UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

Applying Saliency Map Analysis to CNNs on Protein Secondary Structure Prediction

by Guillermo Romero Moreno

The relatively new field of deep learning is slowly being transferred to the biology field and pushing its state-of-the-art achievements. However, improvements in performance come with the drawback of opaqueness, as what deep learning machines learn cannot be fully understood. Although a few authors have already started applying deep network interpretability techniques on biological problems to overcome this issue, none of them has been applied yet to the problem of protein secondary-structure prediction.

The aim of this work is to develop interpretability techniques for state-of-the-art deep networks that have been trained to solve the secondary-structure prediction problem. For doing so, a way to apply and aggregate saliency maps has been construed and applied to a near-state-of-the-art convolutional network, showing some further insights of the relationship between the inputs and the outputs. These results could be of double value: on one side, it may help biologists to get a better understanding on the underlying structural protein processes; on the other, machine learning researchers can understand better their machines and spot their flaws more easily.

Chapter 1

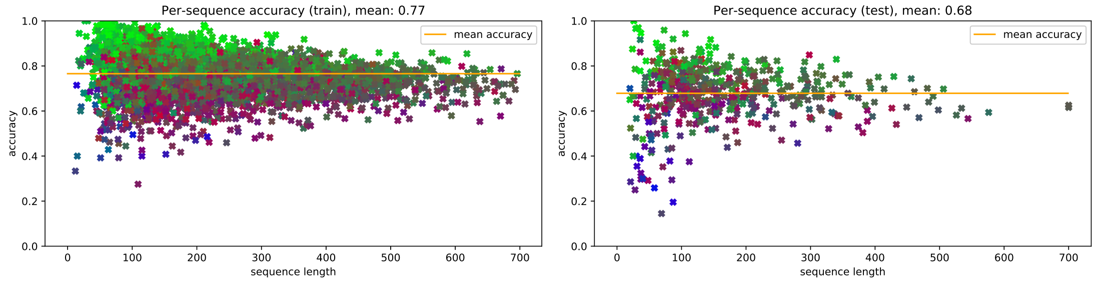
Results & Discussion

I have trained the network described in section **??** for 400 epochs with the same parameters as the ones used by Jurtz et al. (2017); i.e., gradient clipping at 20, regularization term $\lambda = 10^{-3}$, and training-validation split at the 5278th sequence. The resulting network reaches an accuracy of 67.7% on the test set, which is not far from the 71% of the state-of-the-art (see section **??**). I believe that the techniques of analysis here presented and the conclusions withdrawn from them can be transferred to current state-of-the-art methods without losing validity.

1.1 Outlier analysis

In order to analyse the performance space a bit better, the average accuracy per sequence has been calculated and it has been plotted in Figure 1.1 with respect to the sequence length. The distribution exhibits the typical funnel shape that one could expect from processes with random variables forming groups of different sizes: the bigger the groups, the smaller the variance. The funnel ceases to shrink at length about 400, so it would be particularly interesting to understand why the network is classifying worse (60% and below) some of the sequences above that length.

If we observe the colour scheme of the figure, we can understand right away that sequences rich in α-helix are generally better predicted than β-sheets and coils. An explanation could be that while α-helix sizes are up to CHECK NUMBER, which is inside the window size, β-sheets interact with amino-aids further away in the sequence, which is not possible to be captured with the window of the network, of lateral size of 9.



Per-sequence accuracy (train), mean: 0.77

Per-sequence accuracy (test), mean: 0.68

FIGURE 1.1: *The mean accuracy per sequence by sequence length.* The 5504 sequences of the training set are shown on the left and the 514 of the test set on the right. Each point represents a single protein, and its colour corresponds to the amount of β-sheets (red), α-helices (green), and coils (blue) it has. A purple point, for instance, would predominantly have β-sheets and coils.

1.2 Feature visualization

1.2.1 First layer filters

Saliency maps on layers?

