Ranking Beta Sheet Topologies of Proteins

Rasmus Fonseca[*], Glennie Helles[*] and Pawel Winter[*]

Abstract—**One of the challenges of protein structure prediction is to identify long-range interactions between amino acids. To reliably predict such interactions, we enumerate, score and rank *all* β-topologies (partitions of β-strands into sheets, orderings of strands within sheets and orientations of paired strands) of a given protein. We show that the β-topology corresponding to the native structure is, with high probability, among the top-ranked. Since full enumeration is very time-consuming, we also suggest a method to deal with proteins with many β-strands.**

The results reported in this paper are highly relevant for *ab initio* protein structure prediction methods based on decoy generation. The top-ranked β-topologies can be used to find initial conformations from which conformational searches can be started. They can also be used to filter decoys by removing those with poorly assembled β-sheets, and finally they can be relevant in contact prediction methods.

Keywords: beta-sheets, protein structure prediction, β-topology

1 Introduction

Predicting the tertiary structure of a protein from its amino acid sequence alone is known as the *protein structure prediction* (PSP) problem. It is one of the most important open problems of theoretical molecular biology. In particular, *ab initio* PSP (especially needed when a homologous sequence cannot be found in the protein data bank) poses a significant problem. One of the reasons why *ab initio* methods struggle is that the conformational space of most protein structure models increases exponentially with the length of the sequence. The complexity of the PSP problem can be reduced using auxiliary predictions such as secondary structures [1, 2, 3, 4], contact maps [5, 3, 6], structural alphabets [7, 8] and local structure predictions [9, 10]. However, all these predictions have a certain level of inaccuracy so they cannot be used to constrain the conformational space, only to guide the conformational search.

A *β-topology* is a partition of β-strands into ordered subsets (each corresponding to a β-sheet) together with the *β-pair* information (pairing of strands and their orientation) for each β-sheet. The order of β-strands within a β-sheet combined with the β-pair information is referred to as the *β-sheet topology*. If the correct β-topology could be predicted, it would, for instance, assist PSP methods to find the native structure [11, 12, 13, 14, 15].

One approach to predict the β-topology of a protein, in the following referred to as the *pair scoring method* [16], is to assign a pseudo-energy to every β-pair. The problem of determining the best β-topology is then formulated as a maximization problem in a complete graph where nodes correspond to β-strands and edge-weights correspond to the pseudo-energy of pairing two strands [12, 16, 17, 18, 15]. Another approach, referred to as the *topology scoring method*, is to enumerate all β-topologies, and to assign a score to each based on the entire β-topology [19, 20]. In general, the β-topology with highest score is assumed to correspond to the correct one [13]. The topology scoring method has also been used to filter decoy sets from Rosetta [19].

Our objective is not to predict the correct topology, but to generate a small set of β-topologies that will, with high probability, contain the correct one. We, therefore, enumerate all β-topologies and use the scoring methods from [16] and [19] to score and rank them. Our experiments show that for a large percentage of examined proteins, the correct β-topology can be found among the 10% top-ranked β-topologies using the pair scoring method (which outperforms the topology scoring method).

Enumerating all β-topologies is a problem for proteins with more than 7 β-strands due to combinatorial explosion. For such proteins, a subset of the β-topologies is enumerated. This subset is guaranteed to contain a β-topology which is consistent with the correct one, meaning that it has no β-pair which does not exist in the correct one. Such β-topologies can be found among the top 10% top-ranked and can also be found for larger proteins.

2 Methods

The following two subsections describe how a set of β-topologies is generated for a given protein, the first for proteins with 7 strands or less and the second for proteins with more strands. The third subsection describes how a score is calculated for each β-topology, and finally the datasets used in the experiments are described.

[*] {rfonseca, glennie, pawel}@diku.dk. Univ. of Copenhagen, Dept. of Computer Science. Universitetsparken 1, 2100 Copenhagen O, Denmark

Generating small β-topologies

The secondary structure specifies which amino acids are classified as helix, strand or coil. Continuous segments of strand-classified amino acids are simply referred to as strands.

If the secondary structure has 7 strands or less, all possible β-topologies can be generated by enumerating all valid pairings of strands. A valid pairing of strands is characterized by the following rules: Each strand is paired with one or two other strands and each pair of strands is either parallel or anti-parallel. Table 1 shows how many valid pairings exist in a protein with m strands.

