J Math Model Algor (2011) 10:357–369
DOI 10.1007/s10852-011-9162-4

Ranking Beta Sheet Topologies with Applications to Protein Structure Prediction

Rasmus Fonseca · Glennie Helles · Pawel Winter

Received: 31 August 2010 / Accepted: 31 August 2011 / Published online: 21 September 2011
© Springer Science+Business Media B.V. 2011

Abstract One reason why ab initio protein structure predictors do not perform very well is their inability to reliably identify long-range interactions between amino acids. To achieve reliable long-range interactions, *all* potential pairings of β-strands (β-topologies) of a given protein are enumerated, including the native β-topology. Two very different β-topology scoring methods from the literature are then used to rank all potential β-topologies. This has not previously been attempted for any scoring method. The main result of this paper is a justification that one of the scoring methods, in particular, consistently top-ranks native β-topologies. Since the number of potential β-topologies grows exponentially with the number of β-strands, it is unrealistic to expect that all potential β-topologies can be enumerated for large proteins. The second result of this paper is an enumeration scheme of a *subset* of β-topologies. It is shown that native-consistent β-topologies often are among the top-ranked β-topologies of this subset. The presence of the native or native-consistent β-topologies in the subset of enumerated potential β-topologies relies heavily on the correct identification of β-strands. The third contribution of this paper is a method to deal with the inaccuracies of secondary structure predictors when enumerating potential β-topologies. The results reported in this paper are highly relevant for ab initio protein structure prediction methods based on decoy generation. They indicate that decoy generation can be heavily constrained using top-ranked β-topologies as they are very likely to contain native or native-consistent β-topologies.

Keywords Beta-sheets · Protein structure prediction · Topology

R. Fonseca · G. Helles · P. Winter (✉)
Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
e-mail: pawel@diku.dk

R. Fonseca
e-mail: rfonseca@diku.dk

G. Helles
e-mail: glennie@diku.dk

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 similar amino acid sequence with known structure cannot be found in the protein database) 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 primary sequence. The complexity of the PSP problem can be reduced using auxiliary predictions such as secondary structures [2, 3, 9], contact maps [2, 12, 20] or local structure predictions [7, 21]. However, all these predictions have a certain level of inaccuracy so they cannot be used to constrain the conformational space, only to guide the search.

The native *β-topology* of a protein is a partition of β-strands into ordered subsets (each corresponding to a β-sheet) together with the *β-pair* information (indices of paired strands and their orientation in β-sheets)[1] The order of β-strands within a single β-sheet combined with the β-pair information is referred to as the *β-sheet topology*. If the native β-topology could be correctly predicted, it would reduce the search space of PSP and greatly improve the quality of the generated solutions [4, 10, 13, 15, 16]. Furthermore, some PSP methods, such as BuildBeta [13], cleverly use the spatial constraints that a β-topology supplies and can generate a reasonable structure in as little as 10 seconds.

The *pair scoring method* [1] identifies a good β-topology of a protein by assigning 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. The problem is to cover all vertices by disjoint paths (corresponding to β-sheets) and cycles (corresponding to β-barrels). Several other variants of this approach have been suggested [8, 10, 11, 16].

The *topology scoring method* [18] enumerates all *β-topologies*, and assigns a score to each based on properties of the entire β-topology. This can be properties such as the number of hairpin turns and parallel β-pairs. In general, the β-topology with highest score is assumed to correspond to the native [15]. The topology scoring method is also used to filter decoy sets from Rosetta [18].

Since the correct β-topology cannot be predicted accurately using either of these methods, we suggest a different approach: All β-topologies are enumerated and the pair scoring method and the topology scoring method are used to score and rank them. Our experiments show that for a large percentage of examined proteins, the native β-topology can be found among the 10% top-ranked β-topologies using the pair scoring method (which outperforms the topology scoring method). An often used step when solving the PSP problem is to generate a set of decoy structures. Using each of the ranked β-topologies as a constraint (one at a time), a set of decoy structures can be constructed. At least one of these decoy structures will be

[1] It is assumed here that a β-strand can have at most two partners. This is often true, but β-strands with more than two partners also exist; such β-strands cannot be dealt with by our method.

generated using the native β-topology, and will therefore have a very high quality. In this study the focus is on the enumeration and ranking of β-topologies, not on generating decoys.

