BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i75–i84
doi:10.1093/bioinformatics/bti1004

Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms

*Jianlin Cheng and Pierre Baldi**

Institute for Genomics and Bioinformatics, School of Information and Computer Sciences, University of California, Irvine, CA 92697, USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: Protein β-sheets play a fundamental role in protein structure, function, evolution and bioengineering. Accurate prediction and assembly of protein β-sheets, however, remains challenging because protein β-sheets require formation of hydrogen bonds between linearly distant residues. Previous approaches for predicting β-sheet topological features, such as β-strand alignments, in general have not exploited the global covariation and constraints characteristic of β-sheet architectures.

Results: We propose a modular approach to the problem of predicting/assembling protein β-sheets in a chain by integrating both local and global constraints in three steps. The first step uses recursive neural networks to predict pairing probabilities for all pairs of interstrand β-residues from profile, secondary structure and solvent accessibility information. The second step applies dynamic programming techniques to these probabilities to derive binding pseudoenergies and optimal alignments between all pairs of β-strands. Finally, the third step uses graph matching algorithms to predict the β-sheet architecture of the protein by optimizing the global pseudoenergy while enforcing strong global β-strand pairing constraints. The approach is evaluated using cross-validation methods on a large non-homologous dataset and yields significant improvements over previous methods.

Availability: http://www.igb.uci.edu/servers/psss.html

Contact: pfbaldi@ics.uci.edu

1 INTRODUCTION

β-Sheets are a fundamental component of protein architectures, >75% of all protein domains in the Protein Data Bank (Berman *et al*., 2000) contain β-sheets (Zhang and Kim, 2000). β-Sheets are formed by the pairing of multiple β-strands held together by characteristic patterns of hydrogen bonds running in parallel or antiparallel fashion (Fig. 1). These patterns, which are essential for β-sheet and protein stability (Smith and Regan, 1997), involve interactions between residues that are often separated by large distances along the primary sequence.

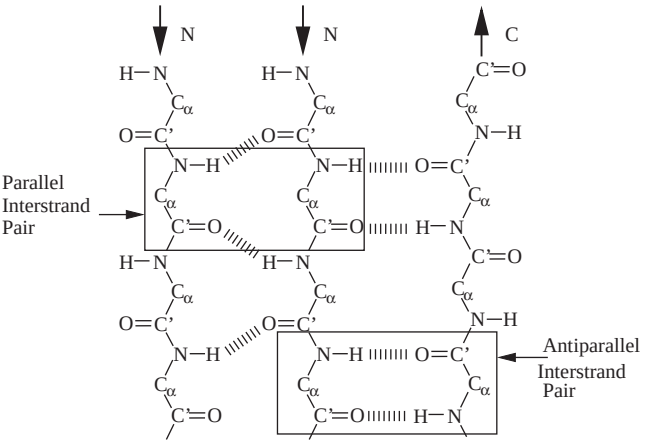


Fig. 1. Illustration of interstrand β-residue pairs and hydrogen-bonding pattern in parallel and antiparallel β-strands. Arrows show the amide (N) to carbonyl (C) direction of β-strands. Hydrogen bonds are represented by hatched blocks.

The β-sheet topology or architecture of a protein, i.e. the pairing organization of all the β-strands contained in a given protein, is essential for understanding its structure (Zhang and Kim, 2000). Prediction of β-sheet topology from amino acid sequence is very useful not only for predicting tertiary structure (Zaremba and Gregoret, 1999; Steward and Thornton, 2002; Ruczinski *et al*., 2002; Rost *et al*., 2003) but also for elucidating folding pathways (Merkel and Regan, 2000; Mandel-Gutfreund *et al*., 2001) and designing new proteins (Smith and Regan, 1995, 1997; Kortemme *et al*., 1998; Kuhlman *et al*., 2003). Many experimental and theoretical studies have been conducted to better understand the formation and stability of β-sheets. For instance, Minor and Kim (1994) report that intrinsic β-sheet propensities of different amino acids contribute to the local structure and stability of β-sheets and that the magnitude and order of β-sheet propensities depend on the local sequence and structural context. Statistical

*To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

studies (Lifson and Sander, 1980; Wouters and Curmi, 1995) reveal non-random distribution and pairing preferences of residue pairs in aligned β-strands, whereas evolutionary conservation of β-residue interactions suggests that pairing preferences depend on structural context, such as solvent accessibility (Zaremba and Gregoret, 1999). Clearly, favorable side-chain interactions between residue pairs contribute to β-sheet stability (Smith and Regan, 1995; Hutchinson *et al*., 1998). However, the evolutionary pressure to maintain complementarity between pairs on neighboring strands appear to be weak (Mandel-Gutfreund *et al*., 2001) and the overall pairing preferences are not very strong and appear to be modulated by the local environment to a high degree.

Several methods, mostly statistical data-driven approaches, have been proposed to predict topological features of β-sheets with moderate accuracy (Rost *et al*., 2003). An early method (Hubbard, 1994) uses a statistical potential approach to predict β-strand alignments with an accuracy level of $\sim$35–45%. Asogawa (1997) proposes to use pairwise statistical potentials of β-residue pairs to improve β-sheet secondary structure prediction by considering clusters of β-residue contacts. Pairwise statistical potentials are used also in the works of Zhu and Braun (1999) to identify up to 35% of native strand alignments from alternative strand alignments. Baldi *et al*. (2000), used elaborate neural networks to improve the prediction accuracy of interstrand β-residue contacts, but the method is not extended to the prediction of strand pairings, strand alignments and β-sheet topologies. Using an information theoretic approach, Steward and Thornton (2002) report an accuracy of 45–48% for strand alignments in β-triplets and 31–37% for any native strand alignments. Although encouraging, all these approaches seem to leave room for major improvements.