This definition of a valid pairing corresponds largely to the definition of 'overall sheet configuration' used in [21] and 'β-sheet topology' from [16]. It is a representation of the β-topology that does not specify precisely which amino acids form hydrogen bonds in two strands, but it focuses on the overall configuration of the β-pairs in the protein.

m	2	3	4	5	6	7
	2	20	156	1744	23800	373008

Table 1: Number of valid β-topologies.

Generating larger β-topologies

If the secondary structure has 8 strands or more the set of β-topologies is generated the following way. First, a subset of six strands is chosen, and the 23800 corresponding β-topologies are added to the set. This process is repeated for all subsets of 6 strands. A total of

$$\binom{m}{6} \cdot 23800$$

β-topologies are therefore enumerated and scored for proteins with m strands. Table 2 shows this value for $m = 8, \ldots, 13$.

8	9	10	11	12	13
666 400	1 999 200	4 998 000	10 995 600	21 991 200	40 840 800

Table 2: Number of valid β-topologies.

The β-topologies in the final set will contain fewer strands than in the native β-topology. However, it can still be guaranteed that at least one β-topology will be very similar to the native β-topology.

To clarify how a β-topology is compared to the native, we introduce the notions of *native-respecting* and *native-matching* β-topologies. A β-topology is native-respecting if each β-pair corresponds to a β-pair in the native. A β-topology, B, is native-matching if it is native-respecting, and if each β-pair in the native furthermore corresponds to a β-pair in B (i.e., B respects the native and the native respects B). Figure 1 illustrates how β-topologies are compared to the native β-topology. For proteins with more than 7 strands the native-matching topology is never among the generated, but several native-respecting β-topologies will still be among them.

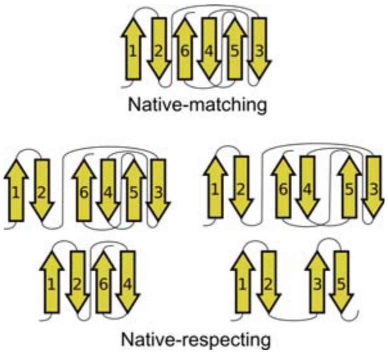

Figure 1: Five β-topologies for 1I8N.

Scoring β-topologies

Two methods of scoring β-topologies have been examined: The topology scoring method and the pair scoring method.

The *topology scoring method* [19], works for proteins with one β-sheet only. It assigns a probability to each β-sheet topology based on the following features: Number of strands, β-pairs, parallel β-pairs, parallel β-pairs with short loops (less than 10 amino acids), jumps (sequential strands that do not form β-pairs), jumps with short loops, the placement of the first strand (near the edge or the center of the sheet) and the helical status of the chain (either all-beta or alpha-beta).

In order to deal with proteins with more than one β-sheet, a more elaborate topology scoring function is needed. In [21], the probabilities of individual β-sheet topologies are combined with two more features, the number of sheets and the number of crossings (consecutive β-strands in different β-sheets), to assign a probability to the entire β-topology.

The *pair scoring method* [16] uses pseudo-energies between pairs of amino-acids in different strands. Neural networks are used to determine these pseudo-energies. The total pseudo-energy of a β-pair is calculated by finding an optimal alignment (either parallel or antiparallel) of the two strands using dynamic programming. The pseudo-energy of the β-pair is then the sum of pseudo-energies for the resulting amino acid pairs. Since a β-topology can be regarded as a set of β-pairs, we calculate the score of a β-topology as the average pseudo-energy of all β-pairs. This ensures that scores of β-topologies are comparable even when they differ in the number of β-pairs.

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

In [22] a third scoring method which is primarily based on hydrophobic packing is discussed. This method, however, is outperformed by both the topology scoring method and the pair scoring method, so the results are not mentioned here.

Datasets

To evaluate how good the scoring of β-topologies is, we generate two datasets. The first is made up of all the chains from PDBSelect25 2009 [23] that contain strands. This is 3305 out of 4423 chains total (75%). Not all the required parameters for the topology scoring method are available in [21], so the dataset is split into a training-set and a test-set (*PDB test-set*), of 161 randomly chosen chains containing between 2 and 7 strands. The training-set is used to learn the parameters in the topology scoring method.

