Matching Protein β-Sheet Partners by Neural Networks

Pierre Baldi * and **Gianluca Pollastri**
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
(949) 824-5809
(949) 824-4056 (FAX)
{*pfbaldi,gpollast*}*@ics.uci.edu*

Claus A. F. Andersen and **Søren Brunak**
Center for Biological Sequence Analysis
The Technical University of Denmark
DK-2800 Lyngby, Denmark
+45 45252477
+45 45931585 (FAX)
{*ca2,brunak*}*@cbs.dtu.dk*

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 of two or more disjoint regions. The exact nature of these long distance interactions remains unclear. Here we introduce two neural-network based methods for the prediction of amino acid partners in parallel as well as anti-parallel β-sheets. The neural architectures predict whether two residues located at the center of two distant windows are paired or not in a β-sheet structure. Variations on these architecture, including also profiles and ensembles, are trained and tested via five-fold cross validation using a large corpus of curated data. Prediction on both coupled and non-coupled residues currently approaches 84% accuracy, better than any previously reported method.

Introduction

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. Such 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 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 (Wouters & Curmi, 1995). 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 (Baldi & Brunak, 1998). 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 (Baldi *et al.*, 1999; Cuff & Barton, 1999; Jones, 1999). Not surprisingly,

and Department of Biological Chemistry, College of Medicine, University of California, Irvine. To whom all correspondence should be addressed.

β-sheet is almost invariably the weakest category in terms of correct percentages, and it is never the highest scoring in terms of correlation coefficients. 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 (Hubbard, 1994; Zhu & Braun, 1999; Street & Mayo, 1999)).

Data Preparation

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 (Hobohm & Sander, 1994) (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 (Abagyan & Batalov, 1997). The selection has been performed by applying the all against all Huang-Miller sequence alignment using the "sim" algorithm (Huang & Miller, 1991), where the chains had been sorted according to their quality (i.e. resolution plus R-factor/20 for X-ray and 99 for NMR).

Assigning β-sheet Partners

The β-sheets are assigned using Kabsch and Sander's DSSP program (Kabsch & Sander, 1983), 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 (Pauling & Corey, 1951). An H-bond is assigned if the Coulomb binding energy is below -0.5 kcal/mol. In wildtype 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.

Statistical Analysis

First Order Statistics

The first order statistics associated with the frequency of occurrence of each amino acid in the data in general, and specifically within β-sheets are displayed in Figures 2 and 3. To enhance the similarities and differences, the ratio of the frequencies in β-sheets over data are given in Figure 4.

Second Order Statistics

Second order statistics are associated with pairings of amino acids in β-sheets. The frequency of pairings can be normalized in different ways. In each row of Figure 5, we have plotted the conditional probabilities $P(X|Y)$ of observing a X knowing that the partner is Y in a β-sheet.

Length Distribution

Interval distances between paired β-strands, measured in residue positions along the chain, are given in Figure 7. A small number of pairs have large distances above 200 amino acids and are not represented. Distances could also be measured circularly but this would not alter the fundamental features of the plot.

Artificial Neural Network Architecture

A number of different artificial neural network (ANN) 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 (Lund *et al.*, 1997).

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 8 (see also (Riis & Krogh, 1996)). 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 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

