

PROTEIN STRUCTURE ASSEMBLY FROM KNOWLEDGE OF β -SHEET MOTIFS AND SECONDARY STRUCTURE

Alessio Ceroni, Paolo Frasconi and Alessandro Vullo
Dipartimento di Sistemi e Informatica, Università di Firenze, Italy
{aceroni,p-f,vullo}@dsi.unifi.it

Abstract We develop and test a new hierarchical approach for the prediction of protein structure. An algorithm is described to assemble the 3D fold of a protein starting from its secondary structure and β -sheet topology. Reconstruction is carried out by energy minimization of a reduced protein model, where β -partners are derived from appropriate distance constraints imposed by the knowledge of β -sheet motifs. Additional constraints are imposed in the (ϕ, ψ) torsion space from secondary structure knowledge. Experiments show how the proposed procedure proves to be a reliable and fast predictive approach for a large fraction of proteins of interest. Arrangements of β -sheets are predicted with special recursive neural networks architectures. We first present a unifying framework for description of a large class of contextual recursive models and then show how it is possible to solve the problem at some extent of success.

Keywords: Protein structure prediction, β -sheets prediction, recursive neural networks, protein structure reconstruction.

1. Introduction

Knowledge of the spatial conformation of a protein can help the study of its function. Unfortunately, the number of resolved structures is still limited by the low throughput of available experimental methods. Prediction tools have the potential to bridge the sequence-structure gap, but no reliable and general methods have yet been proposed. Attempts to simplify the problem have been made by trying to predict the contact map of a protein instead of its atomic positions. It has been demonstrated how good protein models can be derived even with noisy contact maps [Vendruscolo et al., 1997]. Unfortunately, prediction of contact maps is still very unreliable and it is not clear whether the type of errors made by the predictor can be corrected by the reconstruction method. In the attempt to train more efficient predictors, a low-detail representation of protein conformation could extract high level relevant information. Prediction of coarse-grained contact maps, i.e. contacts are defined among secondary

structure segments, has been recently tried [Vullo and Frasconi, 2003], but yet there exists no clear result concerning feasibility of reconstruction using only coarse information.

In this work, we elaborate on efficient and reliable prediction methodologies tailored for a specific class of chains, those that are mainly characterised by residues in strand conformation. The quality of reconstruction for this kind of proteins can be enhanced by the knowledge of secondary structure and indication of which strands are partners. We focus on contacts defined on β -partners, as the geometry and connectivity of β -strands imposes strong constraints on the overall structure of the protein. In section 2 we propose an efficient procedure to find a structure that matches the aforementioned characteristics of a given protein in its native conformation. In order to fully automate structure prediction, in section 3 we also describe an approach for predicting β -sheet motifs using a powerful class of connectionist models. Finally, in section 4 we describe our experiments and show encouraging results in both directions.

2. Backbone Reconstruction Algorithm

The reconstruction procedure performs energy minimization of a reduced protein model, where knowledge about secondary structure and β -partners in the native conformation is enforced as a set of constraints on candidate solutions. The protein model comprises all backbone heavy atoms plus a single atom for the C_β to represent side chain occupation. Free parameters of this model are the dihedral ϕ and ψ angles. The ω angle is set fixed to 180° , while coordinates of the C_β atom and all bond lengths and angles are set to their average values calculated on the PDB dataset.