These approaches, in particular, fail to exploit systematically the global covariation and constraints characteristic of β-sheet architectures. Instead of treating each pair of β-residues or β-strands independent of each other, as previous methods do, one ought to leverage β-sheet constraints, such as the fact that each β residue has at most two partners, that neighboring β-residues in a strand are paired sequentially in parallel or antiparallel fashion with another strand, and that each β-strand has at least one partner strand and rarely more than two or three partner strands.

In the present study, we develop a novel modular approach for predicting interstrand β-residue pairings, β-strand pairings, β-strand alignments and β-sheet topology altogether from scratch by integrating both local and global constraints in three steps. First, 2D-recursive neural networks (2D-RNN) (Baldi and Pollastri, 2003) are trained to predict pairing probabilities of interstrand β-residue pairs using profile, secondary structure and relative solvent accessibility information. Second, dynamic programming techniques are applied to these probabilities to derive pairing pseudoenergies and alignments between all pairs of β-strands. Third, weighted graph matching algorithms are used to optimize the global β-sheet architecture of the protein satisfying the β-strand pairing constraints. While interchain β-sheets play an important role in protein–protein interactions and complex formation (Dou *et al*., 2004), it is worth noting that here, consistent with the available literature, we focus exclusively on the already challenging prediction of intrachain β-sheets. However, we believe that the methods developed here can be adapted to the problem of predicting both intrachain and interchain β-sheets and training datasets for the latter are available through the ICBS database (Dou *et al*., 2004).

2 MATERIALS AND METHODS

2.1 Data

The dataset is extracted from the Protein Data Bank of May 2004. Only structures determined by X-ray diffraction and having resolution better than 2.5 Å are retained. Chains containing unknown or non-standard amino acids, backbone interruptions or whose length is $<$50 amino acids are excluded. DSSP (Kabsch and Sander, 1983) is used to assign secondary structure and relative solvent accessibility values to each residue. Residues with secondary structure E (extended strand) and B (isolated β-bridge) are considered β-residues. Each β-residue may have 0, 1 or 2 partners according to DSSP. A consistency check is used to remove chains containing non-consistent β-residue pair assignments (e_i, e_j), whereby e_i pairs with e_j, but e_j does not pairs with e_i according to DSSP. A filtering procedure is used to select the chains that contain 10–100 β-residues, of which 90% must have at least one partner. The redundancy in the dataset is reduced by the UniqueProt (Mika and Rost, 2003) with a HSSP threshold of 0, which corresponds to sequence identity of roughly 15–20%.

The final dataset contains 916 chains corresponding to 187 516 residues. Of these, 26% (48 996) are β-residues participating in 31 638 interstrand residue pairs. The dataset has 10 745 β-strands with an average length of 4.6 residues and 8172 β-strand pairs, including 4519 antiparallel pairs, 2214 parallel pairs and 1439 pairs involving isolated β-bridges. These strand pairs form 2533 β-sheets. The average sequence separation between residue pairs and strand pairs is 43 and 40, respectively. Sequence separation histograms are displayed in Figure 2a and 2b. Figure 2c and 2d shows that the number of interstrand residue pairs or strand pairs has a strong correlation with the number of β-residues or strands in the chain, as expected.

To leverage evolutionary information, PSI-BLAST (Altschul *et al*., 1997) is used to generate profiles by aligning all chains against the Non-Redundant (NR) database, as in Pollastri *et al*. (2001). Finally, the dataset is evenly and randomly split into 10 folds to perform 10-fold cross-validation studies. The final dataset (β-sheet 916) and the splitted folds are available through http://www.igb.uci.edu/servers/psss.html

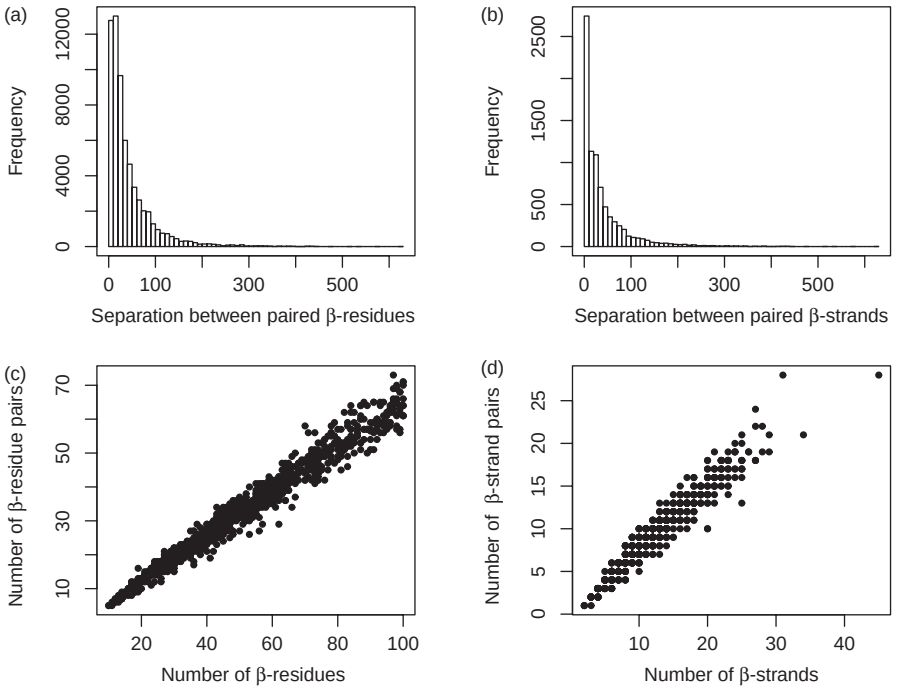


Fig. 2. (a) Amino acid separation between β-residue pairs (mean $= 43$, minimum $= 3$, maximum $= 626$ and standard deviation $= 49$). **(b)** Amino acid separation between β-strand pairs (mean $= 40$, minimum $= 2$, maximum $= 626$ and standard deviation $= 54$). **(c)** Scatterplot of number of β-residue pairs (y) versus number of β-residues (x) per chain. The correlation coefficient is 0.98. Linear regression given by: $y = 0.66x - 0.65$. **(d)** Scatterplot of number of β-strand pairs (y) versus number of β-strands (x) per chain. The correlation coefficient is 0.97. Linear regression given by: $y = 0.74x + 0.27$.

