In Silico Biology 5 (2005) 227–237
IOS Press

SPI – Structure Predictability Index for Protein Sequences

Michal Brylinski[a], Leszek Konieczny[b] and Irena Roterman[c,*]

[a]*Department of Bioinformatics and Telemedicine, Collegium Medicum – Jagiellonian University, Kopernika 17, 31-501 Cracow, Poland*
Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Cracow, Poland
E-mail: brylinsk@chemia.uj.edu.pl
[b]*Institute of Biochemistry, Collegium Medicum – Jagiellonian University, Kopernika 7, 31-034 Cracow, Poland*
E-mail: mbkoniec@cyf-kr.edu.pl
[c]*Department of Bioinformatics and Telemedicine, Collegium Medicum – Jagiellonian University Kopernika 17, 31-501 Cracow, Poland*
E-mail: myroterm@cyf-kr.edu.pl

Edited by E. Wingender; received 11 November 2004; revised and accepted 13 December 2004; published 28 December 2004

ABSTRACT: Estimation of structure predictability for a particular protein is difficult. Many methods estimate it in an *a posteriori* system evaluating the final, native protein structure. The SPI scale is intended to estimate the structure predictability of a particular amino acid sequence in an a *priori* system. A sequence-to-structure library was created based on the complete Protein Data Bank. The tetrapeptide was selected as a unit representing a well-defined structural motif. The early-stage folding structure (a model of which was presented elsewhere) was taken as the object for protein structure classification. Seven structural forms were distinguished for structure classification. The degree of determinability was estimated for the sequence-to-structure and structure-to-sequence relations particularly interesting for threading methods. A comparative analysis of the SPI and $Q7$ scales with the commonly used SOV and $Q3$ scales is presented. The complete contingency table, supplementary materials and all the programs used are available on request.

KEYWORDS: Protein structure prediction, predictability scale, early-stage of folding

INTRODUCTION

One of the problems faced by CASP organizers is estimation of the degree of difficulty in predicting the structure of a particular protein, in two aspects: the difficulty of the fold *per se* present in proteins, and estimation of the goodness of prediction. The method presented in this paper is aimed to help solve these problems.

The method presented here allows estimation of the structure predictability of a given protein's sequence in an *a priori* system without knowledge of the structure. Moreover, the criterion for estimation is the so-called early-stage folding (*in silico*) structure. The background of this method was presented elsewhere (geometry-based aspects [1,2] and theory of information-based [3]).

The model was verified using BPTI [4], ribonuclease [3], human hemoglobin α and β chains [5] and lysozyme [6], taking them as examples to prove the method's reliability. The early-stage folding

*Corresponding author.

Electronic publication can be found in *In Silico Biol.* **5**, 0022 <http://www.bioinfo.de/isb/2004/05/0022/>, 28 December 2004.

structural forms represented reasonable motifs without any structural defects or imperfections. Energy minimization procedure [3–6] and molecular dynamics simulation [6] delivered structural forms acceptable as possible protein structures. To make the model of early-stage folding comprehensible, a summary of the approach is presented in the Appendix.

MATERIALS AND METHODS

Data

The complete set of proteins deposited in Protein Data Bank #2003 [7] was taken for global comparative analysis of the standard $Q3$, *SOV* and newly introduced $Q7$ and SPI. Ten proteins representing different structural characteristics were selected for detailed analysis: 5RAT, 4PTI, 2EQL and 3HHB proteins present in the PDB; 1NEB as an example of "new fold"; and 1H7M, 1KOY, 1M2E, 1NYN and 1O13 – CASP5 targets [8] not present in the PDB #2003.

Structure classification

The structure classification is based on the probability profile presented in Fig. 1b. The basis and explanation of this profile becomes clear after reading the Appendix. The ellipse path shown in Fig. 1a was taken (for reasons presented in the Appendix) as the early-stage conformational sub-space. The commonly observed distribution of ϕ, ψ angles moved to the ellipse path (according to the shortest-distance criterion – Fig. 1a) created the probability distribution as shown in Fig. 1b. This figure shows overlapped profiles of ten amino acids (profiles for all 20 amino acids were shown in [3]). The t-variable (ellipse equation variable) takes its zero value at the point $\phi = 90$ deg and $\psi = -90$ deg, and then increases clock-wise along the ellipse Fig. 1c. Seven well-separated probability maxima can be distinguished. Each of them was given a one-letter code as shown in Fig. 1b.

