CS273B Paper Review

DanQ: A Hybrid Convolutional and Recurrent Deep Neural Network for Quantifying the Function of DNA Sequences

Wen Torng, Junhong Choi, Joy Xiang,
Abhimanyu Banerjee, Raunaq Rewari

Background

Noncoding DNA constitutes over 98% of the human genome and plays crucial roles in regulating gene expression. The majority of single-nucleotide polymorphisms (SNPs) associated with disease and other traits are known to reside in the noncoding region. Despite of this knowledge, how these genetic variants conduct their functional effects still remains poorly understood. Because traditionally defined local sequence motifs have difficulties capturing complicated factors such as cofactor binding sequences, chromatin accessibility and structural flexibility of the DNA-binding site, which often plays crucial roles in transcription factor (TF) binding, it is challenging to predict the functional effect of genetic variants de novo from sequence. Deep learning algorithms offer a promising solution to the problem due to their ability to automatically extract hierarchical features and complex interactions from large scale dataset. Zhou et al. (2015) recently proposed a deep convolutional neural network (CNN) based framework, DeepSEA to predict effects of noncoding variants from sequence alone and achieve state-of-the-art performance. In this paper, Quang et al. propose DanQ, a hybrid convolutional and recurrent deep neural network, attempting to further improve the performance.

Summary

Quang et al. hypothesize that adding bi-directional long short-term memory recurrent neural networks (LSTM-RNNs) on top of the CNN framework can capture temporal dependency of the local sequence motifs, and therefore can achieve better performance than employing CNNs alone. To test the hypothesis, the authors use the same dataset and defined pipeline as in DeepSEA and perform head-to-head comparison on the performances for two different tasks: (1) Predicting chromatic profiles, including TF binding, DNAse I sensitivity, and histone-mark profiles from DNA sequences (2) Prioritizing functional SNP using chromatic features learned from the previous task. They report limited improvement over DeepSEA for predicting chromatic profiles using the ROC-AUC metric. However, the authors argue that PR-AUC is a better metric for evaluating the performance of the highly imbalanced dataset, and show better improvement using the PR-AUC metric. Despite the improvement shown in PR-AUC curves, this result does not immediately translate to an improvement in prioritizing functional SNP. For this task, very limited improvement over DeepSEA is observed. To gain understanding of the network, the authors convert convolutional filters into sequence motifs and show that the learned motifs match well with known motifs.

Critiques

Although the proposed CNN-RNN architecture is not new to the deep learning field, the authors contribute to the problem by introducing the architecture to DNA sequence data for the first time. However, the improvement of performance is rather limited and evidence on how RNNs contribute to the overall predictive performance is not well illustrated. Further insights can be gained by visualizing the temporal dependency captured by the RNNs and further improvement

of performance could be done by improving the design of network architecture, dataset construction, and weight initialization strategy, as will be discussed in the following paragraphs.

I. CNN stage:
For the convolutional stage, the authors propose a single layer CNN, single max pooling stage architecture instead of a three layer CNN with two max pooling stages architecture employed in DeepSEA. It is unclear if the single CNN layer is sufficient to capture the critical local sequence features or if more layers are required. As a separate point, the authors state that training a larger model containing 1024 convolutional kernels and initializing half of the filters with known motifs from JASPAR further improve the performance of DanQ. However, they did not independently verify the effect of initializing the weight with PWM and the effect of increasing the number of filters from 320 to 1024. The benefits of initializing the weights with PWM is unclear.

II. Fixed length LSTM-RNNs:
The authors integrate information from the output of LSTM by first concatenating the forward and backward LSTM outputs and feeding the output vector of each time step individually to the dense layer. This limits the LSTMs to a fixed length architecture since the number of temporal steps has to be known beforehand using this strategy. As a results, the so-called “long range” interaction is only defined in and limited to regions of 1000 bps long. This can pose significant limitations on capturing the interactions between motifs since some genomic elements, such as enhancers, can reside up to million bps away from the TF binding site, and yet 3D architecture of genome might allow close physical interaction between two. Since the architecture design of proposed neural network will not capture such effect, coupling RNN over CNN only benefits the subset of motif interactions that happens within 1000 bps around the TF binding site. Alternative to connecting the output vector of every time step to the dense layer as shown in this paper, it is common to take the output of the last time step as the representative vector or to average the output vector across all time steps to enable arbitrary length sequence input. This approach will allow better modeling of long range interactions given longer sequences.

Furthermore, by connecting the output of all time steps to the dense layer instead of using a representative output, more trainable parameters are introduced, which can increase the computational expense and chances of overfitting. This design can also increase the sensitivity of the network to the exact location of the motif present in the 1000 bps region since the output of each time step have a distinct set of weights connect to the neurons in the next layer. (i.e. The presence of the motif at the 2^{nd} and 3^{rd} temporal step might result in different dense layer activation values from having the exact motifs at the 3^{rd} and 4^{th} temporal steps). Max- or average pooling across output vectors of N time steps can be a potential strategy to alleviate the sensitivity and reduce the number of parameters (if not considering averaging across all output vectors or taking the output from the last time step).

This design choice they make might be based on the fixed length dataset they use. To enable direct comparison, they use the same dataset as constructed in the DeepSEA paper. However, DeepSEA is a CNN-based framework that is not designed to deal with varying length of

sequences and to model long range temporal dependency. Therefore, the constructed dataset might not be the most suitable choice to test the full potential of an RNN architecture. Besides the original dataset, the authors could have constructed a different dataset that includes much longer regions and train a network that allows input of varying length to see if RNN architectures capture additional information from longer sequences.

Further questions to consider include if it is a good strategy to average the scores of template and complementary strands for the prediction or if it is better to do a max operation over the two strand for the purpose of detecting strand-specific binding. Finally, the ultimate question to ask would be can RNNs really tackle long range interactions present in real biology? Although LSTM units can alleviate the vanishing gradient problems, there are still practical limitations on the length of the sequence in which temporal dependency can be modeled. Interactions resulting from the 3D remodeling of chromatin architecture aren't very likely to be captured by currently available RNN architectures, and therefore, another strategy to resolve such problem is needed to fully model de novo long range interactions from sequences alone.

III. Analysis: Network importance scores

The main claim of this paper is the benefits of adding RNNs into the architecture. However, they only visualize the features captured by CNNs and do not perform analysis to show what temporal dependency were captured using the RNN framework. A possible way to analyze this is to use the temporal importance score proposed by Lanchantin et al. (2016.) Lanchantin et al. propose a method to visualize the importance of the ith time-step by feeding in subsequences of the input sequence starting from the ith position till the end of the sequence and recording the corresponding output class score. This approach can allow the authors to observe transitions of scores from negative to positive or vice versa with respect to sequence positions. However, the approach might be subject to thresholding and saturation issues as discussed in class. The DeepLIFT method proposed by Avanti et al. (2016) might be a more robust way to obtain importance scores.

Conclusion

To conclude, the authors introduce a CNN-LSTM architecture to DNA sequence data for the first time and showed slight improvement of performance for predicting chromatic profiles and prioritizing functional SNP over a CNN-based framework, DeepSEA. We propose that various network architecture and dataset design can be improved to fully test the potential of the CNN-LSTM-RNNs framework and also suggest different visualization technique to improve their understanding of what temporal features the recurrent neural networks are learning. Finally, we suggest that additional network architectures that can capture the physicality of human genome might be needed in order to fully model physiological long range interaction among genomic elements, and to really learn the regulatory 'grammar' among them.