On the Role of Long-Range Dependencies in Learning Protein Secondary Structure

Alessio Ceroni and Paolo Frasconi

Dipartimento di Sistemi e Informatica
Università di Firenze, 50139 Firenze, Italy
E-mail: {aceroni,paolo}@dsi.unifi.it

Introduction

Prediction of protein secondary structure (SS) is a classic problem in computational molecular biology and one of the first successful applications of machine learning to bioinformatics. Most available prediction methods use feedforward neural networks whose input is the multiple alignment profile in a sliding window of residues centered around the target position [6, 4]. By construction, predictions obtained with these methods are local. Long-range dependencies, on the other hand, clearly play an important role in this problem. In [2] it was proposed the use of bidirectional recurrent neural networks (BRNN) for the prediction of SS. The architecture in this case allows us to process the sequence as a whole and to "translate" the input profile at each position into a corresponding output prediction for that position. Theoretically, the output at any position in a BRNN depends on the entire input sequence and thus a BRNN might actually exploit long-range information. Unfortunately, well known problems of vanishing gradients do not allow us to *learn* these dependencies.

In this paper, we are interested in developing an architecture that can effectively exploit long-range dependencies assuming some additional information is available to the learner. We start from a rather simple intuitive argument: if the learner had access to information about which positions pairs are expected to interact, its task would be greatly simplified and it could possibly succeed. In the case of SS prediction, a reasonable source of information about long-range interaction can be obtained from contact maps (CM), a graphical representation of the spatial neighborhood relation among amino acids. Of course in order to obtain a CM the protein structure must be known. In addition, it is well known that backbone atoms' coordinates can be reconstructed starting from CMs [7]. Thus, in a sense, using CM information in order to predict SS might appear foolish since most of the information about the 3D structure of the protein is already contained in the map. However, the following considerations suggest that this setting is worth investigation:

- Algorithms that reconstruct structure from CMs are based on a potential energy function with many local minima whose optimization is not straightforward [7]. Thus it is not clear that a supervised learning algorithm can actually *learn* to recover SS from CMs.

- CMs can be predicted from sequence [3] or can be obtained from structures predicted by ab-initio methods such as Rosetta [1]. Although accuracy of present methods is certainly not sufficient to provide a satisfactory solution to the folding problem, predicted maps may still contain useful information to improve the prediction of lower order properties such as the SS.

- Even if CMs are given, the design of a learning algorithm that can fully exploit their information content is not straightforward. For example, Meiler and Baker [5] have shown that SS prediction can be improved by using information about inter-residue distances. Their architecture is a feedforward network fed by *average* property profiles associated with amino acids that are near in space to the target position. In this way, relative ordering among neighbors in the CM is discarded.

The solution proposed in this paper is based on an extended architecture that receives as an additional input a graphical description of the pairwise interactions between sequence positions. We call this architecture interaction enriched BRNN (IEBRNN). Its details are presented in a longer version of this paper.

Results

Prediction from sequence alone

A first set of experiments was performed to obtain a baseline prediction accuracy for the SS problem on this dataset. A "bare bones" BRNN classifier with multiple alignment profiles as input was used for this purpose. Results of the baseline experiment are reported in the upper-left corner of Table 1.

Prediction from CM alone

Here we want to estimate the amount of information about SS the IEBRNN architecture is capable to learn from contacts alone. As reported in Table 1, we obtained $Q_3 \approx 80\%$ and SOV $\approx 73\%$. These results are quite impressive considered that the classifier has no knowledge about the 3D conformation of the hydrogen bonded atoms and the physics behind the formation of SS.

Table 1: Performances of the various methods presented in this paper.

BRNN			IEBRNN		
	Q_3	SOV		Q_3	SOV
Profiles only	74.6%	66.7%	Interactions only	79.9%	73.3%
Profiles + context	82.5%	77.4%	Profiles + interactions	84.6%	79.3%
Profiles + context (no γ)	95.9%	94.6%	Profiles + interactions (no γ)	97.9%	95.5%

Prediction from profiles and contacts together

In this experiment, we trained the IEBRNN with both profiles and contacts as input. For sake of comparison with the results of Meiler and Baker [5] we also trained a standard BRNN with the same kind of inputs they used. In particular, the spatial *context* of each residue was computed by averaging the profile of the amino-acids in a sphere of 6 Å centered on the residue itself. This additional input was then given to the standard BRNN together with the usual profile. This method can be seen as a simplified version of the IEBRNN in which all contacts give the same contribution to a given position. In particular, order among contacts cannot be distinguished. As can be seen from Table 1, the information about contact ordering is efficiently exploited by the IEBRNN, which appreciably outperforms the solution based on the average context. Interestingly, prediction accuracy improves for helices and strands but not for coils.

Effects of interaction robustness

The results of the above experiments show us that IEBRNNs can effectively exploit the information contained in the CM to improve prediction accuracy. However there is still about 15% residual error rate that would be interesting to explain. We conjecture that the reliability of the interactions that were injected as an additional input may play a significant role. In facts, edges in a CM express spatial proximity but do not necessarily imply dependencies between the two close residues. This may be particularly true in the case of contacts that involve coil residues. Instead, contacts between residues that both belong to helices or strands can be expected to encode interactions in a more robust way since they are often maintained by hydrogen bonds.

In order to evaluate the effect of interaction robustness we repeated our experiments using more sparse CM where edges only connect residues that belong to helices or strands. In so doing, we removed about 60% of the edges from interaction graphs. Results are reported in the last row of Table 1 both for the standard BRNN (fed by profiles and context) and for the IEBRNN. The error reduction obtained in this way is dramatic. The residual error is comparable to the disagreement between different SS assignment programs (i.e. DSSP, STRIDE, and DEFINE). These experiments could indicate that part of the information contained in the CM can be misleading, especially for the shorter segments.

References

[1] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–6, 2001.

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

[3] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps with neural networks and correlated mutations. *Prot. eng.*, 14:835–843, 2001.

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

[5] J. Meiler and D. Baker. Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci U.S.A.*, 2003.

[6] B. Rost and C. Sander. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. USA*, 90(16):7558–7562, 1993.

[7] M. Vendruscolo, E. Kussel, and E. Domany. Recovery of protein structure from contact maps. *Fold. Des.*, 2:295–306, 1997.