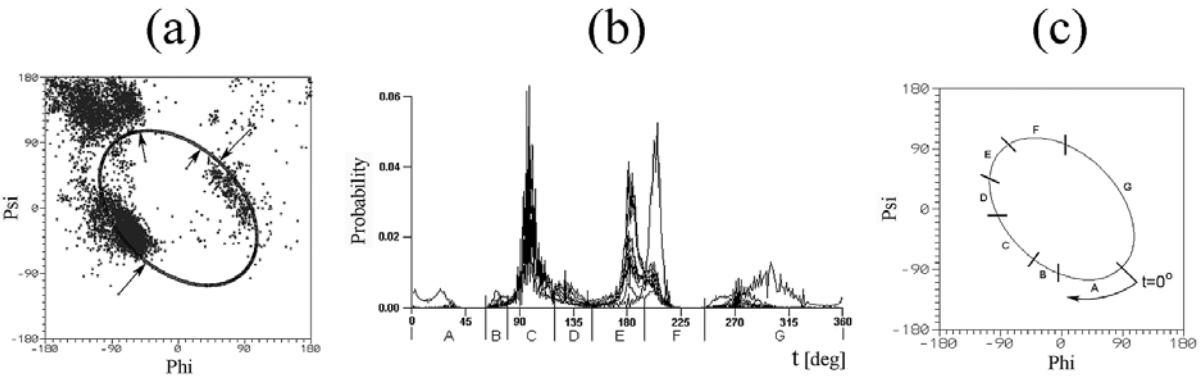

Fig. 1. Structure coding system. (a) ϕ, ψ distribution as it appears in real proteins. The ellipse path represents the limited conformational sub-space. Arrows represent the procedure of moving ϕ, ψ angles toward the ellipse path. (b) Probability profile of ten amino acids along the ellipse path. The t-parameter expresses the ellipse path. (c) Relation between t-parameter value and position on ellipse path; $t = 0$ represents the point $\phi = 90$ deg and $\psi = -90$ deg and then increases along the ellipse path clock-wise.

Structure codes

Each amino acid in the proteins was given a one-letter code (in bold in this paper) expressing the amino acid (sequence), and a one-letter code (in italics) representing the structure.

Contingency table

The tetrapeptide was taken as the shortest unit representing a well-defined structural motif (for example, β-turn, helix and others). All proteins present in the January 2003 release of PDB were analyzed according to their structural classification (following the model presented in the Appendix). Each tetrapetide was described by a four-letter string expressing sequence (bold) and a four-letter string expressing structure (italics). Potentially, 160,000 different sequences for tetrapeptides can occur (columns). Taking seven different structural forms for each amino acid in a tetrapeptide, 2401 structural forms can be distinguished for a tetrapeptide (rows). For all cells, probability values of p^t, p^c and p^r were calculated as follows:

$$p_{ij}^t = \frac{n_{ij}}{N^t} \tag{1}$$

$$p_{ij}^c = \frac{n_{ij}}{N_j^c} \tag{2}$$

$$p_{ij}^r = \frac{n_{ij}}{N_i^r} \tag{3}$$

where i denotes a particular structure (row), j denotes a particular sequence (column), n_{ij} is the number of tetrapeptides belonging to the i-th structure and representing the j-th sequence, N^t is the total number of tetrapeptides, N_i^r and N_j^c denote the number of tetrapeptidess belonging to a particular i-th structure and j-th sequence, respectively. The table expressing all probabilities (p_{ij}^t, p_{ij}^c, p_{ij}^r) is available on request (www interface in preparation). A detailed analysis of the contingency table is presented and discussed also in [9]. The values expressing the probability of the particular tetrapeptide to represent a particular structure, which can be found in each cell of the table, are utilized for the SPI and $Q7$ scales presented below.

Estimation of structure-to-sequence attribution (*Q7 coefficient*)

The structures of proteins appear to represent easy, moderate and hard predictability [10]. Since the structure is sequence-determined, the sequences will also be distinguished as easy, moderately and hard to recognize as structure-determining. Parameter $Q7$ can be introduced to measure the degree of structure determination: $Q7$, in analogy to $Q3$, is based on the fact that three structural forms are distinguished (helix, beta, random coil) in $Q3$ calculation, while seven structural forms are distinguished in the presented model (Fig. 1b). The relation between these two notations is given in Table 1.

$Q3$ measures structure predictability with three structural forms distinguished, and is calculated as follows:

$$Q3 = \frac{N_{r3}}{N} * 100 \tag{4}$$

Table 1
The relation between standard three-state-secondary structure description and newly introduced early-stage structure classification (seven states)

$Q7$ classification	Interpretation	$Q3$ classification
C	Right-handed helix	H
E, F	Strand	E
G	Left-handed helix	H
A, B, D	Random coil	C

where N expresses the total number of amino acids in the polypeptide under consideration, N_{r3} expresses the number of correctly predicted amino acids representing r structural form ($r3$ expresses one among three structural forms: right-handed helix, β-structure, random coil).

$Q7$ can be calculated using exactly the same formula with seven structural forms distinguished:

$$Q7 = \frac{N_{r7}}{N} * 100 \tag{5}$$

where N_{r7} expresses one of seven structural forms distinguished according to the code system presented in Fig. 1b.

