BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Server

Volume 11(10)

Contingency Table Browser – prediction of early stage protein structure

Barbara Kalinowska[1,2], Artur Krzykalski[2], Irena Roterman[1,*]

[1]Department of Bioinformatics and Telemedicine, Collegium Medium, Jagiellonian University, Lazarza 16, 31-530 Krakow, Poland; [2]Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, 30-348, Łojasiewicza 11, Krakow, Poland; Irena Roterman – Email: myroterm@cyf-kr.edu.pl; *Corresponding author

Received October 02, 2015; Revised October 05, 2015; Accepted October 19, 2015; Published October 31, 2015

Abstract:
The Early Stage (ES) intermediate represents the starting structure in protein folding simulations based on the Fuzzy Oil Drop (FOD) model. The accuracy of FOD predictions is greatly dependent on the accuracy of the chosen intermediate. A suitable intermediate can be constructed using the sequence-structure relationship information contained in the so-called contingency table – this table expresses the likelihood of encountering various structural motifs for each tetrapeptide fragment in the amino acid sequence. The limited accuracy with which such structures could previously be predicted provided the motivation for a more in-depth study of the contingency table itself. The Contingency Table Browser is a tool which can visualize, search and analyze the table. Our work presents possible applications of Contingency Table Browser, among them – analysis of specific protein sequences from the point of view of their structural ambiguity.

Background:
The relation between a protein's conformation and its residue sequence is a key problem in protein structure prediction. The most accurate prediction methods, such as those implemented by Rosetta **[1]** or I-Tasser **[2],** combine a knowledge-based approach with molecular dynamics simulations. The process relies on sequence-structure relationship information which relates sequences to known secondary structures. Such information is usually expressed in the form of libraries. The length of the input sequence fragment varies, usually falling between 3 and 9 residues **[3].** Since native conformations depend not only on local interactions but also on interactions with sequentially distant fragments, local sequence-structure information is only partially accurate **[4]** and often ambiguous. Nevertheless, local "lookup" libraries are used by many protein structure prediction algorithms, such as those based on statistical potentials **[2],** Monte Carlo simulations **[1]** or neural networks **[5].**

The Contingency Table Browser presented in this work is intended as a visualization and analysis aid supporting the Early Stage model (proposed by Roterman **[6,7]),** and can be used to produce suitable early-stage folding intermediates on the basis of the so-called Fuzzy Oil Drop (FOD) model. Contrary to other leading methods, our approach relies on restricting the set of potential starting structures and replicating the *in vivo* folding process by taking into account hydrophobicity density distribution throughout the protein body.

Early Stage model
The Early Stage (ES) model bases on the assumption that – at least at the initial folding stage – selection of the optimal conformation of each peptide bond in the protein backbone determines the structure of the emerging intermediate **[8].** We thus search for a limited conformational subspace which expresses the geometry of the polypeptide chain using two parameters. Analysis of the backbone indicates strong correspondence between the dihedral angles formed by adjacent peptide bond planes and the radius of curvature of the resulting chain. The function which defines this relationship also establishes a limited conformational subspace of φ,ψ angles, which manifests itself as an elliptical path on the Ramachandran plot **[Figure 1a**]. Its defining characteristic is that it intersects zones of the plot which correspond to each basic secondary structure (β-sheet, right-handed helix and left-handed helix). Casting actual pairs of φ,ψ angles measured for a large number of native structures onto this elliptical path (using the minimum distance criterion) produces local conformation probability profiles for each amino acid. Such profiles exhibit

seven distinct probability peaks **[9]** to which we ascribe seven structural codes (A to G) denoting specific zones on the Ramachandran plot **(Figure 1a**). Thus, given a set of φ,ψ angles characterizing the input chain, we can assign a structural code to each of its constituent residues. In a similar manner, each known tertiary sequence can be expressed as a set of structural codes. The conformation adopted by each residue is thus accurate to within one of seven zones on the Ramachandran plot, corresponding to various secondary structures. For certain codes (such as C, which corresponds to an α-helix, as well as E and F representing β-sheets) this assignment is quite unambiguous, while in the case of other codes all we can say is that the given residue forms part of a loop.