2.2 Prediction of β-residue pairs using 2D-RNNs

Like contact map prediction (Fariselli *et al*., 2001; Pollastri and Baldi, 2002; Shao and Bystroff, 2003; MacCallum, 2004; Punta and Rost, 2005), we treat prediction of interstrand residue pairing as a binary classification problem on a 2D grid. For each chain, our input is a 2D square matrix **I**, where the size of **I** is equal to the number of β-residues in the chain and each entry $I_{i,j}$ is a vector of dimension 251 encoding the local context information of β-residues (e_i, e_j), as well as their separation. Specifically, we use a local window of size 5 around e_i and e_j. Each position in the window corresponds to a vector of length 25 with 20 positions for the amino acid profile, 3 positions for the secondary structure (Helix, Sheet and Coil), and 2 positions for the relative solvent accessibility (buried or exposed at 25% threshold). The two windows correspond to $250 = 25 \times 5 \times 2$ entries. One additional entry represents the sequence separation between e_i and e_j.

The training target is a binary matrix **T**, whereby each $T_{i,j}$ equals 1 or 0 depending on whether β-residue e_i and e_j are paired or not. Figures 3 and 4 show protein 1VJG in the PDB and its corresponding target matrix which nicely displays the constraints and directions (parallel or antiparallel) of strand pairing. Neural networks or other machine learning methods can be trained on the dataset to learn a mapping from the input matrix **I** onto an output matrix **O**, whereby $O_{i,j}$ is the predicted probability that e_i and e_j are paired. The goal is to make the output matrix **O** as close as possible to the target matrix **T**. The standard approach with feed-forward neural networks is to treat each pair (e_i,e_j) independently and to learn a mapping from a series of independent $(I_{i,j}, T_{i,j})$ examples (Baldi *et al*., 2000). This simplified approach, however, does not explicitly leverage covariations and interactions between β-residue pairs and might not effectively enforce the constraints of β-residue and strand pairings. Here we use a 2D-RNN architecture to exploit covariations and constraints between β-residue pairs globally. This 2D-RNN architecture, previously used in contact map prediction, is described in detail in Baldi and Pollastri (2003) and is not reproduced here for lack of space. Under this architecture, the output $O_{i,j}$ depends on the entire input matrix **I** instead of $I_{i,j}$ only. As for feed-forward neural networks, learning in a 2D-RNN is implemented using gradient descent. In the simulations, the outputs of five models are averaged in an ensemble to produce the predicted probability matrix **O**. Finally, it is important to notice that because our approach is modular—it is not constrained in any way to the use of recursive or even feed-forward neural networks—the output of any algorithm that produces an estimate of the

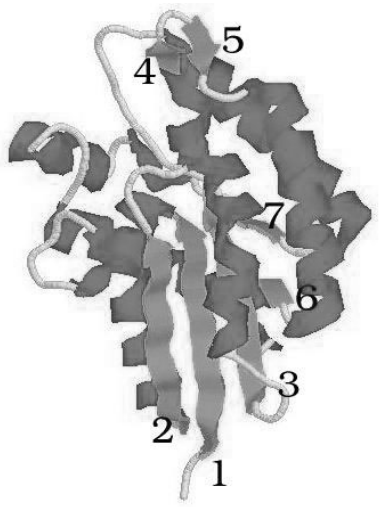


Fig. 3. Protein 1VJG is an α/β protein with seven strands. Strands 1, 2, 3, 6 and 7 form a parallel β-sheet. Strands 4 and 5 form an antiparallel β-sheet. The parallel β-sheet forms the hydrophobic core and is surrounded by tightly packed α-helices.

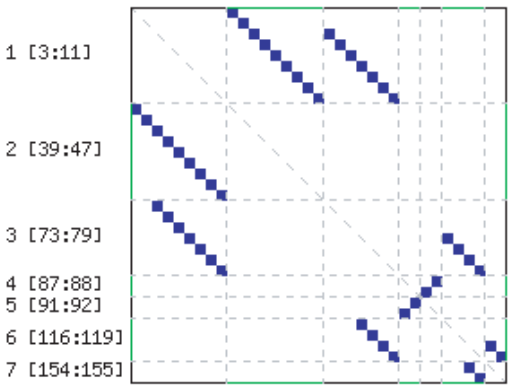


Fig. 4. Interstrand β-residue pairing map of protein 1VJG. The seven strands are ordered along the vertical and horizontal axis. Alternating colors (black and green) are used to distinguish adjacent strands in sequence order. The three numbers associated with each strand on the left are strand number and its starting and ending position along the chain. The map is symmetric. Each blue square represents a native β-residue pairing. A line segment parallel to the main diagonal corresponds to the alignment of a parallel strand pair. A line segment perpendicular to the main diagonal corresponds to the alignment of an antiparallel strand pair. Each row or column has at most two blue squares reflecting the constraint that one residue has at most two partners.

pairing probabilities O_{ij} can be used as input for the second and third steps described below.