There are three serious problems with the suggested approach. First of all, the correct secondary structure has to be known. One solution to this is to use predicted secondary structures. This leads to the second problem; secondary structure predictors are not always fully reliable. They sometimes over- or under-predict β-strands. In such cases, the native β-topology will not necessarily be among those enumerated. Thirdly, even if the prediction of β-strands is correct, the number of β-strands may be so large that the combinatorial explosion will make it impossible to enumerate all β-topologies. In fact, such combinatorial explosion occurs already when eight β-strands are present.

In order to deal with these problems, the notion of a *strand assignment* is introduced. A strand assignment is a set of non-overlapping intervals that specify which parts of the chain are classified as β-strands. One of the best secondary structure predictors, PSIPRED [9], assigns to each amino acid the probability of it being in a strand (pE-levels), helix and in a coil. Amino acids having pE-levels higher than both helix- and coil probabilities are classified as belonging to β-strands. This results in the *predicted strand assignment*. The main reasons why predicted strand assignments differ from the correct ones are over- and under-predictions of strands. A typical example of under-prediction of strands is shown in Fig. 1.

Since the correct strand assignment cannot always be predicted accurately, we suggest a different approach. Using the pE-levels from PSIPRED, *candidate strands* are suggested and potential strand assignments are enumerated and ranked as described in the Methods section. The problem of combinatorial explosion is dealt with by introducing two limitations when generating the set of potential strand assignments. First, only up to 15 candidate strands are considered in the enumeration procedure. Second, only potential strand assignments with up to seven strands are generated.

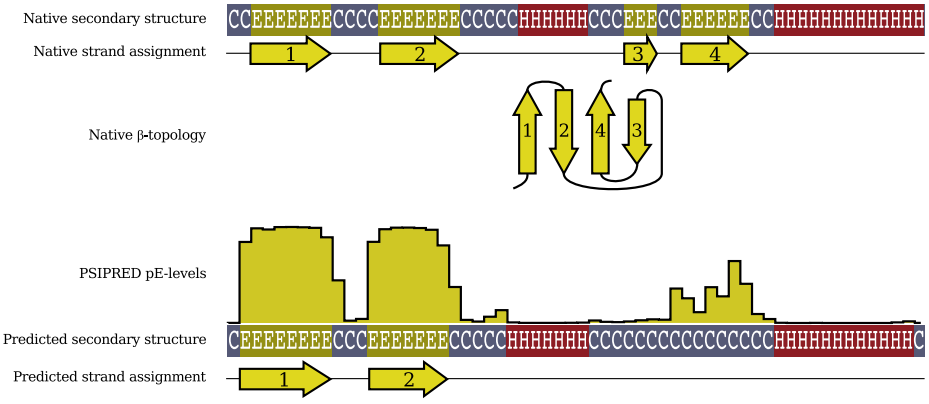

Fig. 1 Comparison of native and predicted strand assignment for the second domain in 3DEV. This example is from CASP8 and is a typical example of β-strand under-prediction. PSIPRED's pE-levels, however, still indicate the presence of a possible strand where the fourth strand should be (although coil probabilities were higher in this region)

Consequently, there will be proteins with eight or more strands whose native β-topologies cannot be generated. However, enumerating potential strand assignments and β-topologies is still relevant for such proteins. To illustrate this, the concepts of *native-respecting strand assignment* and *native-respecting β-topology* are defined. Every β-strand in a native-respecting strand assignment is present in the native strand assignment as well (though the native strand assignment may have more strands). Similarly, every β-pair in a native-respecting β-topology is present in the native β-topology as well (though the native β-topology may contain more β-pairs). For proteins with many strands, a native-respecting strand assignment with up to seven strands can always be found among the potential strand assignments. For most of these, a native-respecting β-topology will be generated. Even though a native-respecting β-topology does not impose as strong a constraint on the PSP problem as a native β-topology itself, it is still a valid constraint that can reduce the search space significantly.

The results reported in this paper are highly relevant for the PSP methods where decoy generation can be constrained or filtered by top-ranked β-topologies. It can also be used in more elaborate contact prediction methods [2, 16, 20].

2 Methods

In the first two subsections the methods for generating potential β-topologies and for calculating their scores are described. Next, it is described how potential strand assignments are generated and how scores are assigned to each of them. The last two subsections describe how to compare both strand assignments and β-topologies and which data sets are used to assess the methods.