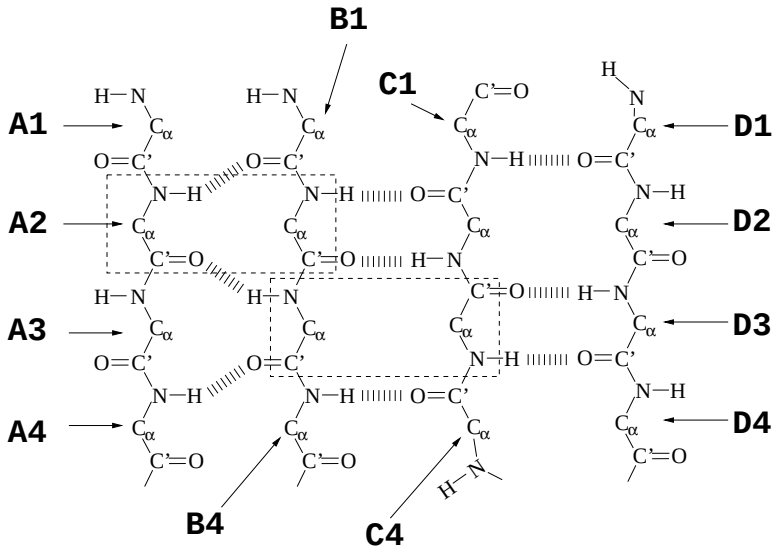

Figure 1: The assignment criteria for sheet partners are shown for two examples by the dashed boxes. That is the A sheet segment binds to the B sheet segment with a parallel sheet and residue A2 is the partner of B2. The other dashed box shows that B3 is the partner of C3, even though none of them has 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 residues A1, A4, B1, B4, C1, C4, D1, D4 are not sheet residues.

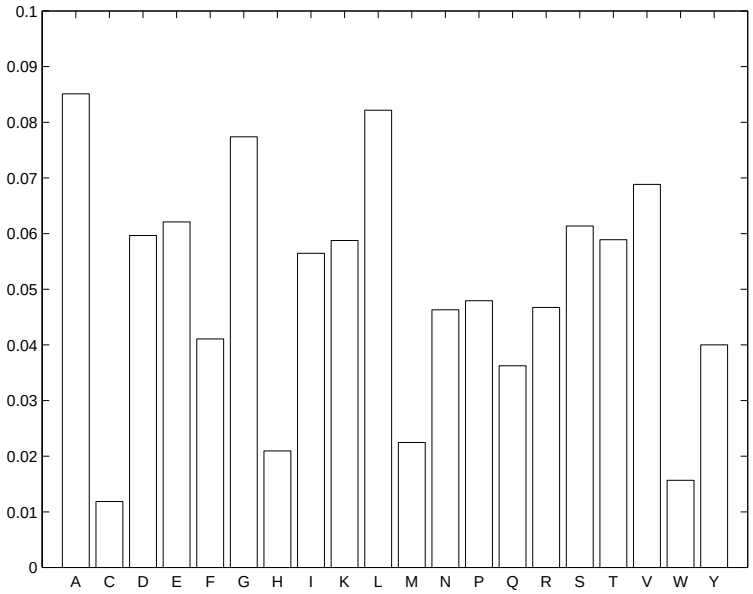

Figure 2: General amino acid frequencies in the data.

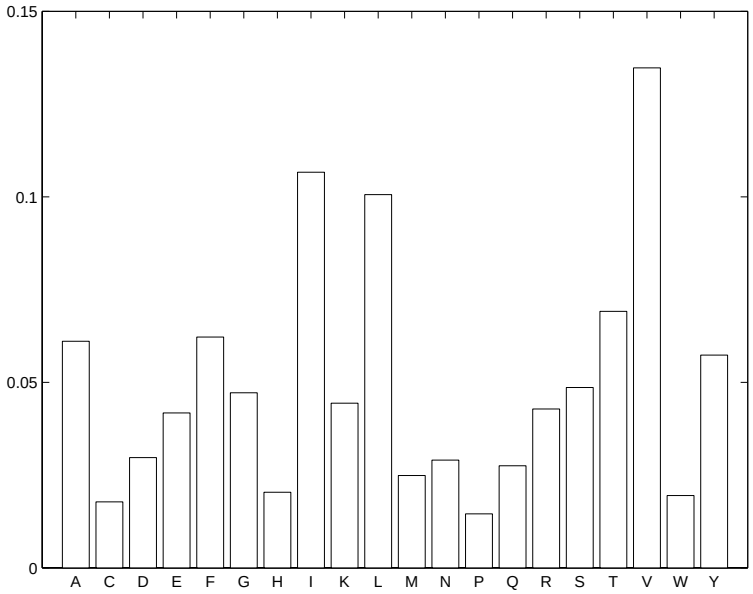

Figure 3: Amino acid frequencies in β-sheets.