Since **I** and **T** are presented to the 2D-RNN as a whole during training, the network can identify pairing constraints encoded in these matrices beyond the local environment of each residue. As a result, by thresholding the values of the

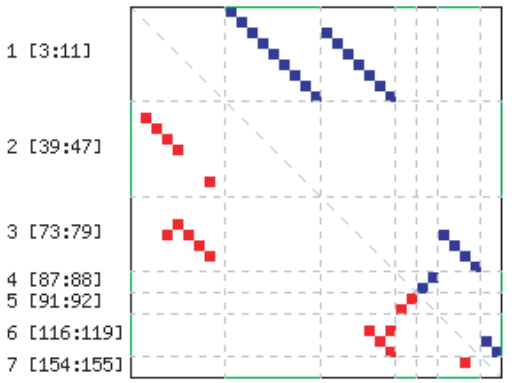
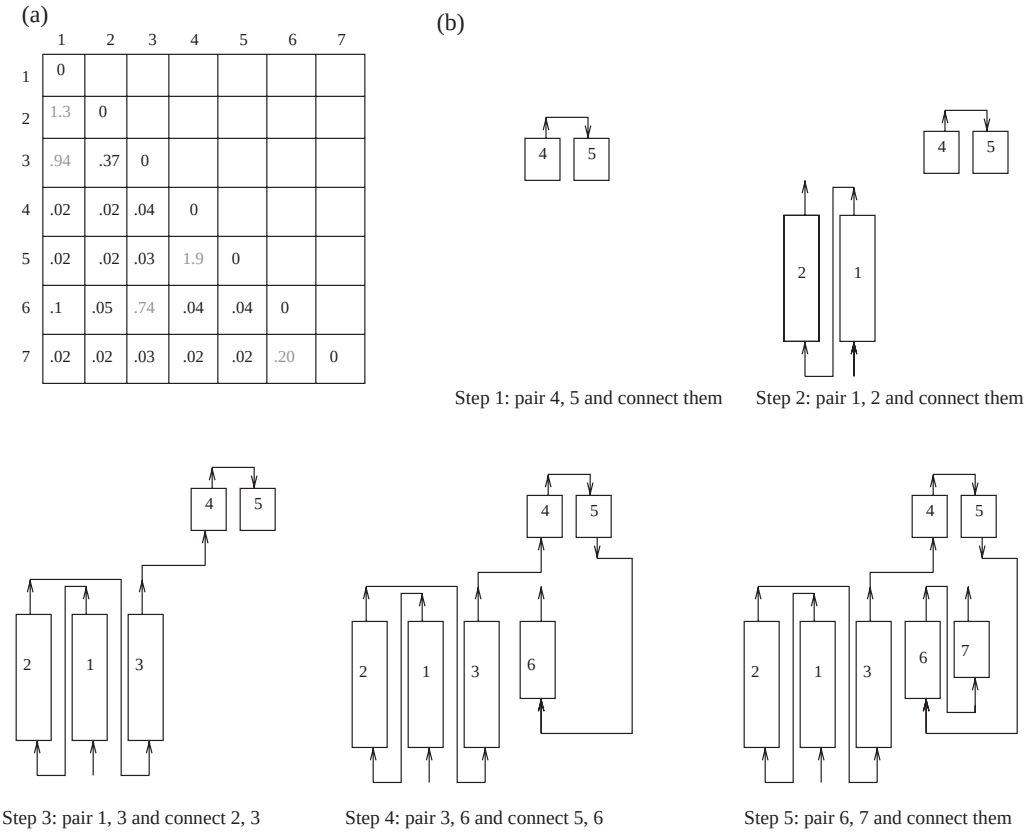


Fig. 5. Predicted β-residue pairing map of 1VJG. Upper triangle (blue) is the true map and lower triangle (red) is the predicted map. The predicted pairs form three segments parallel to the main diagonal corresponding to the true parallel strand pair (1,2), (1,3) and (3,6). Two residue pairs in the true antiparallel strand pair (4,5) are also recalled. One of the two residue pairs in the parallel strand (6,7) is correctly predicted. There are two false positives in strand pair (1,3) and (3,6). For instance, one residue in strand 3 is wrongly predicted as having two partners in strand 1. This error can be detected by checking pairing constraints: a residue can have up to two partners in total, and at most one partner in any single strand. A few residue pairs between strands 1 and 2, which are missing in the predicted map, can be inferred once strands 1 and 2 are predicted to pair.

output **O**, the predicted interstrand residue pairs tend to form line segments parallel or perpendicular to the main diagonal, which correspond to parallel or antiparallel strand pairs. This suggests that aggregate prediction of β-residue pairings can be used to predict β-strand pairings, pairing directions and alignments. Figure 5 shows the predicted interstrand residue pairs of 1VJG with a 0.15 threshold. The predicted map recalls most β-residue pairs and satisfies pairing constraints with few violations. It is worth noting that post-prediction inferences can be used to further enforce some constraints and retrieve some of the missing residue pairs. The predicted interstrand β-residue map can be used directly to infer β-strand pairs. One simple approach we tested is to consider two strands paired if any two of their residues are predicted to be paired. In isolation, however, such an approach cannot be optimal since it disregards global constraints on the number of partners a strand can have (Section 2.4).

2.3 Pseudoenergy for β-strand alignment

For each pair of strands, we can define an optimal alignment and an overall alignment score using dynamic programming techniques in parallel and antiparallel directions with local scores or penalties derived from the matrix **O** of residue-pairing probabilities. Additional intrastrand gap penalties corresponding to β-bulges, as well as penalties for gaps at the end of the strands, can be introduced. The penalty for the bulges can be derived from their frequency. Since β-bulges



	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.1	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Fig. 6. (a) Predicted pseudoenergy matrix **W** of the best alignments of all strand pairs of protein 1VJG. Gray numbers denote the pseudoenergy of the alignments of true strand pairs. **(b)** β-sheet assembly process using graph algorithm. It takes five steps to assemble seven strands into two β-sheets using the energy matrix in (a). In step 1–4, the strand pair with maximum energy is added. In step 5, pair(2,3) has higher energy than pair(6,7). But it is not chosen because strands 2 and 3 have already been selected in previous steps.

tend to be isolated and rare (only 14% of paired strands contain a bulge, and 90% of these contain only a single bulge), to a first-order approximation here we do not allow bulges in the alignments by setting the bulge penalty to infinity. This is also consistent with previous studies (Hubbard, 1994; Zhu and Braun, 1999; Steward and Thornton, 2002). Gaps at the edges of the strands are allowed but are not penalized (penalty $= 0$). Under these assumptions, we can simply search exhaustively through all possible alignments by 'sliding' one strand along the other, in both parallel and antiparallel fashion. Assuming in addition that two paired strands must have at least one residue pairing, two strands with length $m \geq 2$ and $n \geq 2$ have $2(m + n - 1)$ possible alignments, counting parallel and antiparallel directions. If one strand is an isolated bridge ($m = 1$ or $n = 1$), then there are $\max(m, n)$ possible alignments. Without considering β-bulges, one alignment can be uniquely specified by its direction (parallel, antiparallel or isolated bridge) and by one interstrand residue pair.

