Review of “Protein secondary structure prediction using deep convolutional neural fields”

Abubakar Abid and Alex Barron

Summary

A protein’s precise structure and shape greatly affects its interaction with other molecules and is thus incredibly important in the design of drugs or enzymes. The ability to predict this structure purely from amino acid sequence could allow the virtual simulation and screening of potential synthetic molecules purely from their amino acid composition, vastly accelerating the design process.

This paper proposes “DeepCNF”, a new architecture combining a deep convolutional neural network and conditional random field to address this problem. It is able to predict secondary structural features (local folded polypeptide regions) in proteins to a new state of the art accuracy.

Context and Related Literature

The classification task is formulated as a sequence to sequence problem where a sequence of input features describing each amino acid in the protein chain is mapped to an output sequence of secondary structural features which are classified into either 3 or 8 states representing different shapes.

Interestingly, DeepCNF follows in a grand tradition of applying neural networks to protein structure prediction – the first successful approaches date back to the late 1980s (Holley et Al) and by 1993 a 2-layer feed forward network had become the first algorithm to demonstrate greater than 70% accuracy on the 3-state problem (Rost et Al). By 1999, a technique known as PSIPRED had achieved 80% accuracy classification using PSI-BLAST sequence profiles as improved input features and another simple 2 stage network (Jones).

Prior to the release of DeepCNF, PSIPRED had remained the state of the art for 3 state prediction for over 15 years. Some more modern deep learning methods had seen increased success on the 8-state problem through the use of deep belief networks and supervised generative stochastic methods, but neither were able to better 80% accuracy on the 3 state problem. Template based methods which exploit solved protein structures have been able to better the 80% classification accuracy, but these are dismissed by the authors as templates are not known for many proteins. Some more insight into this discarding would have been welcomed.

Methodology

The DeepCNF architectures contains two distinct modules – a deep convolutional neural network (CNN) to extract features from the amino acid inputs and a conditional random field (CRF) to model interactions between input features and labelled structures and among the structures themselves.

The input to the CNN are position specific scoring matrices (as in PSIPRED) generated from PSI-Blast runs. For the 1D CNN, 5 layers of window size 11 and 100 filters are used by default. While the authors seem to have fairly thoroughly explored varying both the number of layers and filters, the window size is always taken to be 11 due to the biological prior that helixes, the largest structural features, tend to be this length. Given that the approach still relies heavily on the PSI-Blast initialization (the authors report no improvement to previous methods without it), it might have been interesting to try increasing the window size so that some structure to structure inferences could be made during the input stage. Then perhaps the reliance on pre-computed sequence profiles could be dropped, allowing the proteins with less complete PSI-Blast entries which DeepCNF currently struggles on to be more accurately classified. The authors might also consider employing the ReLu activation function in the CNN over the current sigmoid/tanh implementation as this has been shown to produce superior performance on the majority of CNNs in any context.

The output of the neural network is fed to the CRF as in the original conditional neural field DeepCNF extends. The CRF uses two potential functions to model dependencies amongst the structural labels and CNN outputs:

$$\psi(Y,X,i) = \sum_{a.b} T_{a,b}\delta(Y_i = a)\delta(Y_{i+1} = b)$$

$$\phi(Y,X,i) = \sum_{a}\sum_{m} U_{a,m}H_m(X,i,W)\,\delta(Y_i = a)$$

Where T, U and W are model parameters to be trained.

The downside to this CRF implementation is that ψ only explicitly models relationships between adjacent structural labels. Using some form of RNN as the output layer would likely provide better modelling of longer distance label to label relationships at the cost of increased computational complexity.

The model is trained end to end on the standard CRF log likelihood with l2 regularization. There is no mention of the use of dropout in the network, but this has been shown to reduce generalization error in many deep neural networks and is likely worth trying.

Experiments

Experiments were performed for both the 3-state and 8-state classification problem on a variety of datasets. DeepCNF performed optimally among PSIPRED and more recent methods for all datasets when excluding template information. DeepCNF shows particularly impressive performance relative to PSIPRED and other methods when training and test proteins showed dissimilar sequence profiles. This suggests the deep architecture is significantly improving generalization over previous methods.

Possible Extensions in Future Work

Given the sequence to sequence nature of the protein structure problem, it seems more natural to replace the CRF used in this architecture with some form of RNN which can better model long distance dependencies in the output sequence. In fact, a recent paper (Li et Al) was able to combine a similar input CNN to that of DeepCNF with a bi-directional GRU RNN to achieve a new state of the art accuracy on the protein structure task. In this architecture, we simply feed the output of CNN into an RNN with a hidden state for each output structure label and pass the RNN outputs through a standard fully connected layer to obtain the final structure labels.

These CNN/RNN hybrid models were originally developed for the combination of natural language and image processing so it would be natural to try some of the NLP field's more recent architectures in the protein structure prediction domain. In particular, various forms of attention-based memory networks (Sukhbaatar et Al, Xiong et Al) have proven incredibly effective in the NLP space and have the potential to further improve the state of the art in this problem. Conceptually, having an attention memory module could allow the model to more directly learn which more distant structure positions are likely to be important in predicting the current output label and focus its attention on those.

Future work might also seek inspiration from the rapid improvement of genomic sequence CNNs. Using a more powerful architecture could remove reliance on the PSI-Blast input features allowing increased performance on a wider range of proteins.