Protein β-Sheet Partner Prediction by Neural Networks

Pierre Baldi* and Gianluca Pollastri
Department of Information and Computer Science, University of California Irvine
Irvine, CA 92697-3425

Claus A. F. Andersen and Søren Brunak
Center for Biological Sequence Analysis, The Technical University of Denmark
DK-2800 Lyngby, Denmark

Abstract

Predicting the secondary structure (α-helices, β-sheets, coils) of proteins is an important step towards understanding their three dimensional conformations. Unlike α-helices that are built up from one contiguous region of the polypeptide chain, β-sheets are more complex resulting from a combination several disjoint regions. The exact nature of these long distance interactions remains unclear. Here we introduce a neural-network based method for the prediction of amino acid partners in parallel as well as anti-parallel β-sheets. The neural architecture predicts whether two residues located at the center of two distant windows are paired or not in a β-sheet structure. The distance between the windows is a third essential input into the architecture. Variations on this architecture are trained using a large corpus of curated data. Prediction on both coupled and non-coupled residues currently exceeds 83% accuracy, well above any previously reported method. Unlike standard secondary structure prediction methods, the use of multiple alignment (profiles) in our case seems to degrade the performance, probably as a result of intra-chain correlation effects.

1 Background

Predicting the secondary structure (α-helices, β-sheets, coils) of proteins is an important step towards understanding their three dimensional conformations. Unlike α-helices that are built up from one contiguous region of the polypeptide chain, β-sheets are built up from a combination of several disjoint regions. These regions, or β strands are typically 5-10 residues long. In the folded protein, these strands are aligned adjacent to each other in parallel or anti-parallel fashion. Hydrogen bonds can form between C'O groups of one strand and NH groups on the adjacent strand and vice versa with C_α atoms successively a little above or below the plane of the sheet. Hydrogen bonds between parallel and anti-parallel strands have distinctive patterns, but the exact nature and behavior of β-sheet long-ranged interactions is not clear.

While the majority of sheets seems to consist of either parallel or antiparallel strands, mixed sheets are not uncommon. A β-strand can have 1 or 2 partner strands,

*and Department of Biological Chemistry, College of Medicine, University of California, Irvine. To whom all correspondence should be addressed (*pfbaldi@ics.uci.edu*).

and an individual amino acid can have 0,1 or 2 hydrogen bonds with one or two residues in a partner strand. Sometimes one or several partner-less residues are found in a strand, giving rise to the so-called β-bulges. Finally, β-strand partners are often located on a different protein chain. How amino acids located far apart in the sequence find one another to form β-sheets is still poorly understood, as is the degree of specificity between side-chain/side-chain interactions between residues on neighboring strands, which seems to be very weak [13]. The presence of a turn between strands is also an essential ingredient.

Partly as a result of the exponentially growing amount of available 3D data, machine learning methods have in general been among the most successful in secondary structure prediction [2]. The best existing methods for predicting protein secondary structure, i.e. for classifying amino acids in a chain in one of the three classes, achieve prediction accuracy in the 75-77% range [3, 4, 8]. Therefore any improvement in β-sheet prediction is significant as a stand-alone result, but also in relation to secondary and tertiary structure prediction methods in general. Here we design and train a neural network architecture for the prediction of amino acid partners in β-sheets (see also [7, 15]).

2 Data Preparation

2.1 Selecting the Data

As always the case in machine learning approaches, the starting point is the construction of a well-curated data set. The data set used here consists of 826 protein chains from the PDB select list of June 1998 [5] (several chains were removed since DSSP could not run on them). All the selected chains have less than 25% sequence identity using the Abagyan-function [1]. The selection has been performed by applying the all against all Huang-Miller sequence alignment using the "sim" algorithm [6], where the chains had been sorted according to their quality (i.e. resolution plus R-factor/20 for X-ray and 99 for NMR).

2.2 Assigning β-sheet Partners

The β-sheets are assigned using Kabsch and Sander's DSSP program [9], which specifies where the extended β-sheets are situated and how they are connected. This is based on the intra-backbone H-bonds forming the sheet according to the Pauling pairing rules [11]. An H-bond is assigned if the Coulomb binding energy is below -0.5 kcal/mol. In wild-type proteins there are many deviations from Paulings ideal binding pattern, so Kabsch and Sander have implemented the following rules: a β-sheet ('E') amino acid is defined when it forms two H-bonds in the sheet or is surrounded by two H-bonds in the sheet. The minimal sheet is two amino acids long; if only one amino acid fulfills the criteria, then it is called β-bridge ('B'). Bulges in sheets are also assigned 'E' if they are surrounded by normal sheet residues of the same type (parallel or anti-parallel) and comprise at most 4 and 1 residue(s) in the two backbone partner segments, respectively.

A standard example of how the partner assignments are made is shown in figure 1. In the case of β-bridges the same rules are followed, while in the special case of β-bulge residues then no partner is assigned.