To discriminate native alignments from alternative ones, the binding pseudoenergy $W(\mathcal{A}[E_r, E_s])$ of each alignment $\mathcal{A}$ of each pair of strands E_r and E_s can be computed by adding the pseudoenergies of each pair of residues i and j in the alignment, derived from the pairing probabilities O_{ij}, or their logarithm $\log O_{ij}$. The binding pseudoenergy W_{rs} of a pair of strands can then be defined by taking the maximum over all their possible alignments: $W_{rs} = \max_{\mathcal{A}} W(\mathcal{A}[E_r, E_s])$. For any pair of strands r and s in a given protein chain, the pseudoenergy is used to identify the best putative alignment, i.e. the one with maximal pseudoenergy W_{rs}, between these two strands. Figure 6a shows the resulting pseudoenergy matrix $\mathbf{W} = (W_{rs})$ for the best alignments between all strand pairs of protein 1VJG. Note how the native strand pairs tend to have higher energy scores suggesting that the pseudoenergy can be used effectively to score and rank strand pairs.

2.4 Prediction of β-strand pairs and β-sheet topology using graph algorithms

Unlike previous methods (Hubbard, 1994; Zhu and Braun, 1999; Steward and Thornton, 2002) which treat strand pairs independent of each other, here prediction of strand pairing and alignment takes into account additional physical constraints characteristic of β-sheet architectures. To illustrate β-sheet topology and its constraints, we use schematic diagrams (similar to Branden and Tooze, 1999) where β-strands

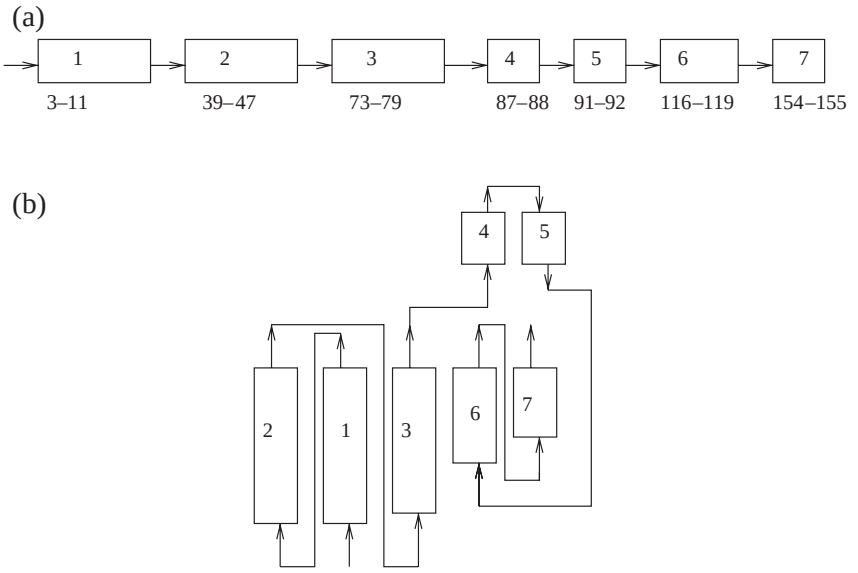


Fig. 7. Schematic diagram of β-sheet topology of protein 1VJG. **(a)** Unpaired strands in sequence order showing the starting and ending positions of the seven strands. **(b)** Topology of β-sheets: paired strands in each β-sheet are aligned side by side. This diagram includes two β-sheets consisting of strands 1,2,3,6 and 7 and strands 4 and 5.

are represented by rectangles of length proportional to the length of the strand. Figure 7 shows the diagram of 1VJG. Lines with arrows connect adjacent strands in sequence order from the N-terminus to the C-terminus. Such schematic diagrams readily reveal several pairing constraints for β-sheet architectures. First, each strand has two edges available for pairing with other strands and, as a result, a β-residue can have at most two partners. It is important to note that this does not imply that a strand can pair at most with two other strands, since a long strand may pair with several short strands on either side. Second, one strand can pair only with one side of another strand sequentially in a parallel or an antiparallel fashion. If two strands pair with the same side of another strand, no overlap is allowed. Third, all strands must have at least one strand partner (ignoring interchain pairings) and we impose the additional condition that they should have at most three strand partners. This condition is not absolute but it is very reasonable since 98.6% of strands have 1, 2 or 3 partners in the large non-redundant dataset. We let $\mathcal{C}$ denote all these constraints.

With these constraints in mind, we develop graph matching algorithms to infer strand pairings and overall β-sheet architecture from the matrix **W** of pseudoenergies of the best alignments of all strands pairs in a given chain. This pseudoenergy matrix defines a completely connected and weighted strand pairing graph (SPG), where vertices represent strands, edges represent possible pairing relations and weights optimal pairing energies. The fully connected SPG of course does not satisfy the set of constraints $\mathcal{C}$. To predict the β-sheet topology, the goal is to prune the complete SPG to derive the true SPG (Fig. 8), where β-sheets appear as maximal connected components. These components are to be derived by maximizing the global pseudoenergy while satisfying all the strand pairing constraints above, i.e. by maximizing $\sum_S W_{rs}$ taken over all subsets S of edges that satisfy $\mathcal{C}$. The global pseudoenergy of an architecture is the sum of the pseudoenergies of each of its β-sheets, and the pseudoenergy of a β-sheet is the sum of the pseudoenergies of all the strand pairs it comprises. To address this constrained optimization problem, we first use a greedy heuristic approach as given in the following table.

