Critical Review 2: Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields

Frank Cipollone, Dave Deriso, Gabriel Maher, Benjamin Nosarzewski,
Nikhil Parthasarathy, and Kushal Ranjan

November 2, 2016

1 Introduction

1.1 Problem Statement

Mapping sequences to protein structure is one of the most difficult problems in computational biology. Previous methods have suffered from dependence on the availability of previously solved structures as templates or relatively shallow architectures that can not capture the complex sequence-structure relationship. The authors present an architecture called DeepCNF that combines the advantages of both conditional neural fields and deep convolutional neural networks. This new model captures not only complex sequence-structure relationship, but also models protein secondary structure label correlation among adjacent residues. The authors show that their method improves upon the state-of-the-art methods, especially on those structure types which are more challenging to predict, such as high curvature regions (S), beta loop (T), and irregular loop (L).

1.2 Previous Work

We found that this paper did a good job of presenting a comprehensive view of the previous work on the topic. The authors described the existing methods, and motivated the changes they wanted to make in order to improve the current state of the art. Emphasis is placed on the variety of mathematical models that have already been tried, but neural networks have been used most effectively. Specifically, a 2-staged neural network called PSIPRED[4] is noted in the paper as achieving an accuracy of approximately 80% for 3-state secondary structure prediction.

Another important method was created by Baldi et al[3]. that was template based. It used solved structures as templates when it ran into similar structures in making predictions. This method actually performed much better than the state of the art, but the authors note that the method performed slightly worse when it predicted on data for which it did not have close templates.

The paper states that although many models have been tried, the 80% value for 3-state secondary structure prediction has not been broken. The authors attribute this plateau to the complexity of the relationship between amino acids and the corresponding secondary structure, and use this idea to motivate their attempted use of a deeper network.

2 Method

The architecture of the deep neural network is designed to capture the complicated interdependencies among input sequences as well as the output labels by combining a deep convolutional neural network (DCNN) with a conditional random field (CRF). The first few layers of the network which make up the convolutional deep neural net are capable of capturing important patterns in the input layer sequences. The authors investigated how the model performed for different depths of the DCNN (up to 7 layers each with 85 different neurons) and found that deeper networks performed better. However, we wonder why the authors did not reduce the dimension of subsequent layers or perform max pooling operations

as is typically done in DCNN, choices which are typically used to prevent over-fitting. Perhaps this would not be easy to implement since the network would need to expand out again in some way before the top layer to complete the condtional neural field (CNF) which has the same number of output labels as input labels. For training, the authors used L2 regularization but did not perform drop out, which could have been useful. Perhaps the L2 regularization and the L-BFGS training method are sufficient to prevent over-fitting. We are also curious if including a recurrent neural network between the DCNN and the CRF could help to capture interdependencies in the input sequence.

Typically, conditional random field (CRF) models use a linear potential function to represent the relationship between input features and output [1]. By adding middle layers between input and output (the DCNN in this case), the model becomes a CNF which is capable of capturing nonlinear relationships between inputs and outputs [1]. The paper assumes knowledge of CNFs without explaining basics such as how label prediction is performed for a new input by applying the Viterbi algorithm [2].

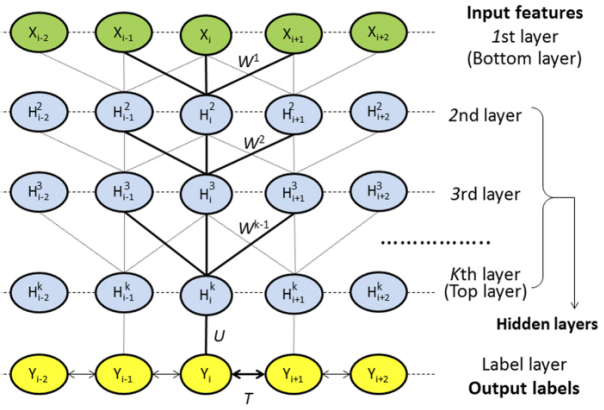

Figure 1: All layers from the first to the top layer form the deep convolutional neural network (DCNN). The top layer and the label layer from a conditional random field (CRF).

3 Results

The authors showed that DeepCNF outperformed all the other predictors. The authors suggest that this improvement comes from the fact that deep convolutional neural networks are better in predicting beta turn (T), high curvature loop (S), and irregular loop (L) states, which appear more often at the boundary of a helix or sheet segment. DeepCNF greatly outperforms other predictors when dealing with protein sequences with more than two effective amino acids across all the residues. The authors take care to ensure that the improvement from their method is not due to similarity between training and test sets by training on a variety of data sets with proteins which belong to different superfamilies.

4 Conclusion and Future Work

The performance of a classifier can be evaluated using sensitivity, which is simply defined as the percentage of correct classifications among a set of possible options. Q3 is the most common sensitivity metric for secondary structure classifiers, and gives the percentage of times an amino acid is classified into one of three states: α-helix (H), β-strand (E), and coil (C) [**?**]. Q8 has eight possible classes: 3-10 helix (G), α-helix (H), π-helix (I), β-stand (E), bridge (B), turn (T), bend (S), and others (C) [**?**]. The 8-state secondary structures convey more precise structural information than 3-state, which is particularly important for a variety of applications [**?**]. Finally, the Segment Overlap Score (SOV) measures sensitivity for segments of continuous structure types instead of sensitivity for individual residues, and also provides a free parameter that sets an allowance for errors near the boundaries of segments [**?**, **?**].