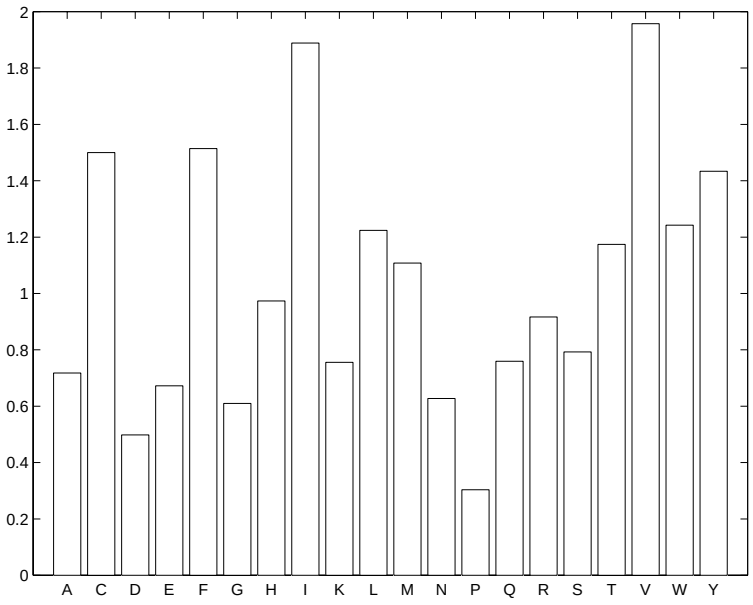

Figure 4: Ratio of amino acid frequencies: β-sheets/data.