Start with a complete SPG with weight matrix **W**. Order all the edges according to the weights into a list L.
$\emptyset \rightarrow S$. S is the set of chosen edges.
Repeat
  Remove one edge e with maximum weight from L.
  If both vertices of e are not in S, add e into S.
  If both vertices of e are in S, discard e.
  If one vertex of e is in S, align the strand of the vertex with the strand of another vertex not in S using their best alignment. If the pair and its alignment satisfy the strand pairing constraints $\mathcal{C}$, add e into S.
  Otherwise discard e.
Until all vertices in G appear in S once.

The greedy algorithm has time complexity $O(N^2 \log N)$, where N is the number of strands. After converging, the edges and vertices in S constitute a spanning subgraph G^* of G.

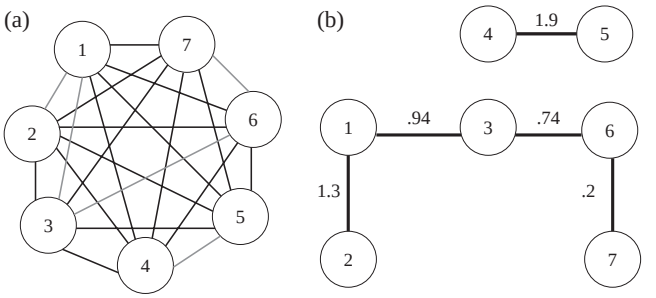


Fig. 8. Strand Pairing Graph of protein 1VJG. **(a)** The complete SPG. Gray edges denote true strand pairs. **(b)** The true weighted SPG. Two components (1,2,3,6 and 7) and (4 and 5) correspond to two β-sheets. The weights are the pseudoenergy of the best alignments of strand pairs.

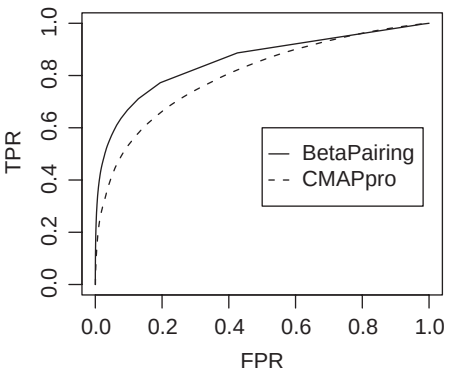


Fig. 9. ROC curve of prediction of interstrand β-residue pairs using the β-residue pairing predictor and CMAPpro.

Connected components in G^* are in 1:1 correspondence with the protein β-sheets and provide the global predicted β-sheet architecture. Figure 6 illustrates how the algorithm assembles the strands of protein 1VJG.

By treating β-sheets as spanning trees of complete SPGs, a variant of the well-known algorithm for finding minimum/maximum spanning tree (MST) (Even, 1979), Kruskal's algorithm (Kruskal, 1956), is also used to predict β-sheets (trees) with maximum pseudoenergy. The only difference between this constrained MST algorithm and the previous greedy algorithm is that it does not always discard edge e when its adjacent vertices are already in the set S. Instead, it adds e into L if its two vertices belong to two disconnected components and the alignment satisfies the strand pairing constraints. Not surprisingly, this algorithm tends to choose more strand pairs (edges) than the greedy graph algorithm. It is worth noting that both the greedy and constrained MST algorithms as described do not allow for cycles and all the components they produce are trees. This approximation is not entirely correct in the case of circular β-sheets, such as β-barrels. To handle β-barrels, we are currently modifying these algorithms to allow up to one cycle in each component.

3 RESULTS AND DISCUSSION

The performance of β-residue pairing prediction is assessed using a variety of standard measures including: area under ROC curve, true positive rate [TPR $=$ TP/(TP $+$ FN)], at 5% false positive rate [FPR $=$ FP/(FP $+$ TN)], specificity [TP/(TP $+$ FP)], sensitivity [TP/(TP $+$ FN)] and correlation coefficient $[(\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN})/((\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP}))^{1/2}]$, and compared with predictions associated with the base-line and with a general purpose contact map predictor. At the break-even point where the total number of predicted β-residue pairs is equal to the true number of β-residue pairs, the specificity and sensitivity of interstrand β-residue pairings are equal to 41% with a correlation coefficient of 0.4. The accuracy of the base-line predictor (the number of true β-residue pairs/total number of interstrand β-residue pairs) is 2.3%. Thus, the improvement factor, i.e. the ratio between the accuracy (specificity or sensitivity) of our method over the base-line (Fariselli *et al*., 2001), is 17.8. To the best of our knowledge, only one method in the literature (Baldi *et al*., 2000) reports quantitive evaluation of β-residue pairing prediction. However, it reports only sepcificity without mentioning the corresponding sensitivity, thus a direct comparison cannot be made. However, we can compare the β-residue pairing predictor with a general purpose contact map predictor (Pollastri and Baldi, 2002) focusing exclusively on β-residue pairings. We use a pretrained 8 Å contact map predictor (CMAPpro) to predict contacts for all chains in the same dataset. To make the comparison even more stringent, we do not take into consideration any homology between the current dataset and the dataset used to train (CMAPpro). We then extract the contact probabilities for β-residue pairings from the full predicted contact map and evaluate them using the same measures. At the break-even point, the specificity and sensitivity of CMAPpro are equal to 27% and the correlation coefficient is 0.26. Thus, our method improves the specificity and sensitivity of CMAPpro restricted to β-residues by 14%. The area under the ROC curve for the beta-pairing predictor is 0.86 versus 0.80 for CMAPpro (Fig. 9). At 5% FPR, TPR for the beta-pairing predictor is 58% versus 42% for CMAPpro. Thus the specialized β-residue pairing predictor significantly improves the predictions of our general purpose contact map predictor restricted to β-strands, consistently with previous expectations (Rost *et al*., 2003).