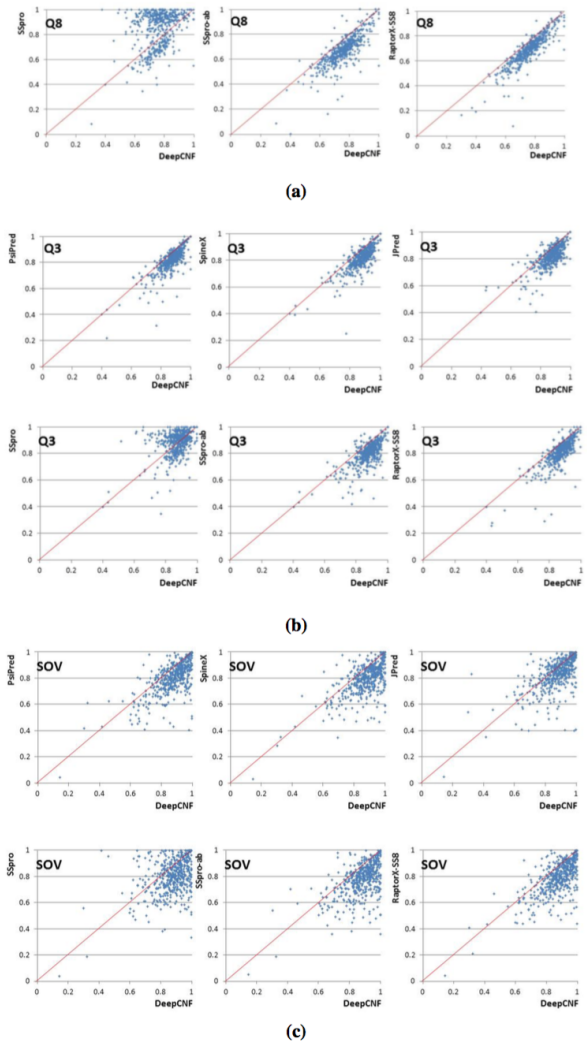

Figure 2: Head-to-head comparison of DeepCNF and three methods: SSpro (with template), SSpro-ab (without template), and RaptorX-SS8. Figure taken from Supplemental Materials Figure S1.

The authors compared the performance of their classifier to that of previous studies using the sensitivity metrics Q3, Q8, and SOV. However, sensitivity only considers the true positive rate, but fails to take into account the number of true negatives, which is measured by specificity. In other words, the fact that one classifier may have a higher proportion of true positives overall does not indicate relative improvements in pairwise class discriminative ability. In this respect, we feel that the authors could have reported a confusion matrix and accuracy instead of just sensitivity. An ROC curve would be ideal, but since there the deep net architecture does not provide a discrimination threshold, it would not be possible to produce an ROC curve in this case.

It's interesting to note that when the authors used SOV as a performance metric, their model outperformed the three other methods more so than when using S3 and S8. One possible reason is that the conditional random fields in the method models the inter-dependency among adjacent residues in a SS segment, which helps reduce erroneous predictions in the middle region of a segment. Our interpretation of this is that the inter-dependencies in the model smooth out sharp variations between predicted structural states thereby producing a prediction that is based on more on a holistic representation of the segment than a prediction based on individual elements of a segment. Clearly, this has better performance in the context of SOV, which again judges based on segment-wise sensitivity.

We feel that this paper has provided significant evidence that convolutional networks for prediction of

secondary structures is a fertile area of inquiry. Future work could be directed at the interaction of local and distal segments, which may be appropriately addressed by such methods as LSTMs, recurrence, and variational kernals. We are already seeing signs of this direction. In fact, since the time this paper was published, one major related paper has been published. Li et al[5] have attempted a new method using Cascaded Convolution and Recurrent neural networks. This paper updated the architecture in two major ways. First, it used convolutional layers with different kernel sizes, looking for local contextual features spanning multiple neurons. Second, the method used an RNN in order to capture interdepencies between amino acids with a high degree of separation. The results reported by this new paper are interesting. The authors of the new paper claim to outperform the Wang et al method in Q8 accuracy by 1.1%, and in Q3 accuracy by 1.7%, using the same training and testing data. This provides support for further studies into architectural optimization that may more faithfully model the complex inter-dependencies within polypeptide chains and their secondary structures.

References

[1] J. Peng, L. Bo, and J. Xu, Conditional Neural Fields, https://research.cs.washington.edu/istc/lfb/paper/nips09b.pdf

[2] C. Sutton and A. McCallum, An Introduction to Conditional Random Fields, http://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf

@articlespencer2015deep, title=A deep learning network approach to ab initio protein secondary structure prediction, author=Spencer, Matt and Eickholt, Jesse and Cheng, Jianlin, journal=IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), volume=12, number=1, pages=103–112, year=2015, publisher=IEEE Computer Society Press

@inproceedingsyaseen2013template, title=Template-based prediction of protein 8-state secondary structures, author=Yaseen, Ashraf and Li, Yaohang, booktitle=Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on, pages=1–2, year=2013, organization=IEEE

[3] Magnan, C. N. & Baldi, P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity, https://www.ncbi.nlm.nih.gov/pubmed/24860169

[4] Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices, https://www.ncbi.nlm.nih.gov/pubmed/10493868

[5] Zhen Li, Yizhou Yu, Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks, https://arxiv.org/abs/1604.07176

@articlerost1994redefining, title=Redefining the goals of protein secondary structure prediction, author=Rost, Burkhard and Sander, Chris and Schneider, Reinhard, journal=Journal of molecular biology, volume=235, number=1, pages=13–26, year=1994, publisher=Elsevier