1.3 Saliency maps on inputs

Before going through the analysis, it is worth commenting that the saliency map outputs have many dimensions, since each position at each sequence has a saliency map with shape 8x42x19, corresponding to the 8 classes (outputs), the 42 inputs and and the total window size of 19. The results can be shown in multiple ways, depending on which dimensions are preserved and which are aggregated.

Analysis on amino-acids and *pssm*

When looking at typical secondary-structure prediction algorithm, there is one point that may raise some suspicion: the inclusion of half of the inputs as one-hot encoded (amino-acids) and the other half as dense vectors (*pssm*). One could thing that this discrepancy may strongly favour the information coming from the dense part, since the weights associated to it will learn much faster in a typical gradient descent learning schema.

Saliency maps can be used to prove whether this hypothesis is right by inspecting which of the input groups is being most decisive in the classification process. For doing so, each saliency map is divided into two groups of $8x21x19$, and all the values inside each group are added up to a single **saliency score**. Thus, each position of each sequence will have two scores, one for the amino-acids and one for the *pssm*, and its comparison will

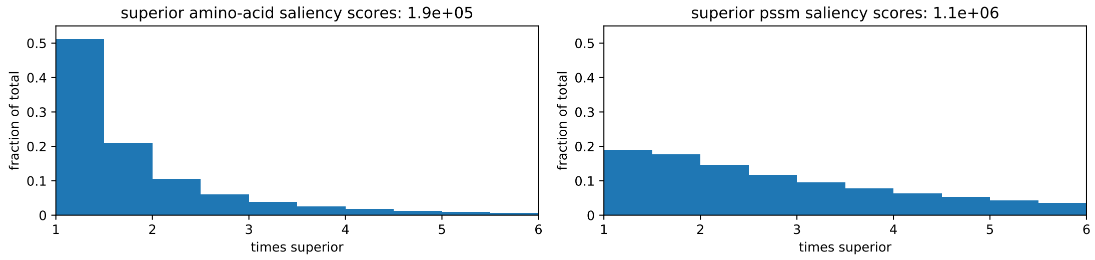


FIGURE 1.2: *Amino-acids versus pssm comparison.* On the left, a histogram of the positions whose amino-acid saliency score was higher than the pssm score, with the total amount in the title. The bins represent the superiority ratio, i.e. how many times bigger was the score. On the right, the analogous histogram for pssm superior scores.

give us which part of the inputs is more decisive. The results are shown in Figure 1.2, from where it can be appreciated that there are considerably more positions in which the input from *pssm* was more determinant (around 1.1 millions) than the ones with more determinant amino-acid inputs (about 200 thousands). Furthermore, in most of the cases amino-acid inputs had more impact it was only by a narrow margin, with 70% of the cases having only up to two times bigger saliency scores. On the contrary, more than 60% of the positions with higher *pssm* saliency score did this with a more than twice higher score.

1.3.1 Sequence-specific saliency maps

This section will present the sort of analysis that can be done for a specific sequence. This can be specially useful for analysing the sequences whose accuracy was remarkably lower (outliers). Each sequence has a number of saliency maps equal to its length, l, and they can either be inspected one by one or be overlapped through the sequence, obtaining a single 8x42xl **sequence map**.

Figure 1.3 shows consecutive saliency maps belonging to a single class. The patterns they present are generally preserved (note the differences in scales), just being shifted by one position on the window axis. This suggests that the algorithm is not differentiating that much between specific amino-acid positions, but rather looking for them to be in the vicinity. For this reason, we can consider that overlapping them in a single sequence map does preserve most of the information, as it is shown in Figure 1.4. This sort of map reveals which amino-acids of the vicinity are mostly responsible for a prediction in a particular position. For instance, the predictions of class E on the left side of the shown fragment have to do with the presence of amino-acids A, V and L, while the failure to predict class H on position 109 is likely related to the presence of amino-acid D at position 110. Note that these saliency values come from the aggregation of all the 42 inputs and not only from the one-hot encoded side, which explains why positions 105 and 106 have different values in spite of having the same amino-acid (D). For a deeper