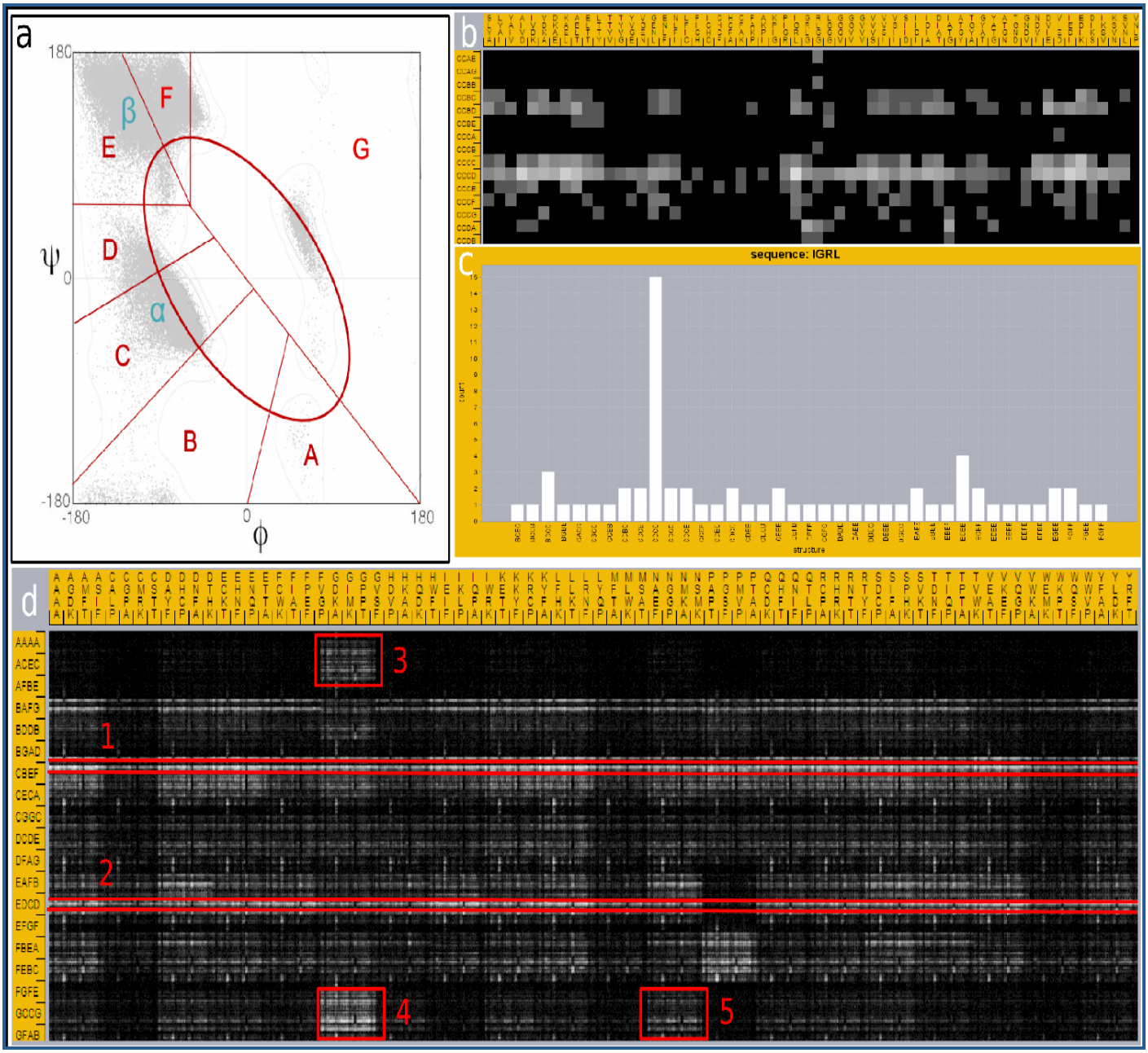

Figure 1: a) Subdivision of the φ,ψ conformational space (Ramachandran plot) into seven zones corresponding to seven structural codes (A to G). Grey patches indicate the most frequently occurring conformations of the protein backbone and its secondary folds (*α-helixes* and *β-strands)*. The elliptical path expresses the limited conformational subspace to which the early stage (ES) intermediate is assumed to belong; **b)** Fragment of the contingency table visualized by Contingency Table Browser. Columns correspond to individual tetrapeptide fragments in protein 2BA2 (PDB code) while rows correspond to structural motifs; **c)** Frequency of occurrence of each four-letter structural motif for a specific tetrapeptide (IGRL) visualized as a bar chart; **d)** Visualization of the entire contingency table (columns correspond to tetrapeptides while rows represent structural motifs). Despite the overwhelming volume of data, preferred conformation zones can clearly be discerned. For example, the two marked bands correspond to α-helixes (1) and β-strands (2) respectively. Additionally, we have highlighted the prevalence of codes A (3) and G (4) for glycine-containing tetrapeptides, as well as the characteristic correlation between proline and code G (5).

