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

Jan Meus\textsuperscript{1,}\textsuperscript{2}, Michał Brylinski\textsuperscript{1,2}, Monika Piwowar\textsuperscript{1,3}, Piotr Piwowar\textsuperscript{3,}\textsuperscript{4}, Zdzisław Wiśniowski\textsuperscript{1,3}, Justyna Stefaniak\textsuperscript{4,5}, Leszek Konieczny\textsuperscript{5,6}, Grzegorz Surówka\textsuperscript{6,7}, Irena Roterman\textsuperscript{1,6,8}

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

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. Tetrapeptide sequences and structures were expressed by letter codes. The transformation of a contingency table of any size (here: 160,000×2,401) to a 2×2 table performed for each non-zero cell of the original table allowed calculation of the \( r \)-coefficient measuring the strength of the relation.

Results: High values of the \( r \)-coefficient extracted sequences of strong structural determinability and structures of high sequence selectivity. The web-site program to calculate the \( r \)-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 • \( r \)-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 5,6-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 χ² 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).
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.

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).
The phi and psi values found for the probability maxima (according to Figure 1C) were used to create a blocked tetrapeptide (ACE-X₁-X₂-X₃-X₄-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 ρ̂-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 ρ̂ values are listed in Table 3. Different fonts were used to distinguish the letter codes for sequence (italics) from those for structure (bold). 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
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-

### Table 3.

Ranking list of the top twenty \( \rho \)-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.

<table>
<thead>
<tr>
<th>Z coefficient *E-001</th>
<th>Structure</th>
<th>Sequence</th>
<th>Position number in alternate method</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.45</td>
<td>AEGD</td>
<td>GNES</td>
<td>7</td>
</tr>
<tr>
<td>6.92</td>
<td>BEBF</td>
<td>ERSY</td>
<td>10</td>
</tr>
<tr>
<td>6.58</td>
<td>AFFB</td>
<td>GFRN</td>
<td></td>
</tr>
<tr>
<td>6.57</td>
<td>AEBB</td>
<td>GPYY</td>
<td></td>
</tr>
<tr>
<td>6.18</td>
<td>AEGE</td>
<td>GIGH</td>
<td>5</td>
</tr>
<tr>
<td>5.78</td>
<td>ABFF</td>
<td>GPHF</td>
<td></td>
</tr>
<tr>
<td>5.71</td>
<td>EAAD</td>
<td>KGGP</td>
<td></td>
</tr>
<tr>
<td>5.49</td>
<td>GAGD</td>
<td>FNAG</td>
<td></td>
</tr>
<tr>
<td>5.34</td>
<td>DBAA</td>
<td>NGGD</td>
<td></td>
</tr>
<tr>
<td>5.19</td>
<td>GCFF</td>
<td>GDSS</td>
<td></td>
</tr>
<tr>
<td>5.16</td>
<td>DDBG</td>
<td>SHHG</td>
<td></td>
</tr>
<tr>
<td>5.10</td>
<td>BBDF</td>
<td>KTRS</td>
<td>1;2</td>
</tr>
<tr>
<td>5.09</td>
<td>GDAE</td>
<td>ESGH</td>
<td></td>
</tr>
<tr>
<td>5.03</td>
<td>DBEB</td>
<td>AERS</td>
<td></td>
</tr>
<tr>
<td>5.02</td>
<td>BACE</td>
<td>GGAE</td>
<td></td>
</tr>
<tr>
<td>4.95</td>
<td>CABF</td>
<td>AGPH</td>
<td></td>
</tr>
<tr>
<td>4.95</td>
<td>AEED</td>
<td>GLRL</td>
<td></td>
</tr>
<tr>
<td>4.91</td>
<td>DAFG</td>
<td>DGPG</td>
<td>3</td>
</tr>
<tr>
<td>4.91</td>
<td>ADEE</td>
<td>GIFR</td>
<td></td>
</tr>
<tr>
<td>4.88</td>
<td>AGAE</td>
<td>IKYG</td>
<td>2</td>
</tr>
<tr>
<td>4.85</td>
<td>CFAG</td>
<td>NTGG</td>
<td></td>
</tr>
<tr>
<td>4.79</td>
<td>EBCB</td>
<td>ELPD</td>
<td></td>
</tr>
<tr>
<td>4.77</td>
<td>CADD</td>
<td>RGRC</td>
<td></td>
</tr>
<tr>
<td>4.67</td>
<td>GGED</td>
<td>KGHH</td>
<td></td>
</tr>
<tr>
<td>4.66</td>
<td>AGCE</td>
<td>GGDD</td>
<td>8</td>
</tr>
<tr>
<td>4.57</td>
<td>DFDA</td>
<td>YNVP</td>
<td></td>
</tr>
<tr>
<td>4.56</td>
<td>BFBE</td>
<td>PEPV</td>
<td></td>
</tr>
<tr>
<td>4.54</td>
<td>GFDD</td>
<td>GQTN</td>
<td></td>
</tr>
<tr>
<td>4.51</td>
<td>FAEA</td>
<td>PGFG</td>
<td></td>
</tr>
<tr>
<td>4.47</td>
<td>AEGF</td>
<td>GCAR</td>
<td>6</td>
</tr>
<tr>
<td>4.47</td>
<td>ADFA</td>
<td>GTQC</td>
<td></td>
</tr>
</tbody>
</table>
Comparison of the results with those of other tools common-ly used for comparative modeling [42,43] is difficult because of the different bases on which the particular mod-
els were constructed. The early-stage conformation is the background for our analysis, while the final, native con-
formation of proteins is used in other methods. The proce-
dure for reducing the contingency table seems to be uni-
versal 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 anx-
iously awaited [44].