Constraints on Protein Structure

Secondary structure information is enforced by constraining the values of the dihedral angles: α -helices (H) and β -strands (E) correspond to two compact regions in the $\phi - \psi$ plot. For every residue in the H and E classes, the distance between its coordinates in the (ϕ, ψ) space and the center of the corresponding region is forced to be lower than a specified threshold:

$$\|(\phi, \psi) - (\phi_s, \psi_s)\| \leq t_s \quad (1)$$

where $s \in \{H, E\}$. For each pair of β -strands we know if they are partners and in this case if they are parallel or anti-parallel. The geometry of two β -partners constrains the distance between hydrogen-bonded residues. Unfortunately, two partner strands can be of different dimensions and we do not want to specify the partnership in terms of connectivity between residues. Let I and J be the sequences of indexes of the residues in two β -strands and I^k and J^k two subsequences of size k . An alignment with parallel orientation is the set

$\{(I_1^k, J_1^k), \dots, (I_k^k, J_k^k)\}$, while an alignment with anti-parallel orientation is the set $\{(I_1^k, J_k^k), (I_1^k, J_{k-1}^k), \dots, (I_k^k, J_1^k)\}$. The procedure must test all possible alignments for each pair of strands. For partner strands, given a particular alignment, the distance between every pair of (supposedly) bonded atoms must be in a strict range of values

$$\forall_{i \in 1, k} : d_{min}^b \leq \| \bar{x}(I_i^k) - \bar{x}(J_i^k) \| \leq d_{max}^b \quad (2)$$

and the alignment that violates less constraints contributes to the solution: this enforces the existence of at least one good alignment between partners. For non-partner strands both orientations are tested; given a particular alignment, the distances between paired atoms must be greater than a specified value

$$\forall_{i \in 1, k} : \| \bar{x}(I_i^k) - \bar{x}(J_i^k) \| > d_{min}^{nb} \quad (3)$$

and the alignment that violates more constraints contributes to the solution; no good alignments must exist between non-partners. Atomic forces impose a lower bound on the distance between two atoms, thus defining an excluded volume for each atom that prevents the protein to collapse in a single point. We introduced these constraints in our procedure by forcing the distances between all pairs of atoms to be higher than a specified threshold:

$$\forall_{i \in 1, k} : \| \bar{x}(I_i^k) - \bar{x}(J_i^k) \| > d_{min}^{nb} \quad (4)$$

Optimization

In order to simplify the optimization task, all the constraints are expressed as quadratic penalty terms:

$$d_{min} \leq d \leq d_{max} \rightarrow \begin{cases} (d - d_{min})^2 & d < d_{min} \\ (d - d_{max})^2 & d > d_{max} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Unfortunately, this gives rise to a highly non linear function of the model free parameters. Global optimisation of non-linear cost functions is generally a difficult task, but we adopt here a simple approach consisting in a quasi-newton local optimization procedures (LBFGS [Liu and Nocedal, 1989]) coupled with a multistart strategies. Our non-linear cost function has in general many local minima. To mitigate this problem, we implement a specific protocol during optimization: firstly, we optimise the cost function with only secondary structure constraints, so that β -strands are formed in the backbone; secondly, we add the constraints for β -partners, relaxing the constraints on secondary structure so that a matching conformation is more easily found; finally, we add the constraints on atomic volumes, that could form barriers and prevent parts of the backbone to reach their final positions.

3. Prediction of β -sheet Motifs

As shown in the previous section, our reconstruction procedure needs a bunch of important ingredients: secondary structure, hence the location of β -strands, and the arrangement of β -sheets in the protein must be known. All these information are obtainable either from the PDB files, when the structure is known, or from predictions. Clearly, the former case is of limited interest and it is considered here to analyze upper bound performance of reconstruction. We assume here that secondary structure is known, for instance using one of several successful methods developed for this problem [Jones, 1999; Baldi et al., 1999], and focus on the more difficult task of prediction of β -sheet configurations.