A second test-set, the *CASP8 test-set*, is compiled from all the CASP8 [24, 25] targets that contain strands. This test-set has no guarantee to be as diverse as the PDB test-set, but it gives a good indication of the practical applicability of our method. At CASP8 there were 119 targets of which 13 contained no strands, so the CASP8 test-set consists of 106 protein chains that all have sheets. 53 of the these have between 2 and 7 strands and the majority of the rest contains between 8 and 12 strands.

3 Results and discussion

The primary tool for analyzing sets of β-topologies is a *rank-plot*. The rank-plot for a set of β-topologies shows the rank of each β-topology (x-axis) and its score (y-axis). The set is sorted by non-increasing score. The rank-plot is therefore a monotonically non-increasing curve (see Figure 2). The position of the native-matching β-topology is highlighted with a circle and native-respecting topologies are highlighted with crosses. The average and median rank of native-matching and native-respecting β-topologies will be the primary tool for reporting results. Since there can be more than one native-respecting topology, we only consider the highest ranked.

Ranking small β-topologies

For every protein in the PDB test-set, the secondary structure is extracted from the PDB file and then used to generate a set of β-topologies and the corresponding rank-plot. For 4 out of the 161 proteins in the PDB test-set, a native-matching β-topology was not among the generated β-topologies because one of their strands paired with more than two other strands.

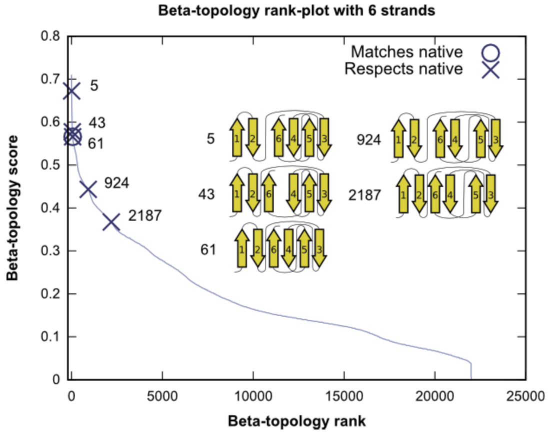

Figure 2: The rank-plot for the all β-topologies of the six-stranded protein 1I8N using the pair scoring method. The native-matching β-topology has rank 61, and the first native-respecting β-topology has rank 5.

The main question when considering the applicability of enumerating β-topologies is: How many of the top-ranked β-topologies does one have to consider before the native-matching is found? Figure 3 shows how many proteins (percentage) have the native-matching β-topology among the top-ranked. The figure illustrates this for both the topology scoring method (top) and the pair scoring method (bottom). Individual curves are generated for proteins containing the same number of strands. For 80% of all 6 stranded proteins it is sufficient to go through roughly 2230 of the top-ranked β-topologies when using the topology scoring method and 232 when using the pair scoring method. This implies that for a large fraction of proteins, enumerating just a relatively small number (hundreds) of β-topologies, results in a set that has a good chance to contain the native-matching β-topology.

The topology scoring method performs well, and at times better, compared to the pair scoring method for proteins with 4 strands or fewer. For proteins with more strands, however, the pair scoring method significantly outperforms the topology scoring method. Therefore, all of the remaining experiments are performed using the pair scoring method.

Table 3 shows more statistics for the rank of the native-matching β-topology using the pair scoring method.

Strands	2	3	4	5	6	7
Proteins	26	33	26	28	27	20
Avg. NM rank	1.08	2.55	4.77	104	213	8850
Median NM rank	1	2	3	49	69	905
Avg. BNR rank	1.08	2.55	1.69	54.3	104	7534
Median BNR rank	1	2	1	13	7	41

Table 3: Average and median ranks of native-matching (NM) and best native-respecting (BNR) β-topologies in PDB test-set.

Ranking larger β-topologies

The native β-topology is among the enumerated for 45% (53 out of 119) of the proteins at CASP8, assuming the secondary structure is predicted correctly. Table 4 (top)