Input data
In order to gather information regarding sequence-structure relationship we selected a set of tertiary structures from the Protein Data Bank (PDB) making sure that no two sequences exhibit structural similarity in excess of 95%. We then prepared a contingency table for tetrapeptide fragments **[10].** The table lists the frequency of occurrence of each structural code for a given tetrapeptide. The dimensions of the table are 160,000 columns (combinations of four peptides from a set of 20) and 2401 rows (combinations of four structural codes from a set of 7).

ES prediction
On the basis of our contingency table we can try to determine the likelihood of encountering a given structural code at each position in an arbitrary polypeptide sequence. The average accuracy of this method is 46% **[10].** One of the reasons for this limited effectiveness is the high degree of structural ambiguity for certain tetrapeptide fragments – indeed, it appears that predicting the conformation of the early stage intermediate requires a more in-depth study of the contingency table and probability profiles for each structural motif separately. To facilitate this process we have developed the CTB tool which supports visualization, browsing and analysis of the entire contingency table, as well as its selected parts.

Description of the Contingency Table Browser
The CTB tool operates on text files which have been prepared for each tetrapeptide separately. An input file consists of two data columns: a list of structural motifs expressed as four-letter sequences and the number of occurrences of each sequence. The program can process an arbitrary number of input files – it can load the entire contingency table at once, or just a fragment thereof. The user can determine the ordering of tetrapeptides and sequences by supplying a text file which contains an ordered list of each. This enables users to focus on specific sequences, including those which exhibit a high degree of structural ambiguity **(Figure 1b).** Contingency Table Browser visualizes the contingency table by applying a grayscale (0 to 255) to each pixel depending on the corresponding frequency of occurrence. The greyscale is normalized in such a way that pure white corresponds to maximum frequency while pure black indicates complete absence of the corresponding motif. Since the maximum value present in the contingency table is 193, the table can be unambiguously visualized using 256 shades of grey. Users may manipulate visualization characteristics by reversing the greyscale, applying gamma correction, visualizing all nonzero values using a single color (white) or enhancing the intensity of either the highest or the lowest values. Given the relatively large area of the contingency table, zooming out may result in multiple values "competing" for a single pixel – in such cases selected subzones can be cropped out for display in a separate window. Another useful tool is the ability to generate frequency bar graphs for a specific tetrapeptide (or a specific structural motif) – this is done by clicking the appropriate column or row (**Figure 1c).**

Despite the large volume of data embodied in the contingency table our tool can be used to draw useful conclusions regarding the statistical properties of individual codes and residues. **Figure 1d** reveals some interesting regularities, showing bars which correspond to α-helixes and β-strands, as well as certain rare structural codes, such as A and G.

Conclusions:
Contingency Table Browser is a tool for visualization and analysis of the contingency table which expresses the correspondence between structural motifs and protein sequences (derived from PDB) in the early stage intermediate. It can aid researchers in applying custom modifications to the ES structure, which would be difficult to achieve solely on the basis of quantitative data regarding the frequency of each structural code. Visual inspection may enhance analysis of certain protein structures and augment statistical methods. It should also be noted that the tool's usability extends beyond the ES intermediate and can include any categorization of the available motifs using letter codes. The tool is freely available at http://www.unique-solutions.pl/ctb/

References :

[1] Gront D *et al. PLoS One* 2011 **6**: e23294 [PMID: 21887241]
[2] Roy A *et al. Nat Protoc.* 2010 **5**: 725 [PMID: 20360767]
[3] Bujnicki JM, *Chembiochem*. 2006 **7**: 19 [PMID: 16317788]
[4] Kihada D, *Protein Sci.* 2005 **14**:1955 [PMID: 15987894]
[5] Lin KL, *IEEE Eng Biol Mag*. 2009 **28**: 38 [PMID: 19622423]
[6] Roterman I *et al. J Theor Biol*. 2011 **283**: 60 [PMID: 21635900]
[7] Roterman I *et al. Int J Mol Sci*. 2011 **12**: 4850 [PMID: 21954329]
[8] Roterman I, *Biochimie*. 1995 **77**: 204 [PMID: 7647113]
[9] Brylinski M *et al. Bioinformatics* 2004 **20**: 199 [PMID: 14734311]
[10] Kalinowska B *et al. J Mol Model*. 2013 **19:** 4259 [PMID: 23812949]

Edited by P Kangueane
Citation: **Kalinowska *et al.*** Bioinformation 11(10): 486-488 (2015)