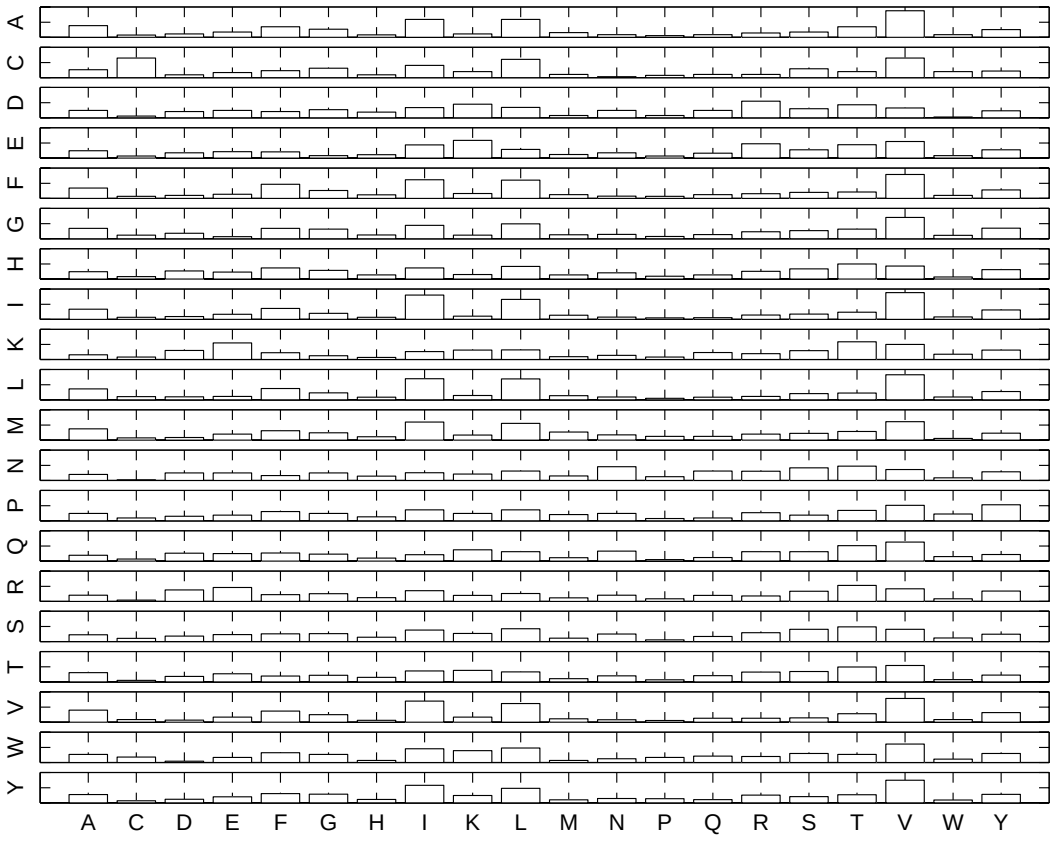

Figure 5: Second order statistics $P(X|Y)$. Conditional probability of observing an XY (or YX) pair in a β-sheet knowing that it contains a Y residue. The sum of the entries along any row is equal to one.

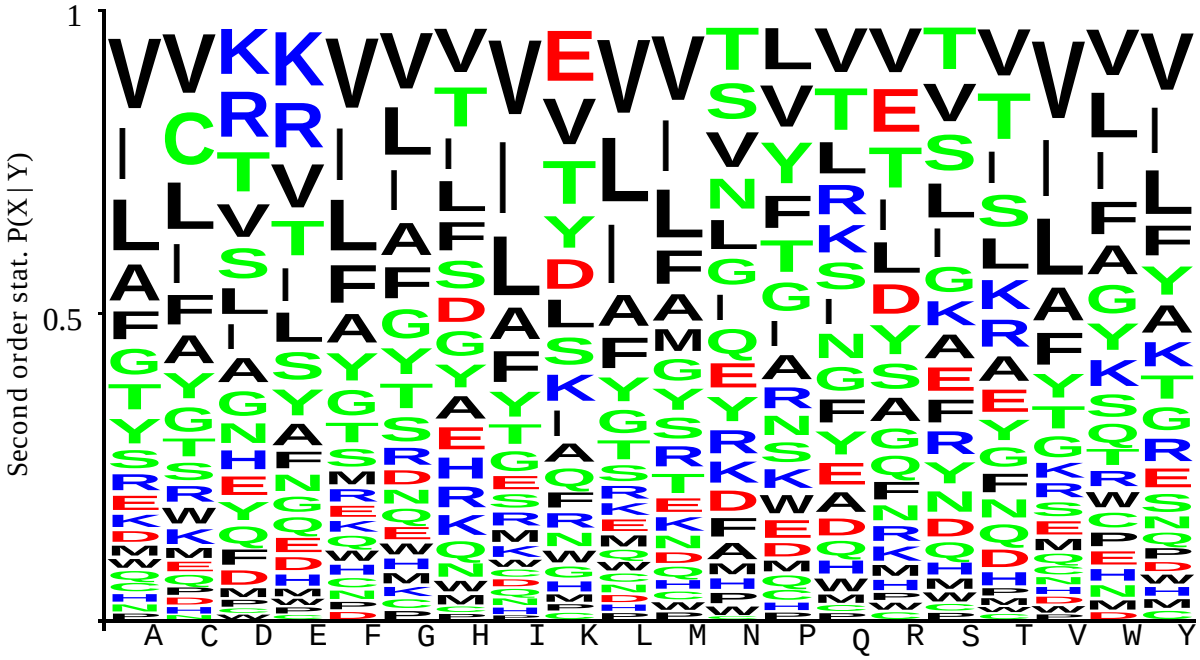

Figure 6: Second order statistics $P(X|Y)$. Conditional probability of observing an XY (or YX) pair in a β-sheet knowing that it contains a Y residue. Logo representation. The sum of each column is equal to one.