The β -sheets motifs inference problem is modelled as a multi-class classification task. Assume we are given a set $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ of strands, each pair of strands (s_i, s_j) defined on \mathcal{S} can be mapped to one of three possible labels: say 0 (no hydrogen bonds between s_i and s_j), -1 (s_i and s_j are antiparallel) and 1 (parallel strands). The connectivity matrix on \mathcal{S} represents the β -sheets motif and is defined as a matrix C whose elements $c_{ij} \in \{0, -1, 1\}$ correspond to the pairs (s_i, s_j) . Here we model connectivity matrices as two-dimensional (square) undirected lattices, where nodes correspond to pairs of β -strands and edges connect adjacent pairs. Modelling β -sheets motifs in this way naturally gives rise to complex structured (graphical) representations than simple fixed size attribute-value pairs. The main advantage of using structured data is the possibility to encode intrinsic dependencies among atomic entities allowing powerful learning algorithms to be employed. The approach adopted here for predicting the connectivity matrix resemble those of [Pollastri and Baldi, 2002; G.Pollastri et al., 2003], where contextual Recursive Neural Networks (RNNs) are used to predict contact maps defined at the amino acid or segment level. In the following, we propose a unifying view of contextual RNNs and derive the predictive architecture used for the present case.

RNNs for undirected graphs

A data structure is a graph whose nodes are marked by sets of domain variables, called labels. A skeleton class, denoted by the symbol $\#$, is a set of unlabeled graphs that satisfy some topological conditions. Let \mathcal{I} and \mathcal{O} denote two label spaces: $\mathcal{I}^\#$ (resp. $\mathcal{O}^\#$) refers to the space of data structures with vertex labels in \mathcal{I} (resp. \mathcal{O}) and topology $\#$. Recursive models such as RNNs [Frasconi et al., 1998] can be employed to compute functions $\mathcal{T} : \mathcal{I}^\# \rightarrow \mathcal{O}^\#$ which map a structure into another structure of the same form but possibly different labels. In the classical framework, $\#$ is contained in the class of bounded DPAGs, i.e DAGs where each vertex has bounded outdegree and whose children are ordered. Recursive models put a causality assumption on data pro-

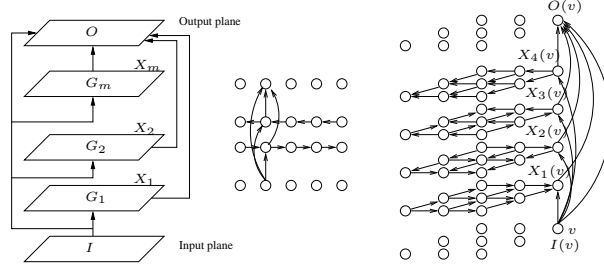


Figure 1. (left): Contextual RNNs, dependencies among input, state and output variables. (center and right): processing of undirected sequences and grids with contextual RNNs (only a subset of connections are shown).

cessing: structures are processed bottom-up according to a reverse topological order of the nodes. Therefore, the state variables associated to these nodes and then their outputs depend only on the sub-structures induced by their children. The above assumption imposes some restrictions on the amount of contextual information that can be tackled and extensions of these models for dealing with more general undirected structures have been proposed [Baldi et al., 1999; Polastri and Baldi, 2002; Vullo and Frasconi, 2003].

A more general assumption is considered here: $\#$ is contained in the class of bounded-degree undirected graphs. In this case, there is no concept of causality and the computational scheme described in [Frasconi et al., 1998] cannot be directly applied. The strategy consists in splitting graphical processing into a set of causal “dynamics”, each one computed over a plausible orientation of U . More formally, assume $U = (V, E) \in \mathcal{I}^\#$ has one connected component. We identify a set of spanning DAGs G_1, \dots, G_m with $G_i = (V, E_i)$ such that:

- the undirected version of G_i is U
- $\forall v, u \in V v \neq u \exists i : (v, u) \in E_i^*$ is E_i^* the transitive closure of E_i

and for each G_i , introduce a state variable X_i computed in the usual way. Fig.1 (left) shows a compact description of the set of dependencies among the input, state and output variables. Connections run from vertices of the input structure (layer I) to vertices of the spanning DAGs and from these nodes to nodes of the output structure (layer O). Using weight-sharing, the overall model can be summarized by $m + 1$ distinct neural networks implementing the output function $O(v) = g(X_1(v), \dots, X_m(v), I(v))$ and m state transition functions $X_i(v) = f_i(X_i(ch_1[v]), \dots, X_i(ch_k[v]), I(v))$. Learning can proceed by gradient-descent (back-propagation) due to the acyclic nature of the underlying graph. Within this framework, we can easily describe all contextual RNNs architecture developed so far. Fig.1 (center) shows that an undirected sequence is