The correlation coefficients of strand pairing by the greedy and constrained MST graph algorithms are virtually identical (0.502 and 0.503, respectively). The specificity and sensitivity of strand pairing using the greedy graph algorithm are 59 and 54%, respectively. In contrast, the specificity and sensitivity of the naive algorithm that always pairs sequentially

adjacent strands are 42 and 50%, respectively. Thus, around similar operating regimes, the greedy graph algorithm yields improvements of 17% in specificity and 4% in sensitivity over the naive algorithm. The small improvement in sensitivity is still very significant because 16% of correctly predicted strand pairs are non-adjacent strand pairs. The constrained MST graph algorithm has specificity and sensitivity of 53 and 59%, respectively. Its sensitivity is 9% higher than the naive algorithm and 20% of correctly predicted strand pairs are non-adjacent strand pairs.

Using the pseudoenergy to align strand predicted to be paired by the greedy graph algorithm, pairing directions (parallel, antiparallel or isolated bridge) of 93% of the correctly predicted strand pairs are correctly identified, 72% of which are correctly aligned (71% of parallel pairs, 69% of antiparallel pairs and 88% of strand pairs involving isolated bridges). The constrained MST graph algorithm yields similar results.

To further evaluate the ability of the pseudoenergy to discriminate true alignments from false alignments, we use it to align all native strand pairs. Pairing directions of 84% native pairs are correctly predicted. Considering only parallel and antiparallel pairs, the pairing directions of 82% of these pairs are predicted correctly, which yields a 15% improvement over the 67% precision achieved by the trivial algorithm which labels all pairs as being antiparallel. Among all strand pairs with correctly predicted directions, 66% of them are aligned correctly (66% of parallel pairs and 63% of antiparallel pairs and 72% of isolated bridges). In comparison, on different datasets, the statistical potential approach in Hubbard (1994) aligns 35–45% of strand pairs correctly, when pairing directions are correctly predicted. If we assume that all pairing directions are known, as some previous methods do (Zhu and Braun, 1999; Steward and Thornton, 2002), then 61% of all native parallel pairs and 60% of all native antiparallel pairs are aligned correctly. The pseudoenergy approach based on pairwise potentials in Zhu and Braun (1999) discriminates 35% of native alignments from alternative alignments, assuming pairing directions are known. Thus on a larger albeit different dataset, the accuracy of the method presented here is significantly higher than previous approaches. Assuming that pairing direction and position of one strand is known, the information theoretic approach of Steward and Thornton (2002), which aligns the known strand with all subsequences in a ± 10 offset around another strand to identify the best alignment, achieves precisions of 48 and 45% for parallel and antiparallel pairs in strand triplets, and 37 and 31% for arbitrary parallel and antiparallel pairs, respectively. Since our methods assume that the position of the two β-strands under consideration is known—in a purely *ab initio* setting, this would have to be predicted (Rost and Sander, 1993; Jones, 1999; Pollastri *et al*., 2001)—the alignment accuracy of our methods cannot be compared directly with the information theoretic approach. However, our results show that it is easier to align parallel strand pairs than antiparallel ones, which agrees with the observations

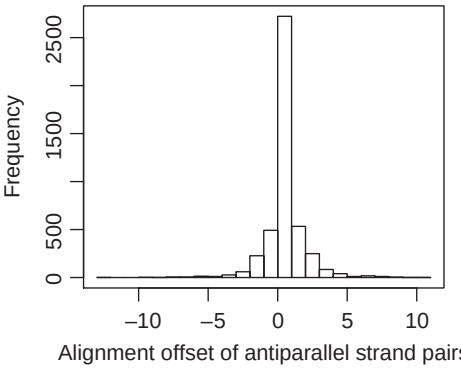


Fig. 10. Histogram of alignment offsets of antiparallel strand pairs. A perfect alignment corresponds to a 0 offset.

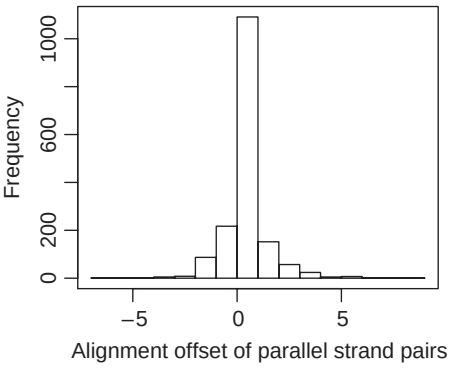


Fig. 11. Histogram of alignment offsets of parallel strand pairs. A perfect alignment corresponds to a 0 offset.

derived using the information theoretic approach. Figures 10 and 11 show the histograms of alignment offsets of all parallel and antiparallel pairs where paring directions are correctly predicted. No simple metric is as yet available for evaluating the prediction of β-sheet topologies. Here we report the strand pairing precision of predicted β-sheets, i.e. the proportion of correctly predicted strand pairs in each β-sheet. Using the greedy graph algorithm, for instance, 51% of predicted β-sheets have β-strand pairing precision $>60\%$.

4 CONCLUSION

We have proposed a new *ab initio* modular approach to the problem of predicting and assembling β-sheets. The method is modular in the sense that alternative algorithms can be 'plugged in' for each one of its stages, for instance in order to predict residue pairing probabilities. Starting from β-residue pairing probabilities, the method provides an integrated prediction of β-sheet architectures by predicting β-strand pairs, β-strand alignments and β-sheets assembly. The pseudoenergy derived from pairing probabilities of β-residue pairs can rather accurately predict β-strand alignments and score β-strand pairs. The greedy and constrained MST graph

algorithms are able to predict strand pair and β-sheet topology from pseudoenergy matrices by globally optimizing the pseudoenergy of β-sheets. While the performance of, for instance, β-strand alignment appears significantly improved over previous statistical data-driven approaches, it is clear that even further improvements should be possible in each one of the three stages. For instance, in the first step, more information about the interstrand sequence can be included (Punta and Rost, 2005). In the second step, gap penalties for β-bulges can be taken into account. In the third step, graph algorithms that allow cycles ought to recover cyclic β-sheets. Furthermore, constrained optimization of the binding pseudoenergy derived here is at best an approximation that will need to be refined to include other packing constraints associated with other secondary structure elements.