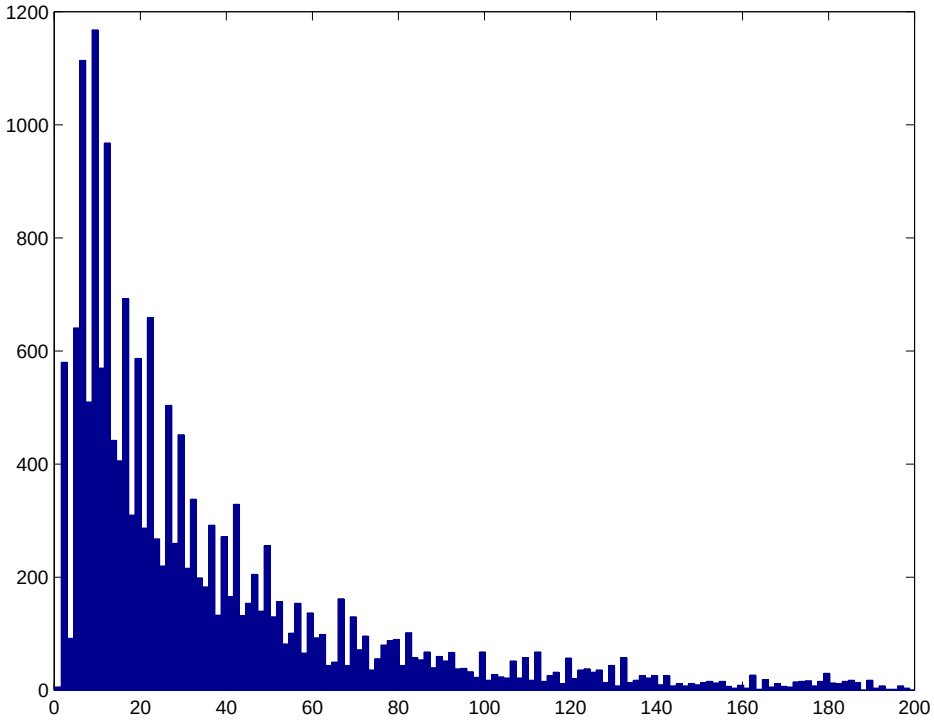

Figure 7: Histogram of distances between partner β-strands along protein chains, truncated to 200.

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.

We have also used a bi-directional recurrent neural network architecture (BRNN), as in (Baldi *et al.*, 1999). This approach can be first described in terms of the Bayesian network shown in Figure 9.

The architecture consists of an input layer where the amino acid sequence is presented, a forward Markov chain (as in standard Markov models of time series and HMMs), a backward chain, and an output layer consisting of all possible N^2 (only one is drawn in the Figure) pairings of identical but distant windows. The output variables correspond to classification into "paired" or "non-paired" categories. Because inference in these Bayesian networks is too slow, we replace the diagram by a recursive neural network, using the techniques described in detail in (Baldi *et al.*, 1999).

Experiments and Results

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

The raw data for our problem is extremely unbalanced. The number of negative examples (amino acid pairs that are not partners) is of course much higher, by a factor of roughly a 1,000. 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. In a typical case, 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. As a general rule, when overfitting begins all the systems we have trained tend to overfit the non-β partner class. We have empirically introduced slight variations in the training schedule as needed. We report also the results of five-fold cross validation tests.

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	129,119	64,017
Positive ex.	37,008	18,198
Negative ex.	44,032,700	22,920,100

Typical performance results are given below for dif-

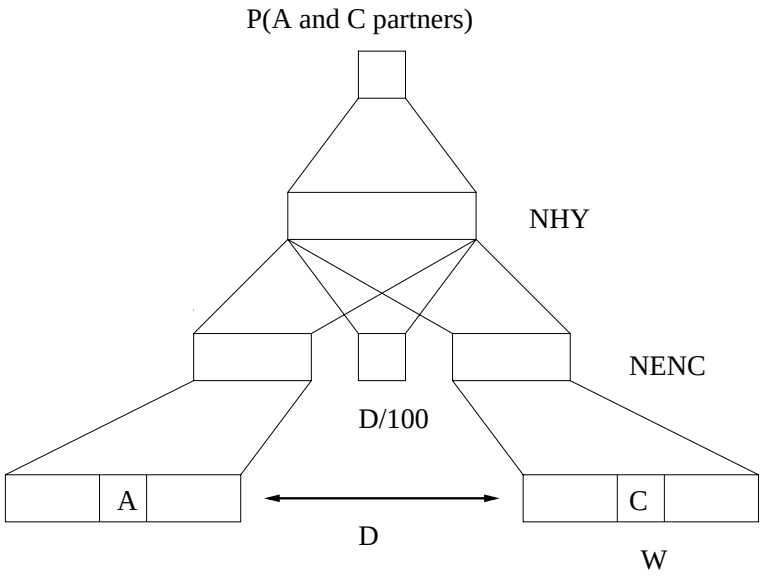

Figure 8: Neural network architecture for amino acid β-partner prediction.