Correct prediction of a residue's structure (according to our model) means that the correct letter-coded probability maximum was found for a particular amino acid in a sequence. The letter code represents the early-stage folding (*in silico*) structural form identified by classifying the ϕ, ψ angle in a real protein as belonging to a particular probability maximum on the ellipse path.

Early-stage structure prediction and Structure Predictability Index (*SPI*)

The data stored in the contingency table obtained according to the calculations presented above can be used for early-stage structure prediction. Since the predictability for each fragment of the whole sequence has been characterized as the potential structural form, the degree of difficulty of structure prediction for a particular amino acid sequence can be also estimated. Examples of early-stage structure prediction for an amino acid sequence are given in Fig. 2. The procedure of structure prediction and SPI calculation was performed as follows: the sequence of each target protein (Fig. 2a) was read using a sliding frame of four amino acid long, 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 highest p_{ij}^c value (Eq. (2), Fig. 2b). Each amino acid in the resulting structure obtained the state with the highest number of tetrapeptide chains belonging to a particular sequence and representing a particular structure (Fig. 2c). In addition, the mean value for all residues was calculated (SPI). SPI (x100) reaches values from 14.29 to 100.00, where 14.29 means completely random prediction.

Comparison of different scales measuring accuracy of prediction

$Q3$ [11] and *SOV* [12,13] are usually used for structure predictability, particularly in CASP projects [14, 15]. The newly introduced indexes $Q7$ and SPI are compared with $Q3$ and *SOV* using the proteins deposited in PDB #2003. The early-stage structure of each protein sequence was predicted using the contingency table described above. Moreover, for each sequence the structure predictability index (SPI) was calculated. Both native and predicted structures were characterized by calculating the $Q3$, $Q7$ and *SOV* parameters and compared to the results obtained using SPI treated as the estimation coefficient. $Q7$ parameter was calculated for seven-state predictions, whereas $Q3$ and *SOV* parameters were calculated for predictions transformed to standard three-state secondary structure description according to Table 1.

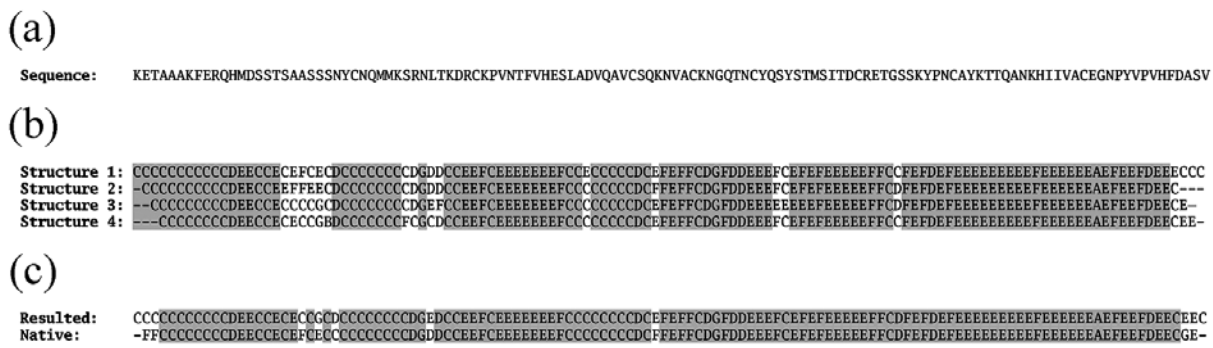

Fig. 2. Example of early-stage structure prediction for amino acid sequence (5RAT). In row (a) amino acid sequence, in row (b) the attributed structure for four different reading frames (coding system as in Fig. 1c) (gray color expresses identity), in row (c) resulting structure compared to the native one (gray shading expresses identity). 4PTI, 2EQL, 3HHB, 1NEB, 1H7M, 1KOY, 1M2E, 1NYN, 1O13 are presented in Supplementary materials.

RESULTS

Contingency table

The contingency table for tetrapeptides representing the sequence-to-structure relation was presented in [9]. A brief explanation needed to clarify the problem presented in this paper is as follows. The total number of protein chains and the corresponding number of residues in PDB #2003 were found to be 36,013 and 8,465,280, respectively. Thus the average length of protein chain was found to be 235aa. The total number of tetrapeptide sequences was calculated to be 160,000 (20^4). 146,940 different tetrapetides happened to occur in real proteins. The total number of structures (as combinations of seven distinguished probability maxima) was calculated to be 2401 (7^4), while in real proteins 2397 different structures occurred in the introduced structure coding system. Finally, the 146,940 $\times$ 2397 contingency table was analyzed for the mutual dependence or correlation between a particular tetrapeptide and its structure. The total number of different tetrapeptides in the PDB #2003 was found to be 1,529,987. The table representing the number of events expressing a particular tetrapeptide fragment sequence revealing a particular letter-coded structural form was converted to a probability scale as described in Materials and Methods.

