles of Protein Conformation, ed G.D. New York: Fasman Plenum Press, 1989; 549–86
4. Fasman GD: The development of the prediction of protein structure. In: Prediction of Protein Structure and the Principles of Protein Conformation, ed G.D. New York: Fasman Plenum Press, 1989; 193–316
5. Garnier J, Robson B: The GOR method for predicting secondary structures in proteins. In: Prediction of Protein Structure and the Principles of Protein Conformation, ed. G.D. New York: Fasman Plenum Press, 1989; 417–65
6. Berman HM, Westbrook J, Feng Z et al: The Protein Data Bank (2003# release). Nucleic Acids Res, 2000; 28: 235–42
7. Kinch LN, Qi Y, Hubbard TJP, Grishin NV: CASP5 target classification. Proteins, 2003; 53(Suppl.6): 340–51
8 Kosinski J, Cymerman IA, Feder M et al: A "Frankenstein's monster" approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. Proteins, 2003; 53: 369–79
9. Venclovas Č: Comparative modeling in CASP5: Progress is evident, but alignment errors remain a significant hindrance. Proteins, 2003; 53(Suppl.6): 380–88
10. Sasson I, Fischer D: Modeling three-dimensional protein structures for CASP5 using the 3-D-shotgun meta-predictors. Proteins, 2003; 53: 389–94
11. Kinch LN, Wrabl JO, Krishna SS et al: CASP5 assessment of fold recognition target predictions. Proteins, 2003; 53(Suppl.6): 395–409
12. Ginalski K, Rychlewski L: Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. Proteins, 2003; 53(Suppl.6): 410–17
13. von Grotthus M, Pas J, Wyrwicz L et al: Application of 3D-Jury GRDB and Verify3D in fold recognition. Proteins, 2003; 53(Suppl.6): 418–23
14. Liu W, Chou KC: Prediction of protein secondary structure content. Prot Engineer, 1999; 12: 1041–50
15. Zhou X, Alber F, Folkers G et al: An analysis of the helix-to-strand transition between peptides with identical sequence. Proteins Struct Func Genet, 2000; 41: 248–56
16. Prakash T, Ramakrishnan C, Dash D, Brahmachari SK: Conformational analysis of invariant peptide sequences in bacterial genomes. J Mol Biol, 2005; 345: 937–55
17. Baldwin RL: Making a network of hydrophobic clusters. Science, 2002; 295: 1657–58
18. Jurkowski W, Brylinski M, Konieczny L et al: Conformational Subspace in Simulation of Early-Stage Protein Folding. Proteins Struct, Funct. and Bioinf, 2004; 55: 115–27
19. Roterman I: Modeling of optimal simulation path in the peptide chain folding – Studies based on geometry of alanine heptapeptide. J Theor Biol, 1995; 177: 283–88
20. Kendall MG, Stuart A: The Advanced Theory of Statistics. vol. 2 New York, Hafner, 1961
21. Kendall MG, Stuart A. The Advanced Theory of Statistics. vol. 2, 4th ed. London, Griffin, 1979
22. Clayton D: Population association. In: Handbook of Statistical Genetics, ed. Balding DJ, Bishop M, Cannings C, John Wiley & Sons Ltd., London, 2001; 519–40
23. Ewens WJ, Spielman RS: The transmission/disequilibrium test. In: Handbook of Statistical Genetics, ed. Balding DJ, Bishop M, Cannings C, John Wiley & Sons Ltd., London, 2001; 507–18
24. Holaman P: Nonparametric linkage. In: Handbook of Statistical Genetics, ed. Balding DJ, Bishop M, Cannings C, John Wiley & Sons Ltd., London, 2001; 487–506
25. Elston R, Olson J, Palmer L: Biostatistical genetics and genetic epidemiology. John Wiley & Sons. Ltd., London, 2002; 29–33
26. Roterman I: The geometrical analysis of polypeptide backbone structure and its deformation. Biochimie, 1995; 77: 204–16
27. Aczel AD: Complete Business Statistics. Boston, New York, San Francisco, St Louis, Bangkok, Bogota, Caracas, Lisbon, London, Madrid, Mexico City, Milan, New Delhi, Seoul, Singapore, Sydney, Taipei, Toronto, McGraw-Hill I, 1996; 460
28. Nemethy G, Gibson KD, Palmer KA et al: Energy Parameters in Polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J Phys Chem, 1992; 96: 6472–84
29. Sippl MJ, Nemethy G, Scheraga HA: Intermolecular Potentials for Crystal Data 6. Determination of Empirical Potentials for 0-H—O=C Hydrogen Bonds for Packing Configurations. J Phys Chem, 1984; 88: 6231–33
30. Crasto CJ, Feng J: Sequence codes for extended conformation. A neighbour dependent sequence analysis of loops in proteins Proteins. Struct Func Gen, 2001; 42: 399–413
31. Prakash T, Ramakrishnan C, Dash D, Brahmachari SK: Conformational analysis of invariant peptide sequences in bacterial genomes. J Mol Biol, 2005; 345: 937–55
32. Moult J, Hubbard T, Fidelis K, Pedersen JT: Critical assessment of methods of protein structure prediction (CASP): round III. Proteins Struct Func Gen, 1999; Suppl.3: 2–6
33. Moult J, Fidelis K, Zemla A, Hubbard T: Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins Struct Func Gen 2001; Suppl.5: 2–7
34. Moult J, Fidelis K, Zemla A, Hubbard T: Critical assessment of methods of protein structure prediction (CASP): round V. Proteins Struct Func Gen, 2003; 53: 334–39
35. Venclovas C, Zemla A, Fidelis K, Moult J: Critical assessment of methods of protein structure prediction (CASP): round V. Proteins Struct Func Gen. 2003; 53: 585–95
36. Brylinski M, Jurkowski W, Konieczny L, Roterman I: Limited conformational space for early stage protein folding simulation. Bioinformatics, 2004; 20: 199–205
37. Jurkowski W, Brylinski M, Konieczny L, Roterman I: Lysozyme folded in silico according to the limited conformational sub-space. J Biomol Struct Dynam, 2004; 22: 149–58
38. Brylinski M, Jurkowski W, Konieczny L, Roterman I: Limitation of conformational space for proteins – early stage folding simulation of human α and β hemoglobin chains. TASK Quarterly, 2004; 8: 413–22
39. Fischer N, Riechmann L, Winter G A: Native-like artificial protein from antisense. DNA PEDS, 2004; 17: 13–20
40. Brylinski M, Konieczny L, Roterman I: SPI – Structure predictability index for protein sequences. In Silico Biology, 2004; 5: 0022 (internet publication)
41. Brylinski M, Jurkowski W, Czerwonko P et al: Early-stage folding in proteins – structure to sequence relation. J Biomed Biotech, 2005; 2; 65–79
42. Kim DE, Chivian D, Baker D: Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res, 2004; 1(32): W526–31
43. Chivian D, Kim DE, Malmstrom L et al: Automated prediction of CASP-5 structures using the Robetta server. Proteins, 2003; 53(Suppl.6): 524–33
44. Tlistone C, Aldhous P: Vital Statistics. Nature, 2003; 424: 610–12