ferent architectural variations. The percentages are computed on the entire test set, including all the negative examples it contains. The total balanced percentage is the average of the two percentages obtained on the positive and negative examples. It is different from the traditional total percentage obtained on the entire data sets, which in this case would not be meaningful. A trivial system that always predict a mismatch would have a percentage of about 99.98% correct in this sense.

The results of several training experiments using different variants of the same feedforward neural network architecture are summarized in Table 2. A typical network has 6500 parameters.

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 β-sheet partners and 70.43% on non-partners). A five-fold cross validation experiment using this architecture is reported in Table 3 and gives a slightly lower percentage of 83.07 %.

The predicted second order statistics of the artificial neural network with the best performance are displayed in Figure 10. The similarity is quite good although there are a few discrepancies, as in the case of double cysteine (C-C) partners. For fairness, these predicted statistics, as well as all the true statistics described above, were in fact computed on the test set only. We did check, however, that true statistics compute on the entire data set are very similar which is also a sign that the test set is representative of the problem.

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	82.44	84.73	83.59
10	12	7	82.31	84.36	83.33
12	11	7	83.41	82.30	82.86

Initial experiments carried with the BRNN architectures show further small but significant improvements. A typical BRNN architecture has a smaller number of parameters, of the order of 3500, but requires longer training times. We varied the size of the two input windows used to make the partnership assignment. Three values were used: 7, 9, and 11. The BRNN with input windows of length 7 yields again the best performance. On the 1/3-2/3 set, this BRNN classifies correctly in

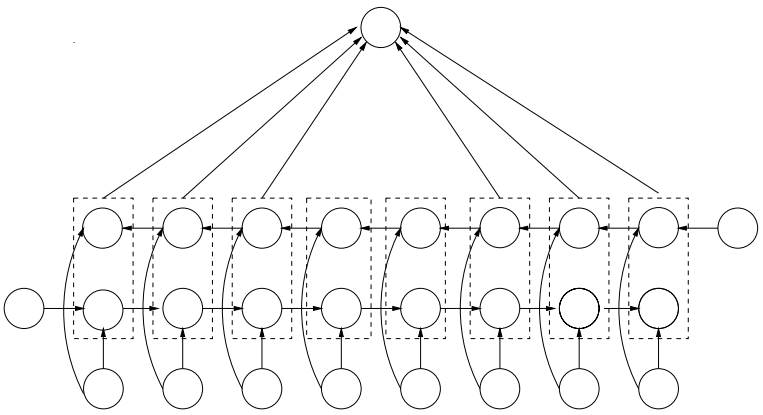

Figure 9: Bayesian network associated with the BRNN (see main text). Only one output node is shown. Two windows of states in both the forward and backward Markov chains are connected to the output unit. Each Markov chain as a unique starting state.

Table 3: Five-fold cross validation results obtained with the optimal NN architecture above. Performance results expressed in percentages of correct prediction. Overall percentage is the simple average of the percentage in each row or column.

Set	Beta	Non-Beta	Total
0	82.88	86.20	84.54
1	81.97	84.53	83.25
2	82.34	84.18	83.26
3	80.82	83.46	82.14
4	81.30	83.05	82.18
Total	81.86	84.28	83.07

84.3% of the cases, 0.7% better than the best static NN. As above, the five-fold cross validation results for this BRNN are slightly lower (83.54%) and given in Table 4. For comparison, the BRNNs architectures with amino acid windows of length 9 and 11 achieve five-fold cross validation percentages of 83.17% and 83.14% respectively.