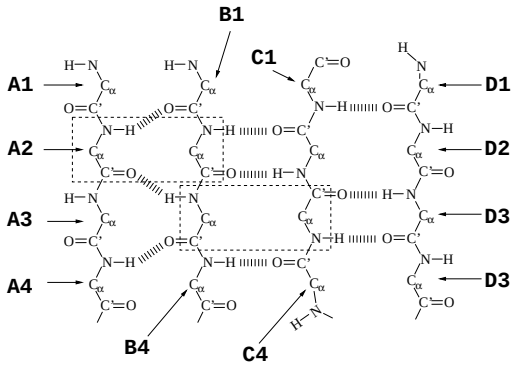

Figure 1: The assignment criteria for sheet partners are shown for two examples by the dashed boxes. That is the A sheet segment binds to the B sheet segment with a parallel sheet and residue A2 is the partner of B2. The other dashed box shows that B3 is the partner of C3, even though none of them has H-bonds in the anti-parallel B-C sheet. The other sheet partners in the example shown are: A3-B3, B2-C2, C2-D2 and C3-D3. Note that the residues A1,A4,B1,B4,C1,C4,D1,C4 are not sheet residues.

3 Neural Network Architecture

A number of different artificial neural network approaches can be considered. Because of the long-ranged interactions involved in beta-sheets, neural architectures must have either very large input windows or distant shorter windows. Very large input windows lead to architectures with many parameters which are potentially prone to overfitting, especially with sparse amino acid input encoding. Overfitting, however, is not necessarily the main obstacle because data is becoming abundant and techniques, such as weight sharing, can be used to mitigate the risk. Perhaps the main obstacle associated with large input windows is that they tend to dilute sparse information present in the input that is really relevant for the prediction [10].

Here we have used a basic two-windows approach. Since the distance between the windows plays a key role in the prediction, one can either provide the distance information as a third input to the system or one can train a different architecture for each distance type. Here, we use the first strategy with the neural network architecture depicted in Figure 2 (see also [12]). The architecture has two input windows of length W corresponding to two amino acid substrings in a given chain. The goal of the architecture is to output a probability reflecting whether the two amino acids located at the center of each window are partners or not. The sequence separation between the windows, measured by the number D of amino acids, is essential for the prediction and is also given as an input unit to the architecture with scaled activity $D/100$. As in other

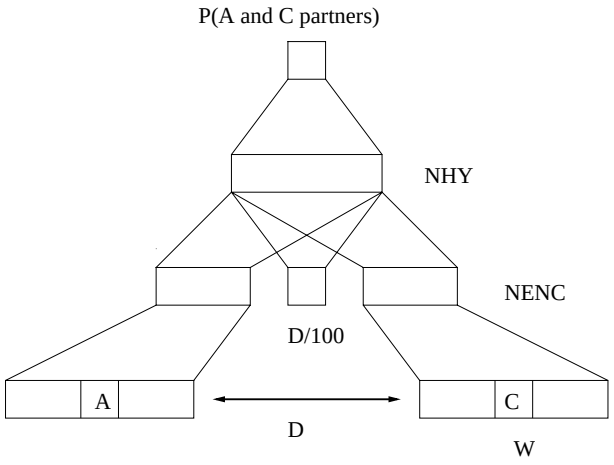

Figure 2: Neural network architecture for amino acid β-partner prediction.

standard secondary structure prediction architectures, we use sparse encoding for the 20 amino acids. Each input window is post-processed by a hidden layer comprising a number NENC of hidden units. Information coming from the input windows and the distance between the windows are combined in a fully interconnected hidden layer of size NHY. This layer is finally connected to a single logistic output unit that estimates the partnership probability. The architecture is trained by back-propagation on the relative entropy between the output and target probability distributions.

4 Experiments and Results

For training, we randomly split the data 2/3 for training and 1/3 for testing purposes. A typical split gives:

Table 1: Training set statistics, with number of sequences, amino acids, and positive and negative examples.

	Training set	Test set
Sequences	551	275
Amino acids	129119	64017
Positive ex.	37008	18198
Negative ex.	44,032,700	22,920,100

The number of negative examples (pairs of amino acids that are not partners) is of course much higher. In order to have balanced training, at each epoch we present all the 37008 positive examples, together with 37008 randomly selected negative examples at each epoch. We use a hybrid between on-line and batch training, with 50 batch

blocks, i.e. weights are updated 50 times per epoch. The training set is also shuffled at each epoch, so the error is not decreasing monotonically. The learning rate per block is set at 3.8×10^{-5} at the beginning and is progressively reduced. There is no momentum term or weight decay. When there is no error decrease for more than 100 epochs, the learning rate is divided by 2. Training stops after 8 or more reductions, corresponding to a learning rate that is 256 times smaller than the initial one. Typical performances are given below for different architectural variations. The percentages are computed on the entire test set, including all the negative examples it contains.

The results of several training experiments using different variants of the same architecture are summarized in Table 2

Table 2: Performance results expressed in percentages of correct prediction. W=input window length, NENC=number of hidden units in the post-processing layers of each window, NHY=number of hidden units in the output hidden layer. The second experiment with the 10/11/7 architecture involves multiple alignments (see text). Overall percentage is the simple average of the percentage on each class.