β-Sheets have remained one of the main stumbling blocks of protein structure prediction over the years. Thus, new methods for the accurate prediction of β-sheets may lead to noticeable improvements in the study of protein structure and folding and in protein design. Our results suggests that the methods presented here can be combined with contact map prediction to generate more accurate contact maps, which in turn can be used in fold recognition and 3D reconstruction. Accurate β-residue and β-strand pairings may also provide strong constraints for improving *ab initio* sampling of tertiary structures and derive energy terms to help select near-native structures from decoys.

ACKNOWLEDGMENTS

Work supported in part by NIH Biomedical Informatics Training grant (LM-07443-01) and NSF MRI grant (EIA-0321390) to PB.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Asogawa,M. (1997) Beta-sheet prediction using inter-strand residue pairs and refinement with hopfield neural network. In *Proceedings of International Conference on Intelligent System of Molecular Biology*, Vol. 5, AAAI Press, Menlo Park, CA, pp. 48–51.

Baldi,P. and Pollastri,G. (2003) The principled design of large-scale recursive neural network architectures—DAG-RNNs and the protein structure prediction problem. *Journal of Machine Learning Research*, **4**, 575–602.

Baldi,P., Pollastri,G., Andersen,C.A.F. and Brunak,S. (2000) Matching protein β-sheet partners by feedforward and recurrent neural networks. In *Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA*. AAAI Press, Menlo Park, CA, pp. 25–36.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Branden,C. and Tooze,J. (1999) *Introduction to Protein Structure, 2nd edn.* Garland Publishing, New York, NY.

Dou,Y., Baisnee,P., Pollastri,G., Pecout,Y., Nowick,J. and Baldi,P. (2004) ICBS: a database of interactions between protein chains mediated by beta-sheet formation. *Bioinformatics*, **20**, 2767–2777.

Even,S. (1979) *Graph Algorithms*. Computer Science Press, Rockville, MD.

Fariselli,P., Olmea,O., Valencia,A. and Casadio,R. (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng.*, **13**, 835–843.

Hubbard,T.J. (1994) Use of β-strand interaction pseudo-potentials in protein structure prediction and modelling. In Lathrop,R.H., (ed.), *Proceedings of the Biotechnology Computing Track, Protein Structure Prediction MiniTrack of the 27th HICSS*. IEEE Computer Society Press, pp. 336–354.

Hutchinson,E.G., Sessions,R.B., Thornton,J.M. and Woolfson,D.N. (1998) Determinants of strand register in antiparallel beta-sheets of proteins. *Protein Sci.*, **7**, 287–300.

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

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kortemme,T., Ramirez-Alvarado,M. and Serrano,L. (1998) Design of a 20-amino acid, three-stranded β-sheet protein. *Science*, **281**, 253–256.

Kruskal,J.B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. In *Proceedings of the American Mathematical Society*, Vol. 7, pp. 48–50.

Kuhlman,B., Dantas,G., Ireton,G., Varani,G., Stoddard,B. and Baker,D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.

Lifson,S. and Sander,C. (1980) Specific recognition in the tertiary structure of beta-sheets of proteins. *J. Mol. Biol.*, **139**, 627–639.

MacCallum,R.M. (2004) Striped sheets and protein contact prediction. *Bioinformatics*, **20** (Suppl 1), i224–i231.

Mandel-Gutfreund,Y., Zaremba,S.M. and Gregoret,L.M. (2001) Contributions of residue pairing to beta-sheet formation:conservation and covariation of amino acid residue pairs on antiparallel beta-strands. *J. Mol. Biol.*, **305**, 1145–1159.

Merkel,J.S. and Regan,L. (2000) Modulating protein folding rates *in vivo* and *in vitro* by side-chain interactions between the parallel beta strands of green fluorescent protein. *J. Biol. Chem.*, **275**, 29200–29206.

Mika,S. and Rost,B. (2003) Uniqueprot: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Minor,D.L. and Kim,S. (1994) Context is a major determinant of beta-sheet propensity. *Nature*, **371**, 264–267.

Pollastri,G. and Baldi,P. (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, **18** (Suppl 1), S62–S70.

Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2001) Improving the prediction of protein secondary strucure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.

Punta,M. and Rost,B. (2005) Toward good 2d predictions in proteins. *FEBS*, in press.

Rost,B., Liu,J., Przybylski,D., Nair,R., Wrzeszczynski,K.O., Bigelow,H. and Ofran,Y. (2003) Prediction of protein structure through evolution. In Gasteiger,J. and Engel,T., (eds), *Handbook of Chemoinformatics—From Data to Knowledge*. Wiley, New York, pp. 1789–1811.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

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

Shao,Y. and Bystroff,C. (2003) Predicting inter-residue contacts using templates and pathways. *Proteins*, **53** (Suppl 6), 497–502.

Smith,C.K. and Regan,L. (1995) Guidelines for protein design: The energetics of β sheet side chain interations. *Science*, **270**, 980–982.

Smith,C.K. and Regan,L. (1997) Construction and design of beta-sheets. *Acc. Chem. Res.*, **30**, 153.

Steward,R.E. and Thornton,J.M. (2002) Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins Struct. Funct. Genet.*, **48**, 178–191.

Wouters,M.A. and Curmi,P.M.G. (1995) An analysis of side-chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins Struct. Funct. Genet.*, **22**, 119–131.

Zaremba,S.M. and Gregoret,L.M. (1999) Context-dependence of amino acid residue pairing in antiparallel β-sheets. *J. Mol. Biol.*, **291**, 463–479.

Zhang,C. and Kim,S. (2000) The anatomy of protein beta-sheet topology. *J. Mol. Biol.*, **2**, 1075–1089.

Zhu,H. and Braun,W. (1999) Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting. *Protein Sci.*, **8**, 326–342.