Table 4: Five-fold cross validation results obtained with a BRNN architecture. Performance results expressed in percentages of correct prediction. Overall percentage is the simple average of the percentage in each row or column.

Set	Beta	Non-Beta	Total
0	82.04	87.26	84.65
1	81.15	85.88	83.51
2	81.80	85.48	83.64
3	81.71	84.33	83.02
4	80.52	85.22	82.87
Total	81.44	85.63	83.54

We then tested a few ensemble architectures, obtained by combining the previous ones. An ensemble made of the three BRNNs described above gives a further small improvement. On the 1/3-2/3 test, the ensemble of 3 BRNNS gives an average performance of 84.7% (83.3% on partners, and 86.09% on non-partners). The five-fold cross validation results for this ensemble are given in Table 5.

Table 5: Five-fold cross validation results obtained with an ensemble BRNN architecture, consisting of three BRNNs. Performance results expressed in percentages of correct prediction. Overall percentage is the simple average of the percentage in each row or column.

Set	Beta	Non-Beta	Total
0	82.01	87.04	84.52
1	81.62	86.36	83.99
2	82.24	85.81	84.03
3	81.82	84.55	83.18
4	81.05	85.18	83.12
Total	81.75	85.79	83.77

Interestingly, virtually no improvement was obtained by using ensembles of standard feedforward neural networks. Table 6 provides a summary of all the five-fold cross validation results for the best architectures. Additional experiments are in progress.

Table 6: Summary of five-fold cross validation results for the best architectures.

Architecture	Beta	Non-Neta	Total
Best ANN	81.86	84.28	83.07
Best BRNN	81.44	85.63	83.54
BRNN ensemble	81.75	85.79	83.77