NHY	NENC	W	beta	non-beta	total
8	7	3	83.00	79.29	81.15
8	7	4	83.00	79.80	81.40
8	7	5	82.92	80.05	81.43
8	7	6	83.27	80.37	81.87
8	7	7	83.55	80.28	81.91
10	9	6	83.25	80.60	81.93
10	9	7	83.38	83.84	83.61
10	9	8	83.49	80.84	82.16
10	11	7	83.93	83.34	83.64
10	11	7	76.32	87.77	82.04
10	12	7	82.31	84.36	83.33
12	11	7	83.41	82.30	82.86

The best overall results (83.64%) are obtained with an architecture with a window length of $W = 7$ and hidden unit layers with $NENC = 11$ and $NHY = 10$. This architecture achieves similar accuracy on both partner and non-partner classes (83.93% and 83.34% respectively). It is worthwhile to notice that a small network with three hidden units trained using the distance between the amino acids alone as input achieves an average performance of 75.39% (80.35% on beta-sheet partners and 70.43% on non-partners).

It is well known that evolutionary information in the form of multiple alignments and profiles significantly improves the accuracy of secondary structure prediction methods. This is because the secondary structure of a family is more conserved than the primary amino acid structure. Notice, however, that in the case of beta-sheet partners, intra-sequence correlations may be essential and these are lost in a profile approach where the distributions associated with each column of a multiple alignment are considered independent. To test these effects, we used the BLAST program with standard default parameters (such as BLOSUM matrix 62) to create multiple alignments of our sequences and retrained the optimal architecture found with the corre-

sponding profiles. As can be seen in the table, the overall performance appears to slightly degrade to 82.04%. More interestingly, however, the performance on the non-partner class is improved (87.77 %), whereas the performance on the partner class is degraded (76.32%). This is consistent with a selective improvement of secondary structure prediction resulting from multiple alignments, which does not extend to beta sheet partners as a result of important intra-sequence correlations that are lost in multiple alignments. This may imply that the actual correlation between sheet sequences is higher than previously thought.

5 Discussion

Perfect prediction of protein secondary structures is probably impossible for a variety of reasons including the fact that a significant fraction of proteins may not fold spontaneously [14], that beta-strand partners may be located on a different chain, and that conformation may also depend on other environmental variables, related to solvent, acidity, and so forth. It is however comforting to observe that steady progress is being made in this area, with an increasing number of folds being solved in the structural data bases, and steady improvement of classification and machine learning methods. Here we have developed a neural network architecture that predicts beta-sheet amino acid partners with a performance of almost 84% correct prediction.

There are several directions in which this work can be extended which are currently in progress. These include:

- The development of secondary structure prediction methods for beta sheets based on sequences rather than profiles, to be combined with the profile-based methods which work better with α-helices and coils.

- The use of the present architecture as a beta-sheet predictor rather than a partner predictor, possibly in combination with another neural network.

- Various combinations of the present architectures with existing secondary structure predictor to improve beta-sheet prediction performance.

- The prediction of beta-strand partners rather than amino-acid partners.

- The combination of alignments with partner prediction in order to better predict beta-strands. In particular, a neural network could be trained to predict for each β-strand an ideal partner strand based on amino acid pairing statistics. True partner strands could then be searched by looking for regions in the sequence that have high parallel or antiparallel alignment scores with the putative ideal partner sequence.

- The use of additional information, such as amino acid properties (hydrophobicity, etc.) to improve prediction accuracy.

Acknowledgements

The work of PB is supported by a Laurel Wilkening Faculty Innovation award at UCI. The work of SB and CA is supported by a grant from the Danish National Research Foundation.

References

[1] R.A. Abagyan and S. Batalov. Do aligned sequences share the same fold? *J. Mol. Biol.*, 273:355–368, 1997.

[2] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, MA, 1998.

[3] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15:937–946, 1999.

[4] J. A. Cuff and G. J. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, 34:508–519, 1999.

[5] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.*, 3:522, 1994.

[6] X. Huang and W. Miller. *Adv. Appl. Math.*, 12:337–357, 1991.

[7] T. J. Hubbard. Use of b-strand interaction pseudo-potentials in protein structure prediction and modelling. In R. H. Lathrop, editor, *In: Proceedings of the Biotechnology Computing Track, Protein Structure Prediction Minitrack of 27th HICSS*, pages 336–354. IEEE Computer Society Press, 1994.

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

[9] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen–bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[10] O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, and S. Brunak. Protein distance constraints predicted by neural networks and probability density functions. *Prot. Eng.*, 10:1241–1248, 1997.

[11] L. Pauling and R.B. Corey. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc. Natl. Acad. Sci. USA*, 37:729–740, 1951.

[12] S. K. Riis and A. Krogh. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol.*, 3:163–183, 1996.

[13] M.A. Wouters and P.M.G. Curmi. An analysis of side chain interaction and pair correlation within anti-parallel beta-sheets: The difference between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins*, 22:119–131, 1995.

[14] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293:321–331, 1999.

[15] H. Zhu and W. Braun. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins. *Protein Science*, 8:326–342, 1999.