spanned by two sequences oriented in opposite directions. We then obtain bi-directional recurrent neural networks [Baldi et al., 1999], or bi-recursive neural networks [Vullo and Frasconi, 2003] if we consider generic undirected graphs. Our is the case of two-dimensional grids, which can be seen as spanned by four directed grids oriented from each cardinal corner (Fig.1, right). The corresponding model is called 2D DAG-RNNs [Pollastri and Baldi, 2002].

4. Experimental Protocol

The experiments were performed using a representative set of non homologous chains from the Protein Data Bank (PDBSelect, december 2002). For every chain we determined the secondary structure class using the same procedure employed for the CATH database [Orengo et al., 1997]. The final dataset contained only *mainly- β* proteins, for a total of 154 chains whose sequence length is between 30 and 300 residues.

Reconstruction from true and predicted β -sheet motifs

We first tested whether our reconstruction procedure is able to reproduce β -sheet motifs of real protein structures. Accuracy was measured as the proportion of pairs of β -strands correctly assigned as partners or non-partners. The average value obtained was 98.5%, with 74% of test proteins with all β -partners correctly assigned. We then tested whether knowledge of β -sheet motifs is sufficient to reconstruct protein native conformations with good quality. We used two measures of quality: the RMSD calculated on the C_α atoms for all the amino-acids in strands, and the GDT_TS measure adopted in the CASP contest [Zemla et al., 2001]. We obtained an average RMSD value of 7.55 Å, and an average GDT_TS of 29.7. The distribution of those measures in the whole data-set is shown in Fig. 2 (left). We then performed the same test on the β -sheet motifs as predicted from the recursive model (see sec.2). In this case, the average number of correctly assigned β -strands pairs dropped to 75%, since the predictor is likely to produce unrealistic structures. The average value of the RMSD became 16 Å, while the average value of GDT_TS was 24.

Prediction of β -sheet motifs

Together with contextual RNNs for grids, we trained and tested multi-layered feed-forward neural networks (FF-NNs) to predict the class of contact for the (i, j) pair of β -strands. In either cases, input was represented by merging an attribute-value representation for the i -th and j -th strand in the sequence. Segments were described by 23-dimensional feature vectors including the average multiple sequence alignment profile, the relative index of one strand and its normalized start and end amino acid positions. In these experiments, we ap-

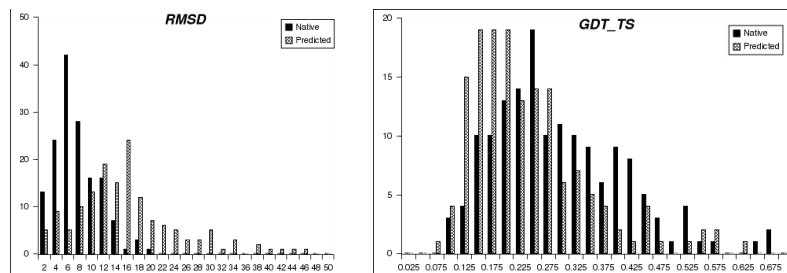


Figure 2. Histograms showing the distribution of RMSD (left) and GDT_TS (right) on the set of reconstructed structures using native and predicted β -sheets topologies.