The complete table (146,940 $\times$ 2397 cells) is treated as a library for linking the amino acid sequence (tetrapeptide) with the possible early-stage folding (*in silico*) structural forms for both dependence of structure on sequence and dependence of sequence on structure. Global analysis of the contingency table shows that the maximum number of different structures attributed to the same tetrapeptide is 144. This tetrapeptide appeared to be **GSAA**. The maximum number of different sequences was found (*CCCC*) 90,587 to be α-helix and 47,809 for β-structure (*EEEE*). Only four structures were not found in the library: *ABAB*, *ABBD*, *ABFB*, and *DBAB*.

Assessment of structure prediction accuracy

Both standard $Q3$ and *SOV* parameters as well as newly introduced $Q7$ parameter were used to evaluate the prediction results. Average prediction rates for $Q3$, *SOV* and $Q7$ were found 69.6, 58.3 and 63.9, respectively. The frequency of each structure class in PDB #2003 compared with their average prediction rates are presented in Fig. 3. The average prediction rate is expressed as the partial accuracy of early-stage structure prediction for A, B, C, D, E, F and G.

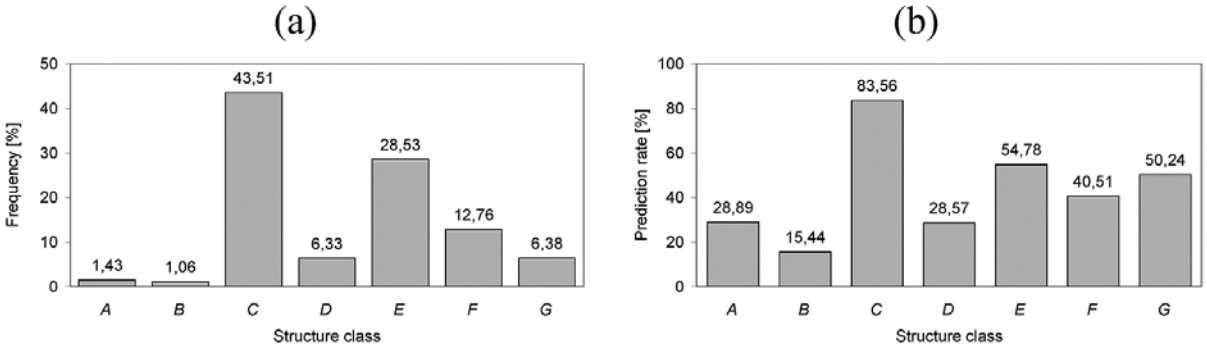

Fig. 3. The profile of seven structure classes distinguished on the basis of early-stage folding model (Fig. 1). Their frequencies in PDB #2003 (a) and average prediction rates (b).

Class C, which represents right-handed α-helix, is obviously the most frequent and its average prediction rate is the highest one. The class including left-handed α-helix (*G*) appeared to represent a very good prediction rate, especially compared with its low frequency. For class G over the half of residues seem to be predicted correctly. Despite the high frequency of β-structure (41.29 for E and F altogether), the average prediction rate is still low. It confirms that β-structure prediction should go beyond analysis of the local sequence-structure relation. An interesting fact is that prediction rates of classes representing loops (A, B, D) are fairly high in comparison with their low frequencies in PDB.

Q7 calculation and Structural Predictability Index (*SPI*)

$Q7$ does not change the general features of structure interpretation, and provides more detailed characteristics of α-helices (C and G for right- and left-handed, respectively), β-structure (two forms distinguished in $Q7$ scale: E, F) and random coil (three forms distinguished: A, B, D). The SPI coefficient estimates difficulty in an *a priori* structure prediction. Since the structural predictability of each tetrapeptide is known, the whole sequence can be estimated. SPI seems to be a good coefficient to estimate the early-stage folding structural predictability of the amino acid sequence. It should be noted that the SPI coefficient can be calculated for amino acid sequences without knowing the final native structure. The results of comparative analysis of standard ($Q3$, *SOV*) and newly introduced (SPI, $Q7$) parameters for the complete set of proteins deposited in PDB #2003 are given in Table 2 and Fig. 4. The R^2 coefficient was calculated for second-degree polynomial approximation for each pair of compared methods (SPI versus $Q3$, SPI versus $Q7$ and SPI versus *SOV*). Its values, always above 0.8, suggest high accordance between the compared parameters. Detailed results for selected proteins given in Materials and Methods are shown in Fig. 2, Table 3 and supplementary materials.