2.1 Generating Potential β-topologies

β-strands are numbered $1, 2 \ldots m$ according to the order they appear in the chain. A potential β-topology generated from a strand assignment with m strands is represented using a binary *β-topology-matrix*, $[a_{ij}]_{m\times m}$. Strands i and j form a parallel pair iff $(a_{ij} = 1) \wedge (i > j)$. They form an antiparallel pair iff $(a_{ij} = 1) \wedge (i < j)$. Entries with 1 in the upper (respectively lower) triangle of the matrix therefore represent antiparallel (respectively parallel) pairs. All other entries are 0. A valid β-topology-matrix is characterized by the following three rules: No strand is paired to itself, no pair of strands is paired both parallel and antiparallel and every strand has one or two partners. Given m strands, the complete set of valid β-topology-matrices is generated beginning with the 0-matrix and adding 1's starting at the top row, from left to right (backtracking when necessary).

Table 1 shows the number of valid potential β-topologies, $V(m)$, for up to seven strands. For a single β-sheet with m strands there are $m!/2$ possible orderings of the strands and $2^m/2$ possible combinations of orientations (ignoring symmetric orderings and orientations). This gives a total of $m! \times 2^{m-2}$ possible β-topologies that contain only a single sheet (not counting barrels). Since this number is a lower bound on $V(m)$, it is clear that $V(m)$ grows exponentially with m. For this reason, it is infeasible to enumerate all potential β-topologies for $m \geq 8$.

Table 1 Number of valid β-topology-matrices, $V(m)$, and number of β-topology-matrices that need to be enumerated to include the native, $B(m)$

m	2	3	4	5	6	7
$V(m)$	2	20	156	1,744	23,800	373,008
$B(m)$	2	11	30	700	1,900	70,000

$V(m)$ is determined computationally. $B(m)$ is a result of the experiments shown later in Fig. 5

2.2 Assigning Scores to β-topologies

Two methods for assigning scores to β-topologies have been examined. The *topology scoring method* assigns a probability to each β-sheet topology based on several more or less complicated topological features [17, 18]. This probability is used as the score in the topology scoring method. The topological features include among other things the number of sheets, the number of times a chain crosses from one sheet to another as well as the number of parallel and anti-parallel β-pairs. This method was reimplemented and the parameters obtained from a training-set of proteins specified in the next section.

The *pair scoring method* uses a feed-forward neural network to obtain probabilities of pairing two amino acids. Dynamic programming is then used to optimally align pairs of β-strands in the best possible way (both parallel and antiparallel alignments are included). The score of each alignment is the sum of pairing probabilities between amino acids. The pseudo-energy of pairing two strands is the maximum over all such alignments [1]. A score is assigned to a β-topology by taking the average of pseudo-energies of all its β-pairs. The neural networks were downloaded from the authors homepage.

2.3 Generating Potential Strand Assignments

A strand assignment is defined as a set of m non-overlapping intervals, $\{(s_1, e_1) \ldots (s_m, e_m)\}$, indicating which parts of the chain are β-strands. To ensure that the β-topology of a protein's native structure can be represented, it is important that each β-strand is identified correctly. PSIPRED can be used to predict the placement of strands. It produces three probability levels for each amino acid, $a = 1, 2, \ldots, n$, indicating the probability of a being either helix (pH_a), strand (pE_a) or coil (pL_a). If $pE_a > \max\{pH_a, pL_a\}$ then a is classified as belonging to a strand. This method often fails to predict a strand entirely or predicts a strand where there is none. However, when PSIPRED fails to predict a strand there is often a hilltop (a segment with local minima at both ends) in the pE-levels (see Fig. 1). A set of *candidate strands*, representing possible placements of strands, are therefore generated around hilltops in the pE-plot (see Fig. 2a).

Similar to a strand in a strand assignment, the candidate strands, $i = 1, 2, \ldots, m_c$, are defined by the indices of their first and last amino acids: (s_i, e_i). A potential strand assignment is generated from a subset of candidate strands. All the potential strand assignments are generated by using all possible subsets of candidate strands. Potential strand assignments with 1 or 0 strands are omitted, as valid β-topologies must have at least two strands. To avoid the combinatorial explosion, potential strand assignments with eight strands or more are omitted as well. Figure 2b shows all possible strand assignments that can be generated using the coil and helix probability

