Poster H-88

Porter+: A Server for Protein Structural Motif Prediction

Authors:
Catherine Mooney (*UCD*)
Gianluca Pollastri (*UCD*)
Alessandro Vullo (*UCD*)

Short Abstract: Porter+ is a server for the prediction of a new alphabet of local structural motifs. The predictive system, based on ensembles of bi-directional recurrent neural networks, and trained on a large non-redundant set of protein structures, achieves 60% correct motif prediction for tetra-peptides into 14 classes.

Long Abstract:

Background:

A significant step towards establishing the structure and function of a protein is the prediction of the local conformation of the polypeptide chain. In a recent study Karchin et al [2] showed that detailed alphabets have greater potential for fold recognition and that the best results can be achieved by combining several alphabets. Developing new, informative alphabets of structural motifs, and efficient methods to predict them from the primary sequence is thus of great interest. These alphabets may help to improve the performances of current algorithms for protein folding, and for ab initio protein structure prediction. We present a new system, Porter+, for the prediction of a new alphabet of local structural motifs. The motifs are built by applying multidimensional scaling (MDS) and clustering to pair-wise angular distances for multiple phi-psi angle values collected from high-resolution protein structures [1].

Dataset:

Porter+ is trained on a data set extracted from the December 2003 25% pdb select list [3]. We use the DSSP program [4] to assign secondary structure and phi and psi angles, and remove sequences for which DSSP does not produce an output due to missing entries or format errors. After processing by DSSP, the set contains 2171 proteins and 344,653 amino acids. For our experiments we split the data into a training set containing 1736 sequences (S1736) and a test set of 435 (S435). Alignments for both sets of proteins were generated by three runs of PSI-BLAST [5].

Clustering:

To cluster sequences of phi and psi angles in the data sets, we followed the scheme devised by Sims et al [1]. In their study, protein conformational space from two to seven residue lengths was mapped into a three-dimensional space employing multidimensional scaling (MDS). Their analysis identified 14 clusters for the case of tetra-peptides, a very significantly smaller number of clusters than suggested by theoretical predictions. Here we adopted the centroids of these clusters as structural motifs and mapped each tetra-peptide in the S1736

and S435 datasets into the motif corresponding to the cluster i that minimises the same distance metric adopted in [1].

Predictive Algorithm and Implementation:

We model the prediction of a residue's closest structural motif as a classification task with multiple classes. To learn the mapping between inputs and outputs (sequence to structural motif) we use an architecture composed of Bidirectional Recurrent Neural Networks (BRNN)[6, 7] of the same length N as the amino acid sequence. In the tests presented in this work the input associated with the j-th residue ij contains amino acid information obtained from multiple sequence alignments of the protein sequence to its homologues, to leverage evolutionary information. Amino acids are coded as letters out of an alphabet of 25 (as in [8]). Beside the 20 standard amino acids, B (aspartic acid or asparagine), U (selenocysteine), X (unknown), Z (glutamic acid or glutamine) and . (gap) are considered. The input presented to the networks is the frequency of each of the 24 non-gap symbols, plus the overall frequency of gaps in each column of the alignment. This input coding scheme is richer than simple 20-letter schemes and has proven effective in [8].

Results and Discussion:

On the S435 set (test set), our systems for structural motif prediction for tetra-peptides into 14 clusters achieve 59.8% correct classification, with the true class being ranked within the 3 most probable for well over 80% of residues. The predictor is also significantly more confident on the correct predictions than on the incorrect ones. This suggests that the full output of the predictors (i.e. the 14 estimated probabilities of the structural motifs for each residue) contains substantially more information than the simple identity of the class ranked first.

Conclusions:

Our structural motif predictor may feed into a number of other stages of ab initio protein structure prediction systems in at least three ways:
- Predictions may be used as an additional input to predictors of other structural features, e.g. solvent accessibility, residue contact maps, etc.
This may lead to improved prediction of these features, and ultimately translate into improved structure predictions.
- They may be directly adopted, in combination with other features, to guide the ab initio reconstruction of protein backbones.
- They may help to select and rank "decoys" obtained by any ab initio structure prediction method.

Availability:

All the predictors are publicly available at the address http://distill.ucd.ie/.

References:
1. Sims GE, Choi I, Kim S: Protein conformational space in higher order phi-psi maps. PNAS 2005, 18:618–621.

2. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K:Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 003, 51:504–514.
3. Hobohm U, Schard M, Schneider R, Sander C: Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Science 1992, 1:409–417.
4. Kabsch W, Sander C: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983, 22:2577–2637.
5. Altschul SF, Madden TL, Schaffer AA: Gapped blast and psi-blast: a new generation of protein database search programs. Nucl Acids Res 1997, 25:3389–3402.
6. Baldi P, Brunak S, Frasconi P, Pollastri G, Soda G: Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999, 15:937–946.
7. Baldi P, Pollastri G: The Principled Design of Large- Scale Recursive Neural Network Architectures – DAG-RNNs and the Protein Structure Prediction Problem. Journal of Machine Learning Research 2003, 4(Sep):575–602.
8. Pollastri G, McLysaght A: Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2005, 21(8):1719–20.
9. Frasconi P, Gori M, Sperduti A: A General Framework for Adaptive Processing of Data Structures. IEEE Trans. on Neural Networks 1998, 9:768–86.
10. Richardson C, Barlow D: The bottom line for prediction of residue solvent accessibility. Protein Engineering 1999, 12(12):1051–4.