Table 2
Second degree polynomial approximations and correlation coefficients (R^2) for each pair of compared parameters calculated for the complete set of proteins deposited in PDB #2003 (Fig. 4)

Compared parameters	Approximation	R^2
SPI vs. $Q3$	$Q3 = -0.0278 * SPI^2 + 6.710 * SPI - 293.667$	0.8464
SPI vs. $Q7$	$Q7 = -0.0372 * SPI^2 + 8.575 * SPI - 388.156$	0.8527
SPI vs. *SOV*	$SOV = -0.0130 * SPI^2 + 4.560 * SPI - 229.033$	0.8031

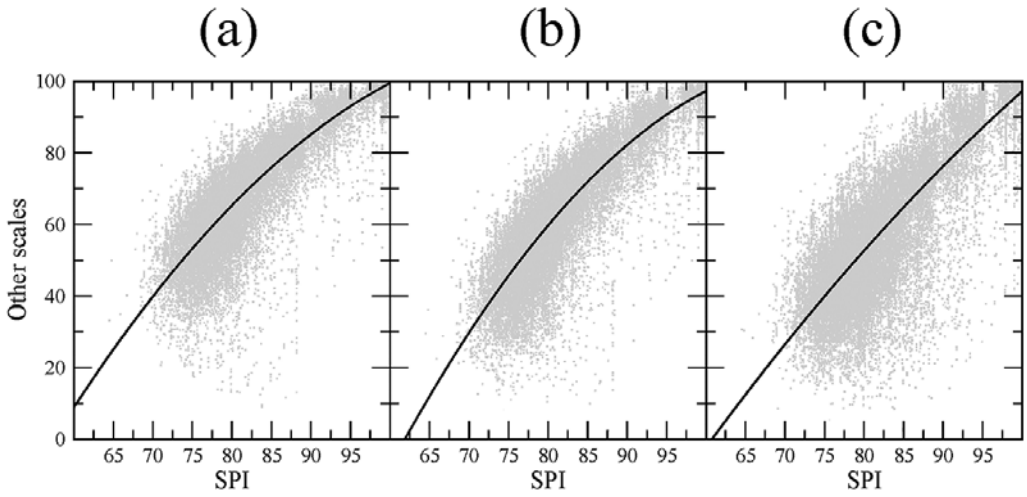

Fig. 4. Structure Predictability Index (SPI) in relation to the accuracy of structure prediction for the complete set of proteins deposited in PDB #2003. (a) SPI versus $Q3$, (b) SPI versus $Q7$, and (c) SPI versus SOV. Solid lines represent second-degree polynomial approximation for each pair of compared methods. The R^2 coefficients and equations are given in Table 2.

DISCUSSION

The model presented in this paper attempts to solve few problems related to protein folding simulation. Generally, two approaches can be proposed to simplify the multidimensional character of the problem: (1) simplification of polypeptide structure and (2) limitation of the conformational space. The first approach is quite frequently presented in many papers [16,17]. The second has been claimed to be necessary [18,19]. The model of basins distinguished on the Ramachandran map was presented and proposed as the solution of the hyper-dimensionality of the conformational space [20,21].

The presented model seems to link both approaches: the geometry is treated as the sequence of rigid peptide bond planes, with the radius of curvature (shape of polypeptide chain) dependent on the dihedral angle between peptide bond planes, with an elliptical limited conformational sub-space.

Models known in the literature concerning the problem of the sequence-to-structure relation discuss the structure of proteins as it appears in the final native form of the protein [22–30]. The model introduced in this paper represents an approach for the relation between sequence and structure in the early-stage folding (*in silico*) structural form (the basis for the model is presented in detail in [1–3] and verified by BPTI [4], ribonuclease [3], lysozyme [6] and hemoglobin [5] folding).

The tetrapeptide was selected as the unit because it represents the shortest chain that can represent a well-defined structural form (helix, β-sheet, β-turn) [31,32]. The structure-coding system, which treats all possible structural forms in a common, unified model, includes all irregular random forms in the same scale together with regular conformations. This enabled us to distinguish quite unexpected loop-creating sequences, which in the traditional three-category classification (helix, beta, coil) could get lost. The traditional models do not distinguish different forms of random coiled fragments. The coding system introduced here can very easily distinguish different unstructured forms. The high correlation between the traditional and newly introduced models makes them good tools to use together for structure classification.

Most of local structure prediction methods have focused on three-state secondary structure prediction. Statistical [22,23], information theory [33,34], pattern recognition [35,36], neural networks [37,38] and

Table 3
Different scales adopted to measure the accuracy of structure prediction of selected proteins