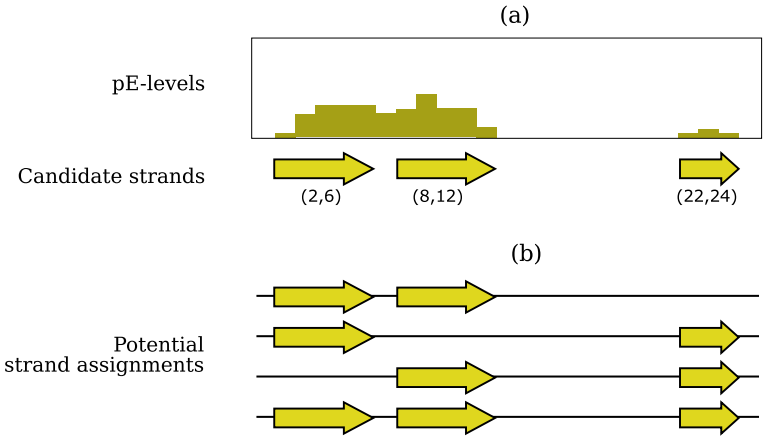

Fig. 2 **a** Probability levels for β-strands (pE-levels of 1ZEC) predicted using PSIPRED. Three candidate strands are identified from the hilltops. **b** All potential strand assignments for 1ZEC

levels and candidate strands from Fig. 2a. The total number of potential strand assignments that are generated for a protein with m_c candidate strands is

$$\sum_{i=2}^{m_c} \binom{m_c}{i} = 2^{m_c} - m_c - 1 \tag{1}$$

2.4 Assigning Scores to Potential Strand Assignments

The pE-levels are used to calculate a score for every potential strand assignment. The average pE value for each strand is calculated as

$$\langle pE \rangle_i = \frac{1}{l_i} \sum_{a=s_i}^{e_i} pE_a \tag{2}$$

where $l_i = (e_i + 1) - s_i$. The score of a strand assignment is then the average of $\langle pE \rangle_i$ for all i, i.e.,

$$\langle pE \rangle = \frac{1}{m} \sum_{i=1}^{m} \langle pE \rangle_i \tag{3}$$

By using averages it is ensured that strand assignments with different number of strands have comparable scores.

2.5 Comparing both Strand Assignments and β-topologies

Two strands, i and j, from different strand assignments are said to overlap iff any part of the interval $[s_i, e_i]$ overlaps $[s_j, e_j]$. Two strand assignments *match* iff there exists a pairing of every strand in the first with every strand in the second such that each pair of strands overlap. A strand assignment is furthermore said to *respect* another strand assignment iff there exists a pairing of every strand in the first with a subset of strands

in the second such that each pair of strands overlaps. This definition will prove useful because potential strand assignments that respect the native strand assignment can be considered 'almost native'. Figures 3 and 4a both give examples of strand assignments that respect another strand assignment.

A β-topology is given by a strand assignment with m strands and a valid β-topology-matrix, $[a_{ij}]_{m\times m}$ specifying which strands are paired. Two β-topologies *match* if their strand assignments match and they have identical β-topologies. Note that if the strand assignments match then the β-topologies will always be of the same dimension. One β-topology, with matrix $[a_{ij}]$, is said to *respect* another, with matrix $[a'_{kl}]$, iff its strand assignment respects that of the second and $(a_{ij} = 1) \Rightarrow (a'_{kl} = 1)$ where i and k are indices of strands that overlap, and j and l are indices of strands that overlap. Figure 3 illustrates how strand assignments and β-topologies are compared. Figure 4b also shows four β-topologies that respect the native.

2.6 Data Sets

For evaluating the quality of the scoring of strand assignments and β-topologies, we generate three data sets. The first two are made up of chains from PDBSelect25 2009 [6] that contain strands. There are 3,305 of these (out of 4,423 chains in total). The topology scoring method is a probabilistic model that has a set of parameters extracted from PDB-files. Not all these parameters are given in [17] so a training-set is needed for the topology scoring method. The proteins from PDBSelect25 2009 are therefore split into a training-set and a test-set (the *PDB test-set*. The *PDB test-set* consists of 161 randomly chosen chains with between two and seven strands and the training-set is the rest of the proteins.

The third data set, the *CASP8 test-set*, is compiled from all the CASP8 [14] targets that contain β-strands. This test-set has no guarantee to be as diverse as PDBSelect25 but gives an indication of the practical applicability of our method. At CASP8 there were 119 targets, but 13 contained no strands, so the CASP8 test-set consists of 106 protein chains that all have β-sheets. 53 of the these have between two and seven strands and the majority of the rest contains between 8 and 12 strands.