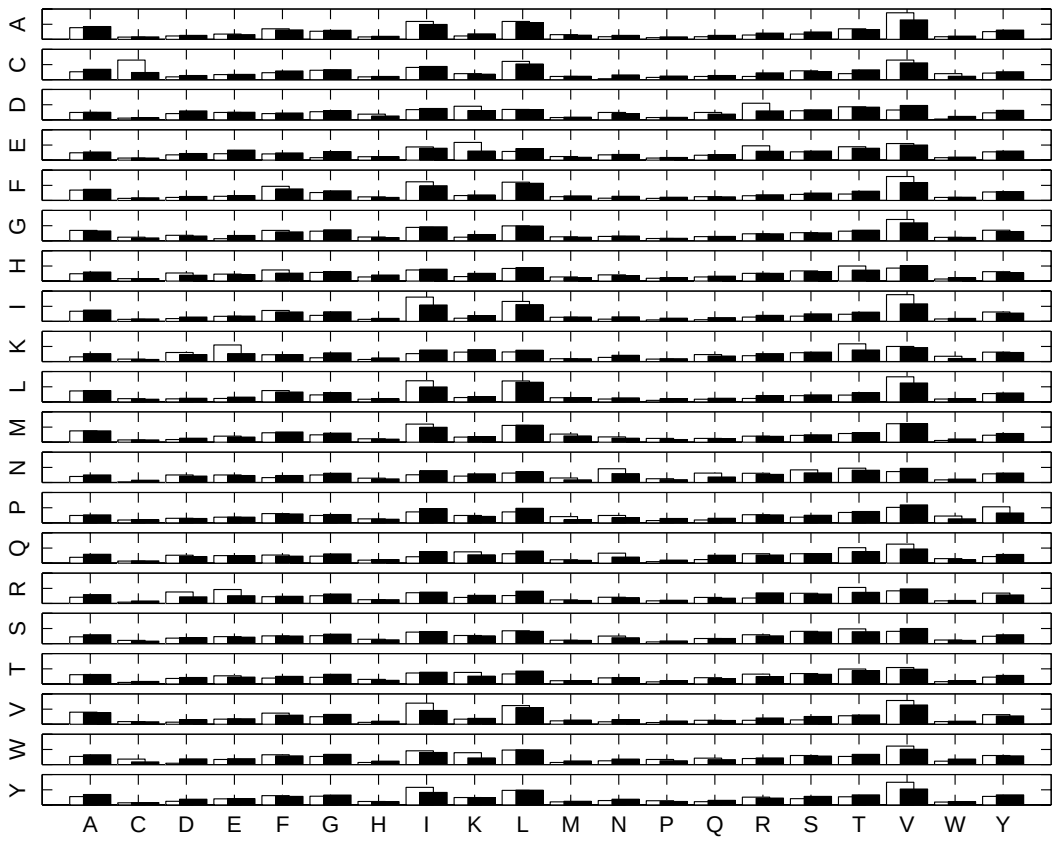

Figure 10: Comparison of true and predicted second order statistics. Black bars represent predicted frequencies. Empty bars represent true frequencies, as in Figure 5.

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 start testing these effects, we used the BLAST program (Altschul *et al.*, 1990) with standard default parameters to create multiple alignments of our sequences (BLOSUM matrix 62, Expectation value (E) = 10.0, Filter query sequence with SEG = True, Database = NR). The profile matrix was then used as input to the artificial neural network, instead of the sequence, in a retraining of the optimal architecture. As can be seen in Table 2, the overall performance of 83.59% is comparable, but not any better, to the performance of the feedforward neural network trained on sequences rather than profiles. This significant departure from what is observed in the case of general secondary structure prediction schemes seems to suggest that, in the case of β-sheets there are tradeoffs associated with the type of input. Profiles may provide more robust first order statistics, but weaker intrasequence correlations, and vice versa for the raw sequences. We are in the process of further testing these tradeoffs and trying to leverage the advantages of each approach.

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 (Wright & Dyson, 1999), that β-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. Nevertheless it is 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 β-sheet amino acid partners with a balanced performance close to 84% correct prediction, above previously

reported results.

β-strands are difficult to predict. Furthermore, reliable β-strand pairing would go a long way towards solving the protein structure prediction problem. The systems developed here can be viewed as a first step in this direction since they pair amino acids rather than strands. A balanced performance of 84% amino acid pairing prediction is insufficient by itself to reliably predict strand pairing because of the large number of false positive predictions in a large unbalanced data set. Our results, however, indicate several clear directions for future work towards improved prediction of β-strand partners and secondary structures.

Some of these directions are being explored and include:

- The testing of the effect of multiple alignments and profiles on the BRNNs architectures.
- The development of techniques to reduce the number of false positive predictions. Two experiments that need to be tried in this direction is to train similar systems on the partner/non-partner problem but on data extracted exclusively from beta sheets, ignoring α-helical and coil regions. Another even more restrictive possibility is to use only data associated with β-sheets that are neighbors along the primary sequence, and the inclusion of other categories, such as parallel or antiparallel strands.
- The construction and training of 20 partner prediction architectures, one per amino acid to try to further improve the quality of the match in Figure 10.
- The exploration of the use of raw sequence information, in addition to profiles, in secondary structure prediction systems for the β-sheet class.
- The use of the present architecture as a β-sheet predictor rather than a partner predictor, possibly in combination with another post-processing neural network.
- Various combinations of the present architectures with existing secondary structure predictor to improve β-sheet prediction performance.
- The combination of alignments with partner prediction in order to better predict β-strands.
- The possible leverage of additional information, such as amino acid properties (hydrophobicity, etc.)

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

Abagyan, R. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.*, **273**, 355–368.

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Baldi, P. & Brunak, S. (1998). Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge, MA.

Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. & Soda, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937–946.

Cuff, J. A. & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**, 508–519.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522.

Huang, X. & Miller, W. (1991). *Adv. Appl. Math.*, **12**, 337–357.

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

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

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

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

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

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

Street, A. G. & Mayo, S. L. (1999). Intrinsic β-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci. USA*, **96**, 9074–9076.

Wouters, M. & Curmi, P. (1995). An analysis of side chain interaction and pair correlation within antiparallel beta-sheets: The difference between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins*, **22**, 119–131.

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

Zhu, H. & Braun, W. (1999). 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.