plied a 5-fold cross validation procedure. In order to control overfitting we applied early-stopping, using for each fold part of the available training data as validation set. After cross validating the two methods, we obtained the results indicated in Table 1. We report several indices together with their 95% confidence intervals: micro-averaged global classification accuracy (Q_3) and accuracy of prediction of anti-parallel (Q_{ap}), parallel (Q_p) and unpaired strands (Q_{nb}). The index Q_3 indicates consistently higher performance than a baseline approach: predicting the class randomly, but with the frequencies observed in the training sets would lead to an expected 72.6% accuracy. It clearly results that contextual RNNs predict anti-parallel and parallel strands consistently better than simple non-recursive nets. The latter model tends to predict the more numerous class in the majority of cases and then it shows better global results (Q_3). A trivial predictor always assigning a pair of strands to the more numerous class (non-contact) would achieve 84.0% but with no use. Finally, the values of Q_{ap} and Q_p give a measure of the difficulty of the problem, the major obstacle being the unbalance among the classes: anti-parallel and parallel pairs represent 13.99% (resp. 1.96%) of the total set of pairs.

| Method | Q_3 | Q_{ap} | Q_{nb} | Q_p |
|--------|--------------|----------------|--------------|---------------|
| C-RNNs | 80.5 ± 9 | 39.1 ± 3.0 | 89.1 ± 8 | 7.1 ± 4.2 |
| FF-NNs | 85.1 ± 8 | 24.2 ± 2.6 | 97.2 ± 4 | 0.0 ± 0 |

Table 1. Experimental comparison of contextual RNNs and feed-forward neural nets for the problem of predicting β -strands pairings.

5. Concluding Remarks

Experimental results demonstrates how the proposed approach is able to build protein models matching the available characteristics of a native confor-

mation. The proposed algorithm is inherently fast – reconstruction takes on average 20 minutes on standard workstations – because it is based on an efficient local optimization procedure combined with a multi-start strategy. By this, differently from other de-novo methods, we were able to test our approach over a large non-redundant set and statistically significant quality measures were obtained. Moreover, reconstruction of β -sheets topology does not require fine-grained information about contacts between single residues. Therefore, our algorithm can be used even if there are incomplete information about the native structure, e.g. during NMR modelling. Unfortunately, the reconstructed structures are quite distant from corresponding native conformations, but we believe the quality of reconstruction could be improved by the addition of different types of contacts between secondary structure elements.

We also explored the case in which nothing is known about strands topology. We built a predictor of partnership between β -strands using recursive neural networks. Reconstruction was then tested using the predicted topology instead of the real one. Unfortunately, the algorithm did not prove to be sufficiently reliable to correct the errors of the predictor, which in turn can be substantially improved with the use of richer input descriptions. This led to a substantial decrease in the quality of reconstruction compared to the previous case. However, we have no knowledge of similar experiments being conducted before, so this can be considered as a first step toward a complete reconstruction procedure based on coarse-grained information alone.

References

- Baldi, P., Brunak, S., Frasconi, P., Pollastri, G., and Soda, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946.
- Frasconi, P., Gori, M., and Sperduti, A. (1998). A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, 9(5):768–786.
- G.Pollastri, Baldi, P., Vullo, A., and Frasconi, P. (2003). Prediction of protein topologies using GIOHMMs and GRNNs. *Advances in Neural Information Processing Systems*, 15.
- Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J.Mol.Biol.*, 292:195–202.
- Liu, D.C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997). CATH - A hierarchic classification of protein domain structures. *Structure*, 5:1093–1108.
- Pollastri, G. and Baldi, P. (2002). Prediction of contact maps by recurrent neural network architectures and hidden context propagation from all four cardinal corners. *Bioinformatics*, 18(Supplement 1):S62–S70.
- Vendruscolo, M., Kussel, E., and Domany, E. (1997). Recovery of protein structure from contact maps. *Fold. Des.*, 2:295–306.
- Vullo, A. and Frasconi, P. (2003). Prediction of protein coarse contact maps. *J. Bioinf. Comp. Biology*, 1(2):411–431.
- Zemla, A., Venclovas, C., Moulton, J., and Fidelis, K. (2001). Processing and evaluation of predictions in CASP4. *Proteins*, 5:13–21.