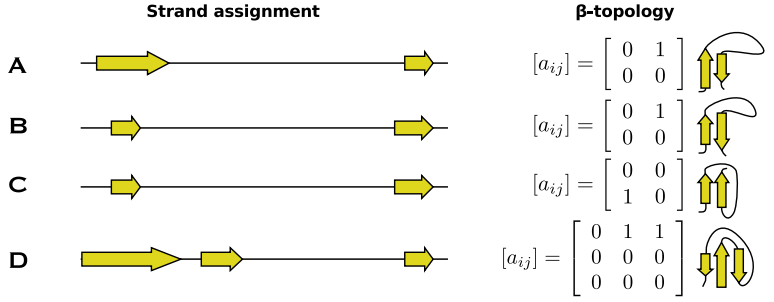

Fig. 3 Examples of comparing strand assignments and β-topologies. Four strand assignments are shown in the *left column*. Strand assignments A, B and C match each other and they all respect D. The *right column* shows examples of β-topologies for each strand assignment. The β-topologies A and B match each other. They neither respect nor match C but they both respect D

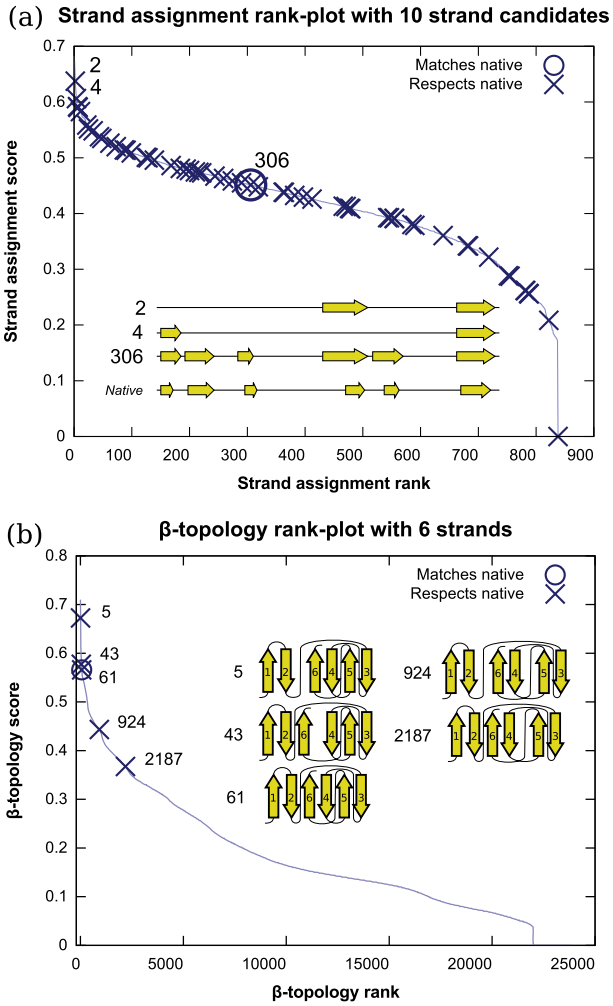

Fig. 4 **a** The strand assignment rank-plot for the six-stranded protein 1I8N. The native strand assignment has rank 306. However, potential strand assignments with ranks as low as 2 and 4 respects the native, and will likely be used to generate β-topologies that respects the native. **b** The β-topology rank-plot for the six-stranded protein 1I8N. The native strand assignment has been used, and the scores are calculated using the pair scoring method. The native β-topology has rank 61, but the β-topology with rank 5 respects the native, and thus provides a constraint that is nearly as good as the native. All topologies that either match or respect the native are highlighted and shown inside the plot

3 Results and Discussion

Given a protein, the *rank-plot* of potential strand assignments illustrates the rank of each strand assignment plotted against its score, as defined in Section 2.4. The rank-plot is therefore a monotonically non-increasing curve as shown in Fig. 4a. The first potential strand assignment that matches the native strand assignment (the *native-matching strand assignment*) is highlighted using a circle. Potential strand assignments that respect the native (*native-respecting strand assignments*) are highlighted using crosses.

Given a protein and a strand assignment, the rank-plot of potential β-topologies illustrates the rank of each β-topology plotted against its score, as defined in Section 2.2 (See Fig. 4b). Only a single β-topology can match the native and only β-topologies with two sheets or more (more than three strands) can respect (and not

match) the native β-topology. These β-topologies are referred to as *native-matching β-topologies* and *native-respecting β-topologies*, respectively.

The average and median rank of native-matching and native-respecting strand assignments and β-topologies will be the primary tool for reporting results.

