Protein secondary structure prediction: efficient neural network and feature extraction approaches

J.C.B. de Melo, G.D.C. Cavalcanti and K.S. Guimarães

A simple and efficient approach to the protein secondary structure prediction problem is presented and evaluated with four established measures: Q_3, Matthews coefficients, $Q_{observed}$ and $Q_{predicted}$. They are applied to the raw data and also to features extracted with the PCA and the ICA methods. The results obtained are better than any predictor trained in similar conditions.

Introduction: In this Letter, an important step towards determining the 3D conformation of a protein is studied: the prediction of its secondary structure, i.e. recurrent arrangements in the tertiary structure that can be distributed in three classes: α-helix, β-strand and coils.

Different approaches have been proposed to this classical problem. Those that use machine learning methods, neural networks, in particular, have achieved the best results [1]. Nonetheless, the computational resources required are becoming higher and higher. On the other hand, it is common to have a new database developed for each new predictor, and the comparison of the performance of these predictors under different conditions can be misleading [1].

In an effort to switch that trend, we developed a protein secondary structure predictor, GMC [2], which ensembles three classifiers in a simple architecture. Expressive results were achieved using the established databases RS126 and CB396 [3] for training and testing the networks. Afterwards, a differential point was introduced in the GMC predictor: a preprocess phase for compressing the input data through one of the two methods: principal components analysis (PCA) [4] and independent component analysis (ICA) [5]. The significant Q_3 accuracy achieved in the three experiments motivates a more detailed analysis of their results, which is presented in this Letter.

Data set and evaluation method: The experiment was realised with the database developed by Cuff and Barton [3] labelled CB396. The database is composed by 396 dissimilar protein sequences.

In the evaluation method used, the set of proteins was split into M subsets, training the network with the $(M-1)N/M$ remaining proteins, and performing the test with the N/M removed proteins. The value used for M was seven, in order to make the experiments as close as possible to the ones applied by Cuff and Barton [3]. The results reported here are the average prediction accuracies from the seven different testing sets.

In this Letter, besides Q_3, that gives the percentage of correctly classified residues, three alternatives measures are reported: the Matthews correlation coefficient [6], $Q_{observed}$ and $Q_{predicted}$. The Matthews coefficient helps to minimise the effects of over- and under-predictions due to the inherited unbalanced distributions of the classes in the databank.

$$C_x = \frac{(p_x n_x) - (u_x o_x)}{\sqrt{(n_x + u_x)(n_x + o_x)(p_x + u_x)(p_x + o_x)}} \quad (1)$$

The Matthews coefficient is defined by (1), where $p_x = M_{x,x}$, $n_x = \Sigma_{j\neq x}^{3}\Sigma_{k\neq x}^{3} M_{j,k}$, $o_x = \Sigma_{j\neq x}^{3} M_{j,x}$, $u_x = \Sigma_{j\neq x}^{3} M_{x,j}$, x is one of the tree classes involved: Helix, Strand or Coil, and M is the confusion matrix. The coefficient C_x is in the range $+1$ (indicating classes totally correlated) to -1 (to indicate classes totally anti-correlated). The accuracy for each class x was evaluated through two per-state percentages $Q_{observed}$, represented by $Q_x^{\%obs}$ (2) and $Q_{predicted}$, represented by $Q_x^{\%pdr}$ (3).

$$Q_x^{\%obs} = \left(\frac{p_x}{p_x + u_x}\right) \times 100 \quad (2)$$

Considering all residues observed in a particular class x, $Q_x^{\%obs}$ gives the percentage of residues correctly predicted. For $Q_x^{\%pdr}$, this percentage is calculated considering all the residues predicted in a particular class.

$$Q_x^{\%pdr} = \left(\frac{p_x}{p_x + o_x}\right) \times 100 \quad (3)$$

Architecture and experiments: Three fully connected neural networks with one hidden layer were used. They were trained with the RPROP algorithm, using as input data PSI_Blast profiles [4, 5, 7] of the sequences in the same data set, CB396. The differential between the networks was the fact that each one had a distinct number of nodes (30, 35 or 40) in the intermediate layer, which were established after many experiments. The output layer had three nodes in all networks, one for each class (Helix, Strand and Coil). The rules used to combine the nets were: Voting, Product, Average, Maximum and Minimum. For the last four rules the function Softmax [8] was used to normalise the outputs of the neural networks (Fig. 1).