Protein	Residues	SPI	$Q7$	Q_A	Q_B	Q_C	Q_D	Q_E	Q_F	Q_G	Q_3	Q_{helix}	Q_{beta}	Q_{coil}	SOV	$\mathrm{SOV}_{\mathrm{helix}}$	$\mathrm{SOV}_{\mathrm{beta}}$	$\mathrm{SOV}_{\mathrm{coil}}$
3HHB	141	99.2	97.1	100.0	–	99.1	72.7	100.0	100.0	100.0	97.1	99.1	100.0	76.9	95.7	97.8	100.0	73.1
5RAT	124	95.0	93.4	100.0	–	97.4	90.9	97.9	81.0	66.7	94.3	97.4	94.2	85.7	86.0	80.6	93.5	72.2
4PTI	58	95.0	87.5	–	–	94.7	66.7	94.7	80.0	60.0	91.1	100.0	93.1	62.5	89.7	96.1	93.1	62.5
2EQL	129	79.2	56.7	0.0	0.0	84.1	25.0	33.3	23.1	50.0	62.2	84.8	37.8	37.5	56.0	71.7	38.7	37.5
1NEB	60	77.8	37.9	–	–	92.9	20.0	25.0	0.0	40.0	44.8	92.9	26.5	40.0	38.6	63.4	29.4	35.0
1KOY	62	89.0	75.0	–	0.0	93.8	0.0	0.0	0.0	–	81.7	91.8	50.0	0.0	85.8	98.0	43.8	0.0
1M2E	135	82.7	53.4	0.0	0.0	79.4	0.0	36.8	7.7	66.7	60.2	80.0	43.1	16.7	45.5	48.0	46.6	16.7
1H7M	99	80.6	63.9	0.0	–	92.5	0.0	45.8	11.1	20.0	71.1	92.6	51.5	20.0	63.5	82.9	48.9	20.0
1O13	105	78.4	46.6	0.0	–	85.7	16.7	35.3	17.6	22.2	51.5	81.1	37.3	26.7	38.7	71.0	25.9	22.2
1NYN	111	73.4	48.6	0.0	0.0	77.1	0.0	35.3	10.0	60.0	54.1	73.7	29.5	50.0	44.6	55.3	31.7	41.7

Q_A, Q_B, Q_C, Q_D, Q_E, Q_F and Q_G express the partial accuracy of early-stage structure prediction for A, B, C, D, E, F and G (on the basis of the coding system introduced in Fig. 1). Global measurement is expressed by $Q7$. Q_{helix}, Q_{beta}, Q_{coil}, $\mathrm{SOV}_{\mathrm{helix}}$, $\mathrm{SOV}_{\mathrm{beta}}$ and $\mathrm{SOV}_{\mathrm{coil}}$ represent the partial accuracy of helix, β-sheet and random coil structural forms prediction. The global estimation is expressed by both $Q3$ and SOV. SPI expresses the structure predictability of given sequence.

nearest-neighbour methods [39,40] have been developed. With the best methods, residues in a particular sequence can be assigned to one of three structural categories (helix, strand, coil) with average success rates of roughly 60–70% [41]. The application of multiply aligned sequences brings about a gain in prediction accuracy of 6–8%, relative to the single case, insisting that the secondary structure must be the same for all of the family members [11,12,28,42,43]. Our approach achieved almost 70% of the average prediction rate in standard three-state description for a single amino acid sequence. It gives the opportunity to improve the accuracy of prediction using sequence alignment as input. Simultaneously, the extension of local structure description to seven classes slightly decreased the accuracy by 5.7%. Interestingly, the prediction rate of classes not belonging to repetitive secondary structures as well as the class representing left-handed helix are significant high in relation to their frequencies in PDB. However, prediction rates of loops are still lower than those of repetitive secondary structures. There is an evidence, that the utilization of a special loop library yields better accuracy in loops prediction [44]. A special loop library according to our model will be created and applied for loop prediction in the future.

Only the most probable sequence-structure combinations were presented in this paper, although some alternative structural forms can be constructed (lower probability in contingency table structural attribution), allowing prediction of nonstandard structural motifs. The early-stage folding (*in silico*) model verified here on the basis of the whole PDB seems to offer a tool for starting structure definition (for further energy minimization, molecular dynamics simulation and other procedures).

A separate analysis for species-dependent contingency tables (human, mammalian, insect, bacteria, etc.), for the particular biological activity of groups of proteins (trans-membrane, interacting with DNA, particular enzymes, etc.) will be done in the near future, together with comparative analysis. The present paper was focused on the practical usefulness of this approach.

Our approach also revealed that unordered structures represent high determinability. This may mean that the folding pathway is initiated by turns and bends, which are the strategic points in the polypeptide, probably followed by the second step in the folding process, which is the creation of highly ordered structures.

ACKNOWLEDGEMENTS

Many thanks to Prof. Marek Pawlikowski, Faculty of Chemistry, Jagiellonian University, for fruitful discussions. The work was financially supported by Collegium Medicum grants (501/P/133/L,

WL/222/P/L).

An example visualizing the relation between the ϕ, ψ distribution on the Ramachandran map (SER) and the ellipse path is shown in Fig. 1App. The probability profile after moving all ϕ, ψ angles to the ellipse path (shortest-distance criterion) is shown in Fig. 1(a). The details concerning the geometric basis of the model can be found in [1,2,45]. The information entropy analysis is presented in [3]. The model has been successfully applied to BPTI [4], ribonuclease [3], hemoglobin [5] and lysozyme [6] folding to prove the model's reliability.