3.1 Ranking β-topologies Using Native Strand Assignments

An important question when considering the practical applicability of enumerating β-topologies is: How many of the top-ranked β-topologies does one have to enumerate, on average, before the native-matching is found? Using the PDB test-set, Fig. 5 shows how many proteins (percentage) have the native-matching β-topology among the top-ranked. The figure illustrates this for both scoring methods—the topology scoring method and the pair scoring method. Individual curves are generated for proteins containing the same number of strands. For example, for 80% of all 6 stranded proteins it is sufficient to go through roughly 2,230 of the top-ranked β-topologies (out of 23,800 in total) when using the topology scoring method and 232 when using the pair scoring method. This implies that for a large fraction of proteins going through just a relatively small number (hundreds) of β-topologies

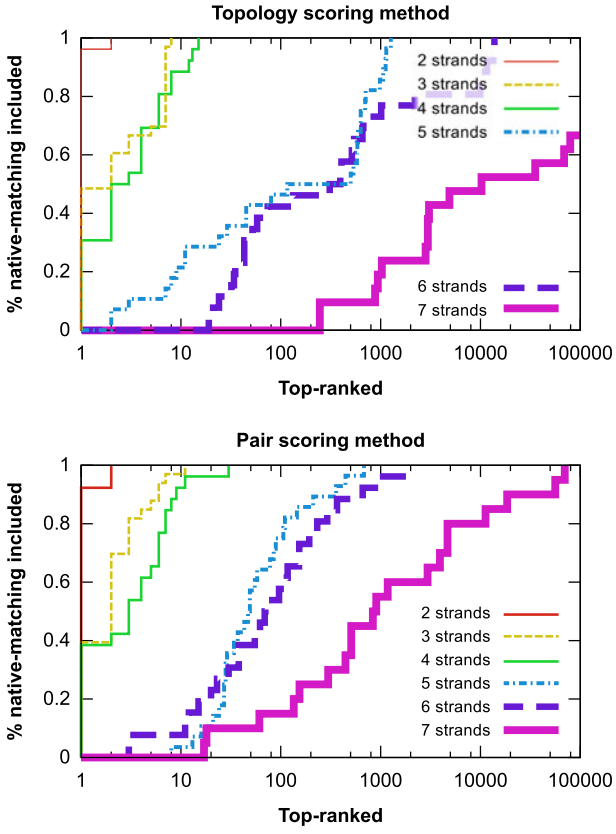

Fig. 5 Percentage of native-matching β-topologies among the top-ranked potential topologies using the topology scoring method and the pair scoring method. The x-axis shows the number of top-ranked topologies on a logarithmic scale

Table 2 Average and median ranks of native-matching β-topologies in PDB test-set (pair scoring method)

Strands	2	3	4	5	6	7
Proteins	26	33	26	28	27	20
Avg. rank	1.08	2.55	4.77	104	213	8,850
Median rank	1	2	3	49	69	905

gives a constraint for the PSP problem that can significantly reduce the size of the conformational search space.

The topology scoring method performs equally well as the pair scoring method for proteins with up to four strands. For proteins with more β-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 2 shows statistics for the rank of the native-matching β-topology. 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 β-topology is among the 10% highest ranked potential β-topologies.

3.2 Ranking Potential Strand Assignments

PSIPRED [9] was used to generate pH, pE and pL-levels for all proteins in the PDB test-set. From the pE-levels, candidate strands are identified and potential strand assignments generated. For every potential strand assignment, a score is calculated using the pE-levels, and a rank-plot is generated for every protein (161 in total). The number of potential strand assignments that one has to enumerate before the native-matching strand assignment is encountered is shown in Fig. 6. The red curve converges on ≈81% after 3,000 potential strand assignments (out of approximately 15,000 on average for each protein), which indicates that for only 19% of the proteins in the PDB test-set, no potential strand assignment that matches the native is generated. The typical reason for this is that PSIPRED fails to identify one or more strands. For a majority of the proteins, however, it is enough to enumerate

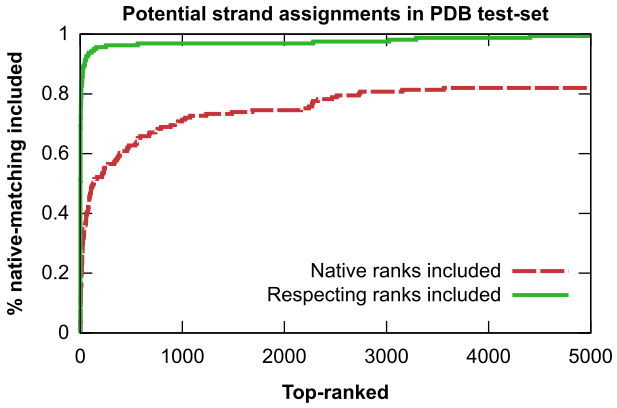

Fig. 6 Percentage of proteins for which the native-matching strand assignment (*red curve*) and native-respecting (*purple curve*) is included among the top-ranked strand assignments

less than 1,000 potential strand assignments. The proteins in the PDB test set have 15,369 potential strand assignments on average. It is therefore observed that for a majority of the proteins, a native-matching strand assignment can be found among the top 7% of the generated strand assignments.