**Conclusions**

The early-stage model has been verified using BPTI [36],
lysozyme [37], and ribonuclease [18]. The verification ex-
cluded any forbidden structural forms in these proteins
(for example knots). The approach to the native struc-
tural forms after application of an energy minimization pro-
cedure or molecular dynamics simulation seems to be sat-
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 con-
tingency 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/). 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-struc-
ture 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 struc-
tural 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 se-
quence. 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 con-
tingency table is the observation shown in Figure 2. The se-
quence 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 accord-
ing to high values of \( \rho \) coefficient, seem to be promising.

The presented analysis of the relation between sequence
and structure in early-stage folding made possible quantita-
tive 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 ap-

**Figure 2.** Ten top structures of tetrapeptides with the highest
structure-to-sequence determinability according to Table 3 (ten highest values of \( \rho \) 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.
peared 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 $\rho$-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:

9. Venclovas Č: Comparative modeling in CASP5: Progress is evident, but alignment errors remain a significant hindrance. Proteins, 2003; 53(Suppl.6): 380–88
Index Copernicus
Global Scientific Information Systems for Scientists by Scientists

www.indexcopernicus.com

Index Copernicus integrates

IC Scientists
Effective search tool for collaborators worldwide. Provides easy global networking for scientists. C.V.’s and dossiers on selected scientists available. Increase your professional visibility.

IC Journal Master List
Scientific literature database, including abstracts, full text, and journal ranking. Instructions for authors available from selected journals.

IC Patents
Provides information on patent registration process, patent offices and other legal issues. Provides links to companies that may want to license or purchase a patent.

IC Conferences
Effective search tool for worldwide medical conferences and local meetings.

IC Grant Awareness
Need grant assistance? Step-by-step information on how to apply for a grant. Provides a list of grant institutions and their requirements.

IC Virtual Research Groups [VRG]
Web-based complete research environment which enables researchers to work on one project from distant locations. VRG provides:
- customizable and individually self-tailored electronic research protocols and data capture tools,
- statistical analysis and report creation tools,
- profiled information on literature, publications, grants and patents related to the research project,
- administration tools.

IC Lab & Clinical Trial Register
Provides list of on-going laboratory or clinical trials, including research summaries and calls for co-investigators.

IC Patents
Provides information on patent registration process, patent offices and other legal issues. Provides links to companies that may want to license or purchase a patent.