APPENDIX

The polypeptide chain can be described by a representation other than ϕ, ψ angles. Two geometric parameters seem to describe the polypeptide conformation: V-angle [deg] – dihedral angle between two sequential peptide bond planes, and R [Å] – radius of curvature, which was found to depend on the V-angle. The dependency between these two parameters appeared to accord with a second-degree polynomial. The structures fitting this relation localized on the Ramachandran map revealed the part of the map creating the conformational sub-space. This sub-space, which appeared to be ellipse-shaped, represents the polypeptide chain structures depending only on the backbone conformation. This is why it was assumed to represent the early-stage folding structures. The sub-space satisfies two important conditions: (1) it links all structurally important areas (right-handed helix, C7eq energy minimum and left-handed helix); and (2) the amount of information stored in amino acid sequence appeared to be equilibrated with the amount of information necessary to predict the structure to the extend of early-stage folding conformation.

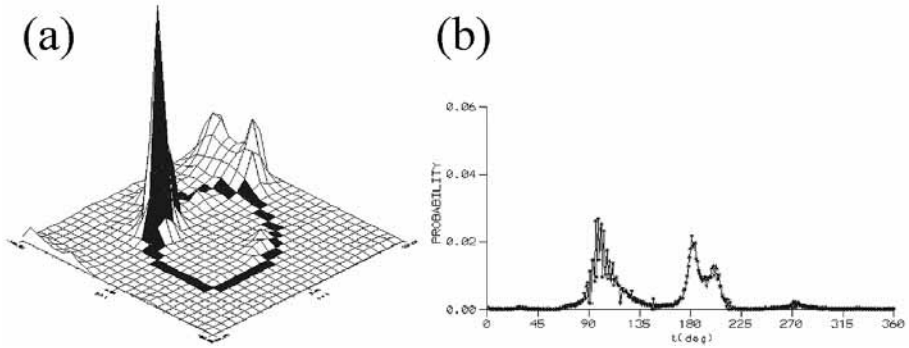

Fig. 1 App. The ϕ, ψ angle of SER distribution as found in the complete PDB (January 2003 release). (a) All over the Ramachandran map, black line distinguishes the ellipse-path, (b) After moving all ϕ, ψ angles toward the ellipse path. The variable called t expresses the variable of the ellipse equation; zero value of t expresses the point $\phi = 90$ deg, $\psi = -90$ deg and then increases clock-wise along the ellipse.

REFERENCES

[1] Roterman, I. (1995). Modelling of optimal simulation path in the peptide chain folding - Studies based on geometry of alanine heptapeptide. J. theor. Biol. **177**, 283-288.

[2] Roterman, I. (1995). The geometrical analysis of polypeptide backbone structure and its deformation. Biochimie **77**, 204-216.

[3] Jurkowski, W., Brylinski, M., Konieczny, L., Wisniowski, Z. and Roterman, I. (2004). The conformational sub-space in simulation of early-stage protein folding. Proteins **55**, 115-127.