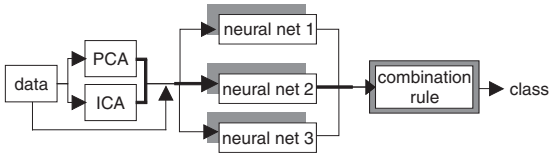

Fig. 1 *System architecture*

Three experiments were performed: one over the raw data [2], and two after feature extraction procedures, PCA [4] and ICA [5], respectively.

Results: The best results were reached with the ensemble of the three neural networks using the Product and the Average combination rules.

Using the Q_3 measure, the predictor achieved the best result reported in the literature using the raw data, 75.9%. For the experiments with a feature extraction phase, the results with ICA were, on average, 0.5% better than the results with PCA. The best result using PCA for feature extraction, 74.5%, was achieved when 180 principal components were informed to the networks. Using the same number of independent components, the Q_3 percentual for the ICA experiments was 74.9%. In all experiments the standard deviation was low, remaining in the range of 1.73 for the experiments with raw data, 1.57 when 180 principal components were used, and 1.83 for 180 independent components. In the three experiments, the GMC performance was superior to the best known result in the same conditions, 72.9%, obtained by CONSENSUS (a combination of four predictors, PHD, NNSSP, PREDATOR and DSC) [3].

Table 1: Comparative analysis between different implementations of GMC predictor and CONSENSUS using Mathews coefficients, $Q_{observed}$ and $Q_{predicted}$

Method	GMC (without PCA—Product rule)	GMC (PCA180—Product rule)	GMC (ICA180—Product rule)	CONSENSUS
C_α	0.70	0.67	0.67	0.63
C_β	0.61	0.58	0.59	0.55
C_{coil}	0.57	0.55	0.56	0.50
$Q_\alpha^{\%obs}$	81.4	79.9	80.2	70
$Q_\beta^{\%obs}$	73.8	71.6	74	55
$Q_{coil}^{\%obs}$	72.6	71.2	71.1	81
$Q_\alpha^{\%prd}$	79	78.4	77.8	81
$Q_\beta^{\%prd}$	64.4	62.3	61.9	73
$Q_{coil}^{\%prd}$	79.5	77.5	79.5	65

A comparison among the predictors that had presented the four best Q_3 accuracies over CB396 database is shown in Table 1. The measures used were Matthews coefficients, $Q_{observed}$ and $Q_{predicted}$. The values used for CONSENSUS, nowadays called JPred, were the most recent reported by EVA [9]. Comparing GMC with and without the feature extraction, the implementation with raw data obtained the best results for all the coefficients used, except for the measure $Q_{observed}$ applied to the data coming from the experiment with the ICA method. The Matthews coefficients of GMC were higher than the coefficients of CONSENSUS for all the classes. The percentages $Q_{observed}$ were better for GMC except for the class Coil, where CONSENSUS had a percentage 8% better. For the $Q_{predicted}$ measure the percentage for Coil class was better than for the GMC predictor. Regarding the Strand and Helix classes, the CONSENSUS obtained higher value, but the results for Helix class remain comparable. Table 2 reveals the Matthews correlation coefficients for the best predictors available nowadays. The GMC predictor trained with raw data presented C_α and C_{coil} superior to

the others. Only the C_β coefficient was worse. Using PCA as a preprocessing phase, the coefficient C_α is inferior only that one of GMC. The C_{coil} is in the average of other predictors and C_β presented the less relevant result.