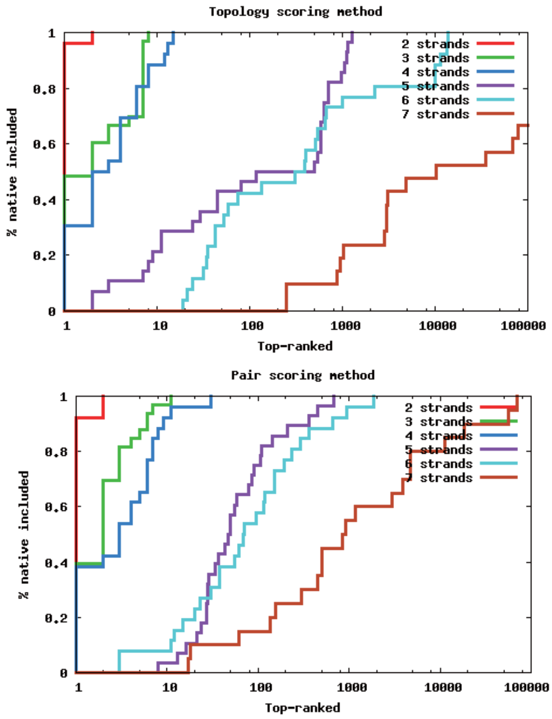

Figure 3: Percentage of native-matching β-topologies among the top-ranked potential topologies using the pair scoring method and the topology scoring method. The x-axis shows the number of top-ranked topologies on a logarithmic scale. The pair scoring method outperforms the topology scoring method for chains with 5 to 7 strands.

shows statistics for the rank of the native-matching and native-respecting β-topologies. Comparing these numbers to those for the PDB test-set in Table 3, it is observed that the ranks of the native-matching β-topologies are higher for the proteins with 6 strands, but notably lower for those with 5 and 7 strands. By comparing the median ranks to the total number of valid β-topologies, shown in Table 1, it is observed that, for a vast majority of the proteins, the native-matching β-topology is among the 10% highest ranked potential β-topologies.

Strands	2	3	4	5	6	7
Proteins	2	4	4	13	11	16
Avg. NM rank	1.5	2.0	5	73	872	4240
Median NM rank	1.5	2.0	2	28	149	768
Avg. BNR rank	1.5	2.0	1.25	31	525	1033
Median BNR rank	1.5	2.0	1.25	4	3	9
Strands	8	9	10	11	12	
Proteins	11	2	7	5	19	
Avg. BNR rank	7628	213	626	6982	11821	
Median BNR rank	59	213	211	464	1582	

Table 4: Average and median ranks of native-matching (NM) and best native-respecting (BNR) β-topologies in CASP8 test-set. For proteins with more than 7 strands, a subset of β-topologies is generated, which is guaranteed to contain a native-respecting β-topology

The ranks of the best native-respecting β-topologies are typically significantly lower than the ranks of the native-respecting. Furthermore, the median ranks are much lower than the average ranks, which indicates that for a majority of proteins the native-respecting β-topology is among the top-ranked, but for a few, the rank is very big.

If, for instance, only the 200 highest ranked β-topologies were considered for each protein in the CASP8 test-set, then the native-respecting β-topology would be among these for 50% of the proteins and the native-matching would be among them for 31%.

4 Conclusions and future work

We presented a method to enumerate β-topologies such that it is guaranteed that a native-respecting β-topology is always among the generated. Furthermore, for proteins with 7 strands or less, a native-matching topology is also guaranteed to be among those generated. The enumerated β-topologies have been scored and ranked using two different scoring methods: The pair scoring method and the topology scoring method. The pair scoring method is shown to outperform the topology scoring method. It is shown that the native-matching β-topology is among the top 10% highest ranked β-topologies, with native-respecting topologies frequently found among the very highest ranked.

There are a number of ways to improve and extend this work. First of all, a better method for scoring β-topologies could be developed by combining the topology scoring method [19] and the pair scoring method [16]. Features and concepts from other sources such as [26, 20, 17, 15] could be used as well. Furthermore, disulphide bindings could be incorporated into the model. This could significantly limit the number of β-topologies for cysteine-containing proteins.

It is assumed that the secondary structure can be predicted correctly. This assumption does not always hold. Particularly the placement of strands is important when enumerating β-topologies. To ensure that at least one β-topology is native-respecting, it should be investigated how the accuracy of strand predictions could be improved.

Finally, the natural extension of this work is to design a PSP method that can use the top-ranked β-topologies to constrain the conformational search and generate high quality protein structure decoys. [14] presents an interesting approach that, using the entire set of β-topologies from [19] and inverse kinematics, can generate high quality decoys. Similar methods, using e.g., only the 200 top-ranked β-topologies, can run longer experiments on each β-topology and possibly give better results for proteins with many strands.

References

[1] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.

[2] M. Ouali and R. D. King. Cascaded multiple classifiers for secondary structure prediction. *Prot. Sci.*, 9:1162–1176, 2000.