[4] Brylinski, M., Jurkowski, W., Konieczny, L. and Roterman, I. (2004). Limited conformational space for early stage protein folding simulation. Bioinformatics **20**, 199-205.
[5] Brylinski, M., Jurkowski, W., Konieczny, L. and Roterman, I. (2004). Limitation of conformational space for proteins - early stage folding simulation of human α and β hemoglobin chains. TASK-Quarterly **8**, 413-422.
[6] Jurkowski, W., Brylinski, M., Konieczny, L. and Roterman, I. (2004). Lysozyme folded in silico according to the limited conformational sub-space. J. Biomol. Struct. Dyn. **22**, 149-158.
[7] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acid Res. **28**, 235-242.
[8] Kinch, L. N., Qi, Y., Hubbard, T. J. P. and Grishin, N. V. (2003). CASP5 target classification. Proteins **53**, 340-351.
[9] Brylinski, M., Konieczny, L., Czerwonko, P., Jurkowski, W. and Roterman, I. (2005). Early-stage folding in proteins (*in silico*) - Sequence to structure relation. J. Biomed. Biotech., in press.
[10] Orengo, C. A., Bray, J. E., Hubbard, T. J., LoConte, L. and Sillitoe, I. (1999). Analysis and assessment of ab initio three-dimensional prediction, secondary structure and contacts prediction. Proteins **3** (suppl. 3), 149-170.
[11] Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. **232**, 584-599.
[12] Rost, B., Sander, C. and Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. J. Mol. Biol. **235**, 13-26.
[13] Zemla, A., Venclovas, C., Fidelis, K. and Rost, B. (1999). A modified definition of Sov, a segment based measure for protein secondary structure prediction assessment. Proteins **34**, 220-223.
[14] Zemla, A., Venclovas, C., Reinhardt, A., Fidelis, K. and Hubbard, T.J. (1997). Numerical criteria for the evaluation of ab initio predictions of protein structure. Proteins **1** (suppl. 1), 140-150.
[15] Aloy, P., Stark, A., Hadley, C. and Russell, R. B. (2003). Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins **53**, 436-456.
[16] Liwo, A., Czaplewski, C., Pillardy, J. and Scheraga, H. A. (2001). Cummulation-based expression for the multibody terms for the correction between local and electrostatic interaction in the united residue force field. J. Chem. Phys. **115**, 2323-2347.
[17] Liwo, A., Arlukowicz, P., Czaplewski, C., Oldziej, S., Pillardy, J. and Scheraga, H. A. (2002). A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape - Application to the UNRES force field. Proc. Natl. Acad. Sci. USA **99**, 1937-1942.
[18] Colubri, A. (2004). Prediction of protein structure by simulating coarse-grained folding pathways: A preliminary report. J. Biomol. Struct. Dyn. **21**, 625-638.
[19] Alonso, D. O. V. and Daggett, V. (1998). Molecular dynamics simulations of hydrophobic collapse of ubiquitin. Protein Sci. **7**, 860-874.
[20] Fernandez, A., Colubri, A., Aqpigmanesi, G. and Burastero, T. (2001). Coarse semiempirical solution to the protein folding problem. Physica A **293**, 358-384.
[21] Sosnick, T. R., Berry, R. S., Colubri, A. and Fernandez, A. (2002). Discriminating foldable proteins from nonfolders: when and how do they differ? Proteins **49**, 15-23.
[22] Chou, P. Y. and Fasman, G. D. (1974). Conformational parameters for amino acids in helical, Beta-sheet and random coil region calculated from proteins. Biochemistry **13**, 211-222.
[23] Chou, P. Y. and Fasman, G. D. (1974). Prediction of protein conformation. Biochemistry **13**, 222-245.
[24] Efimov, A. V. (1984). A novel super-secondary structure of proteins and the relation between the structure and the amino acid sequence. FEBS Letters **166**, 33-38.
[25] Unger, R., Harel, D., Wherland, S. and Sussman, J. L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins **5**, 355-373.
[26] Han, K. F. and Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc. Natl. Acad. Sci. USA **93**, 5814-5818.
[27] Rychlewski, L. and Godzik, A. (1997). Secondary structure prediction using segment similarity. Protein Eng. **10**, 1143-1153.
[28] Salamov, A. A. and Solovyev, V. V. (1997). Protein secondary structure prediction using local alignments. J. Mol. Biol. **268**, 31-36.
[29] de Brevern, A. G., Valadie, H., Hazout, S. and Etchebest, C. (2002). Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci. **11**, 2871-2886.
[30] de Brevern, A. G., Benros, C., Gautier, R., Valadie, H., Hazout, S. and Etchebest, C. (2004). Local backbone structure prediction of proteins. In Silico Biol. **4**, 0031.
[31] Bystroff, C., Thorsson, V. and Baker, D. (2000). HMMSTR: a Hidden Markov Model for local sequence-structure correlations in proteins. J. Mol. Biol. **301**, 173-190.
[32] Zhu, Z. Y. and Blundell, T. L. (1996). The use of Amino acid patterns of classified helices and strands in secondary structure prediction. J. Mol. Biol. **260**, 261-276.

[33] Garnier, J., Osguthorpe, D. J. and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. **120**, 97-120.
[34] Gibrat, J. F., Garnier, J. and Robson, B. (1987). Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs. J. Mol. Biol. **198**, 425-443.
[35] Taylor, W. R. and Thornton, J. M. (1983). Prediction of super-secondary structure in proteins. Nature **301**, 540-542.
[36] Presnell, S. R., Cohen, B. I. and Cohen, F. E. (1992). A segment-based approach to protein secondary structure prediction. Biochemistry **31**, 983-993.
[37] Qian, N. and Sejnowski, T. J. (1988). Predictind the secondary structure of globular proteins using neural network models. J. Mol. Biol. **202**, 865-884.
[38] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**, 195-202.
[39] Zhang, X., Mesirov, J. P. and Waltz, D. L. (1992). Hybrid system for protein secondary structure prediction. J. Mol. Biol. **225**, 1049-1063.
[40] Yi, T. M. and Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbour methods. J. Mol. Biol. **232**, 1117-1129.
[41] Shortle, D. (2000). Prediction of protein structure. Curr. Biol. **10**, R49-R51.
[42] Mehta, P. K., Heringa, J. and Argos, P. (1995). A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. Protein Sci. **4**, 2517-2525.
[43] Salamov, A. A. and Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. **247**, 11-15.
[44] Fourrier, L., Benros, C. and de Brevern, A. (2004). Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics **5**, 58.
[45] Roterman, I. and Konieczny, L. (1995). Geometrical analysis of structural changes in immunoglobin domains' transition from native to molten state. Computers and Chemistry **19**, 247-252.