Table 2: Comparative analysis between protein structure prediction methods using Mathews coefficients

Method	C_α	C_β	C_{coil}
GMC (without PCA—Product Rule)	0.70	0.61	0.57
GMC (ICA180—Product Rule)	0.67	0.59	0.56
GMC (PCA180—Product Rule)	0.67	0.58	0.55
PROF	0.67	0.65	0.56
SSpro	0.67	0.64	0.56
PSIPRED	0.66	0.64	0.56
PHDpsi	0.64	0.62	0.53
CONSENSUS	0.63	0.55	0.50
PHD	0.59	0.59	0.49

Conclusion: A simple and efficient secondary structure predictor was presented and evaluated. Three experiments were performed, one with the raw data and two others using PCA and ICA for feature extraction.

Considering the raw data, the best Q_3 performance for predictors trained with the CB396 was achieved, 75.9%. Reducing the dimension with PCA and ICA, the accuracy maintained good levels, only 2% inferior when 30% of the original data was informed to the networks. The analysis with Matthews correlation coefficients revealed that the GMC with raw data achieved the best C_α and C_{coil} indices in the literature, being the value of the C_β index comparable to the others. The index for the experiment with PCA kept good levels for C_{coil} and C_α, overcome only by GMC without feature extraction. For ICA experiments, the C_{coil} and C_β values were slightly better than those obtained with PCA. For all measures used, the GMC predictor presented an excellent performance for helices. But this fact was not observed for the class Strand, owing to the long-range interactions, motivating new research directions in order to avoid this problem. The combination that presented the best results for the GMC predictor was implemented as a web server, and can be accessed at http://biolab.cin.ufpe.br/tools.

Acknowledgments: J.C.B. Melo and K.S. Guimarães thank the Brazilian sponsoring agency CNPq.

15 June 2004
Electronics Letters online no: 20045764
doi: 10.1049/el:20045764

J.C.B. de Melo, G.D.C. Cavalcanti and K.S. Guimarães (*Center of Informatics, Federal University of Pernambuco, P.O. Box 7851, Recife—PE 50.732-970, Brazil*)

J.C.B. de Melo: Also with Physics and Mathematics Department, Federal Rural University of Pernambuco, Dois Irmãos, Recife—PE, 52.171-900, Brazil

References

1 Rost, B.: 'Review: protein secondary structure prediction continues to rise', *J. Structural Biology*, 2001, **134**, pp. 204–218
2 Guimarães, K.S., Melo, J.C.B., and Cavalcanti, G.D.C.: 'Combining few neural nets for effective secondary structure prediction'. Proc. 3rd IEEE Symp. on Bioinformatics and Bioengineering, Maryland, USA, pp. 415–420
3 Cuff, A.J., and Barton, J.G.: 'Evaluation and improvement of multiple sequence methods for protein secondary structure prediction', *Proteins Struct. Function Genet.*, 1999, **34**, pp. 508–519
4 Melo, J.C.B., Cavalcanti, G.D.C., and Guimarães, K.S.: 'PCA feature extraction for protein structure prediction'. Proc. of IEEE 2003 Int. Joint Conf. Neural Networks, Oregon, USA, pp. 2952–2957
5 Melo, J.C.B., Cavalcanti, G.D.C., and Guimarães, K.S.: 'Protein secondary structure prediction with ica feature extraction'. Proc. 2003 IEEE Int. Workshop on Neural Networks for Signal Processing, Special Section on Bioinformatics, Toulouse, France, pp. 13–22
6 Matthews, B.W.: 'Comparison of predicted and observed secondary structure of T4 phage lysozyme', *Biochem. Biophys. Acta.*, 1975, **405**, pp. 442–451
7 Altschul, S., *et al.*: 'Gapped Blast and Psiblast: a new generation of protein database search programs', *Nucleic Acids Res.*, 1997, **25**, pp. 3389–3402
8 Duda, R.O., Hart, P.E., and Stork, D.G.: 'Pattern classification' (John Wiley & Sons, New York, 2001)
9 Eyrich, V.A., *et al.*: 'EVA: continuous automatic evaluation of protein structure prediction servers', *Bioinformatics*, 2001, **17**, (12), pp. 1242–1243