[3] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi. Scratch: a protein structure and structural feature prediction server. *Nucl. Acids Res.*, 33:W72–W76, 2005.

[4] C. Cole, J. D. Barber, and G. J. Barton. The Jpred 3 secondary structure prediction server. *Nucl. Acids Res.*, 36:W197–W201, 2008.

[5] R. M. MacCallum. Striped sheets and protein contact prediction. *Bioinformatics*, 20 Suppl 1, 2004.

[6] A. N. Tegge, Z. Wang, J. Eickholt, and J. Cheng. NNcon: Improved protein contact map prediction using 2D-recursive neural networks. *Nucl. Acids Res.*, 37(37):W315–W318, 2009.

[7] C. Etchebest, C. Benros, S. Hazout, and A. G. de Brevern. A structural alphabet for local protein structures: improved prediction methods. *Proteins*, 59:810–827, 2005.

[8] M. Tyagi, A. Bornot, B. Offmann, and A. G. de Brevern. Protein short loop prediction in terms of a structural alphabet. *Comput. Biol. Chem.*, 33:329–333, 2009.

[9] O. Zimmermann and U. H. E. Hansmann. Support vector machines for prediction of dihedral angle regions. *Bioinformatics*, 22:3009–3015, 2006.

[10] G. Helles and R. Fonseca. Predicting dihedral angle probability distributions for protein coil residues from primary sequence using neural networks. *BMC Bioinformatics*, 10:338, 2009.

[11] Y. Cui, R. S. Chen, and W. H. Wong. Protein folding simulation with genetic algorithm and supersecondary structure constraints. *Proteins*, 31:247–257, 1998.

[12] J. L. Klepeis and C. A. Floudas. ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.*, 85:2119–2146, 2003.

[13] G. Porwal, S. Jain, S. D. Babu, D. Singh, H. Nanavati, and S. Noronha. Protein structure prediction aided by geometrical and probabilistic constraints. *J. Comput. Chem.*, 28:1943–1952, 2007.

[14] N. Max, C. Hu, O. Kreylos, and S. Crivelli. BuildBeta-A system for automatically constructing beta sheets. *Proteins*, 78:559–574, 2009.

[15] R. Rajgaria, Y. Wei, and C. A. Floudas. Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD. *Proteins*, Early View, 2010.

[16] J. Cheng and P. Baldi. Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, 21 Suppl 1:75–84, 2005.

[17] J. Jeong, P. Berman, and T. M. Przytycka. Improving strand pairing prediction through exploring folding cooperativity. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 5:484–91, 2008.

[18] M. Lippi and P. Frasconi. Prediction of protein beta-residue contacts by Markov logic networks with grounding-specific weights. *Bioinformatics*, 25:2326–33, 2009.

[19] I. Ruczinski, C. Kooperberg, R. Bonneau, and D. Baker. Distributions of beta sheets in proteins with application to structure prediction. *Proteins*, 48:85–97, 2002.

[20] A. S. Fokas, I. M. Gelfand, and A. E. Kister. Prediction of the structural motifs of sandwich proteins. *Proc. Nat. Acad. Sci. USA*, 101:16780–16783, 2004.

[21] I. Ruczinski. *Logic regression and statistical issues related to the protein folding problem.* PhD thesis, Univ. of Washington, 2002.

[22] J. L. Klepeis and C. A. Floudas. Prediction of beta-sheet topology and disulfide bridges in polypeptides. *J. Comput. Chem.*, 24:191–208, 2003.

[23] S. Griep and U. Hobohm. PDBselect 1992-2009 and PDBfilter-select. *Nucl. Acids Res.*, 38:D318–319, 2010.

[24] S. Shi, J. Pei, R. I. Sadreyev, L. N. Kinch, I. Majumdar, J. Tong, H. Cheng, B. Kim, and N. V. Grishin. Analysis of CASP8 targets, predictions and assessment methods. *Database*, 2009(0):bap003–, April 2009.

[25] M. L. Tress, I. Ezkurdia, and J. S. Richardson. Target domain definition and classification in CASP8. *Proteins*, 77 Suppl 9(S9):10–17, 2009.

[26] J. A. Siepen, S. E. Radford, and D. R. Westhead. Beta edge strands in protein structure prediction and aggregation. *Protein Sci*, 12(10):2348–59, 2003.