Using only the top-200 ranked potential strand assignments, a native-respecting strand assignment can be found for more than 95% of the proteins.

3.3 Combining Potential Strand Assignments and Potential β-topologies

This subsection seeks to determine the applicability of enumerating both potential strand assignments and potential β-topologies. Since the CASP8 test-set contains proteins that state-of-the-art PSP methods are benchmarked on, we will use this test-set. The combinatorial explosion of β-topologies is dealt with by mainly looking for the native-respecting β-topologies. This ensures that the experiment can be run on proteins with more than seven strands. The experiment seeks to determine how many β-topologies it is necessary to enumerate to find a native-respecting β-topology. It does so without assuming that the native strand assignment is known in advance. Given a potential strand assignment with m strands, $B(m)$ is defined as the number of top-ranked β-topologies which it is necessary to enumerate before the native-matching β-topology is included. The values of $B(m)$ are read off the curves in Fig. 5 and shown in the third row of Table 1. For each of the 106 proteins in the CASP8 test-set, the following experiment is performed: The potential strand assignments are generated, scored and ranked. Starting from the top-ranked potential strand assignment, with m_1 strands, all its β-topologies are generated, scored and ranked. The $B(m_1)$ top-ranked β-topologies are examined. This process is repeated for the lower-ranked strand assignments until the first native-respecting β-topology is encountered. The number of examined β-topologies is then reported. The average and median of these numbers are shown in the second row of Table 3. There is a huge difference between the average number of β-topologies that has to be examined ($\approx$80,000) and the median (44). This indicates that only a limited number of outliers needs to have many β-topologies examined. For a majority of the proteins, less than 50 β-topologies need to be examined before a native-respecting β-topology is found. In many cases, however, this first native-respecting β-topology will only have two strands. This does not provide a very strong constraint on the PSP problem. The experiment above is therefore repeated, but for potential strand assignments

Table 3 Combining potential strand assignments and β-topologies for the CASP8 test-set

Min. m	μ(SA)	$\mu_{\frac{1}{2}}$(SA)	μ(β-sum)	$\mu_{\frac{1}{2}}$(β-sum)
2	102	7	80,634	44
3	271	41	255,956	9,725
4	337	48	361,101	23,586
5	503	198	691,917	242,925

μ(SA) and $\mu_{\frac{1}{2}}$(SA) denotes the average and median rank of the first native-respecting strand assignment from which a native-respecting β-topology can be generated. μ(β-sum) and $\mu_{\frac{1}{2}}$(β-sum) denote the average and median number of β-topologies that have to be examined before a native-respecting β-topology is located. For each row, only topologies with at least 'Min m' strands are considered

with at least three, four and five strands. As a result, approximately 10,000, 22,000 and 240,000 β-topologies, respectively, have to be enumerated for a majority of the proteins before a native-respecting β-topology is found. Although these numbers are high, it is still realistic to generate that many decoys in a reasonable PSP method.

The focus of this subsection has, so far, been solely on native-respecting β-topologies because these can be found for proteins containing any number of strands. The above experiment is repeated for proteins with seven strands or fewer and the number of examined topologies is reported only when the native-matching β-topology is examined. The average and median number of topologies that have to be examined are around 13,900,000 and 4,800,000 and this can only be done for 33 out of the 53 proteins (62%). While these numbers are rather large, any PSP method that efficiently takes advantage of β-topologies, such as [13], will be able to go through that many topologies in a limited amount of time. Furthermore, for a few proteins (3DFD, 3DED, 3DEX, 2KDM and 3DO8), the native β-topology is found after only examining a few thousand β-topologies.

4 Conclusions and Future Work

We have presented a method to enumerate and rank potential β-topologies for proteins with up to seven strands 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.

If the correct secondary structure assignment (strand assignment) is not known in advance, the output from PSIPRED is used to generate and rank potential strand assignments with up to seven strands. The results show that the native strand assignment is among the top 7% highest ranked strand assignments for the majority of proteins. Potential strand assignments are then used to generate potential β-topologies. Given the correct strand assignment, it is shown that the native β-topology is among the top 10% highest ranked β-topologies, with native-respecting topologies frequently found among the very highest ranked. Using predicted strand assignments, non-trivial (more than two β-strands) native-respecting β-topologies can be found within the top 10,000 highest ranked β-topologies.

There is 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 [18] and the pair scoring method [1]. Features and concepts from other sources such as [5, 8, 16, 19] could be used as well. Furthermore, disulphide bonds could be incorporated into the model. This could significantly limit the number of β-topologies for cysteine-containing proteins.

The results indicate that a relatively large number of strand assignments has to be examined before the native strand assignment is located. The method used for scoring potential strand assignments is very simple. A huge improvement of the results could be achieved by refining the scoring of potential strand assignments using, for instance, machine learning methods like neural networks or support vector machines. Furthermore, a secondary structure predictor that overpredicts strands could also help to ensure that the native strand assignment is among the potential strand assignments for more than 81% of the proteins.

Finally, the very important and 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.

Acknowledgement We thank Marcus Brazil for his valuable comments and suggestions.

References

1. Cheng, J. Baldi, P.: Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. Bioinformatics **21**(Suppl 1), 75–84 (2005)
2. Cheng, J., Randall, A.Z., Sweredoski, M.J., Baldi, P.: SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. **33**, W72–W76 (2005)
3. Cole, C., Barber, J.D., Barton, G.J.: The Jpred 3 secondary structure prediction server. Nucleic Acids Res. **36**, W197–W201 (2008)
4. Cui, Y., Chen, R.S., Wong, W.H.: Protein folding simulation with genetic algorithm and super-secondary structure constraints. Proteins **31**, 247–257 (1998)
5. Fokas, A.S., Gelfand, I.M., Kister, A.E.: Prediction of the structural motifs of sandwich proteins. Proc. Natl. Acad. Sci. USA **101**, 16780–16783 (2004)
6. Griep, S. Hobohm, U.: PDBselect 1992–2009 and PDBfilter-select. Nucleic Acids Res. **38**, D318–319 (2010)
7. Helles, G. Fonseca, R.: Predicting dihedral angle probability distributions for protein coil residues from primary sequence using neural networks. BMC Bioinformatics **10**, 338 (2009)
8. Jeong, J., Berman, P., Przytycka, T.M.: Improving strand pairing prediction through exploring folding cooperativity. IEEE/ACM Trans. Comp. Bio. Bioinf. **5**, 484–491 (2008)
9. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices'. J. Mol. Biol. **292**, 195–202 (1999)
10. Klepeis, J.L. Floudas, C.A.: 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)
11. Lippi, M. Frasconi, P.: Prediction of protein beta-residue contacts by Markov logic networks with grounding-specific weights. Bioinformatics **25**, 2326–2333 (2009)
12. MacCallum, R.M.: Striped sheets and protein contact prediction. Bioinformatics **20**(Suppl 1), i224–i231 (2004)
13. Max, N., Hu, C., Kreylos, O., Crivelli, S.: BuildBeta—a system for automatically constructing beta sheets. Proteins **78**, 559–574 (2009)
14. Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Tramontano, A.: Critical assessment of methods of protein structure prediction—Round VIII. Proteins **77**(S9), 1–4 (2009)
15. Porwal, G., Jain, S., Babu, S.D., Singh, D., Nanavati, H., Noronha, S.: Protein structure prediction aided by geometrical and probabilistic constraints. J. Comput. Chem. **28**, 1943–1952 (2007)
16. Rajgaria, R., Wei, Y., Floudas, C.A.: 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 **78**(8), 1825–1846 (2010)
17. Ruczinski, I.: Logic Regression and Statistical Issues Related to the Protein Folding Problem. Ph.D. thesis, Univ. of Washington (2002)
18. Ruczinski, I., Kooperberg, C., Bonneau, R., Baker, D.: Distributions of beta sheets in proteins with application to structure prediction. Proteins **48**, 85–97 (2002)
19. Siepen, J.A., Radford, S.E., Westhead, D.R.: Beta edge strands in protein structure prediction and aggregation. Protein Sci. **12**(10), 2348–2359 (2003)
20. Tegge, A.N., Wang, Z., Eickholt, J., Cheng, J.: NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Res. **37**(37), W315–W318 (2009)
21. Zimmermann, O., Hansmann, U.H.E.: Support vector machines for prediction of dihedral angle regions. Bioinformatics **22**, 3009–3015 (2006)