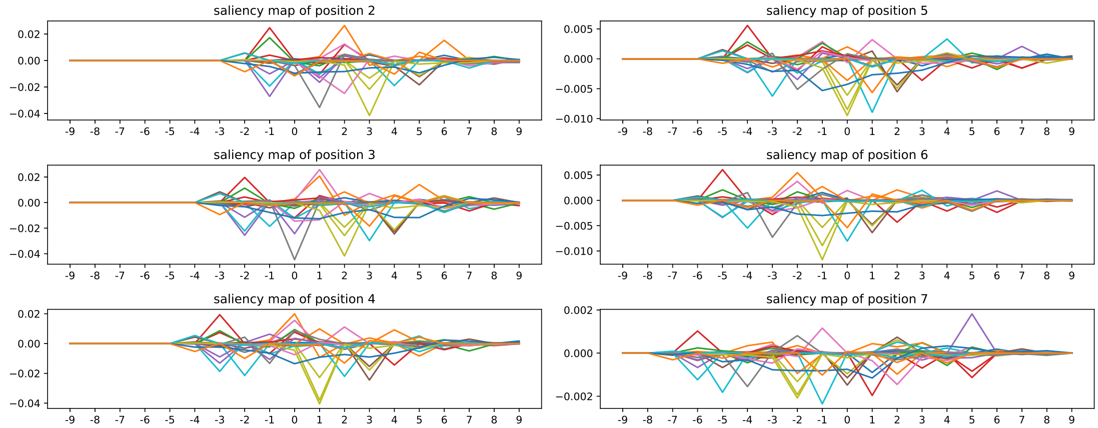


FIGURE 1.3: *Saliency maps for consecutive positions in a sequence.* In every sub-figure there are 42 lines of the 42 inputs, corresponding to their saliency values for class H.

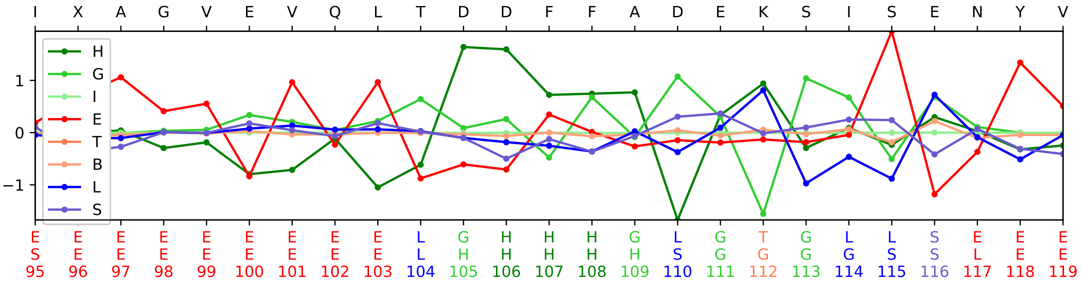


FIGURE 1.4: *Fragment of a sequence map aggregated by input.* Each line corresponds to the aggregated saliency values of one of the eight output classes. The upper x axis displays the amino-acids in the sequence. The labels at lower x axis contains three sets of labels in this order: the predictions, the true values, and the positions in the sequence. The code of colours of the labels is the same as the lines and is set according to the class of the predictions.

look into the whole *pssm* spectrum, we would need to narrow the scope down to a single class, as it is done in Figure 1.5. This figure reveals that is indeed common that the real amino-acid in the sequence is not among the most influential ones from the *pssm* for making the decision. Note that the pssm values of the amino-acids not always need to have contributions of the same sign (such as P or D in the figure), revealing that the network is capturing something more than pure presence: location and combinations of amino-acids are also important.

1.3.2 Class-specific saliency maps

Instead of looking at the saliency information along single sequences, aggregating the information from all the 1.3 millions saliency maps could give us some broader information of what the network has learned. Again, the high dimensionality of the data to explore requires us to seek for meaningful ways to aggregate and represent the data.

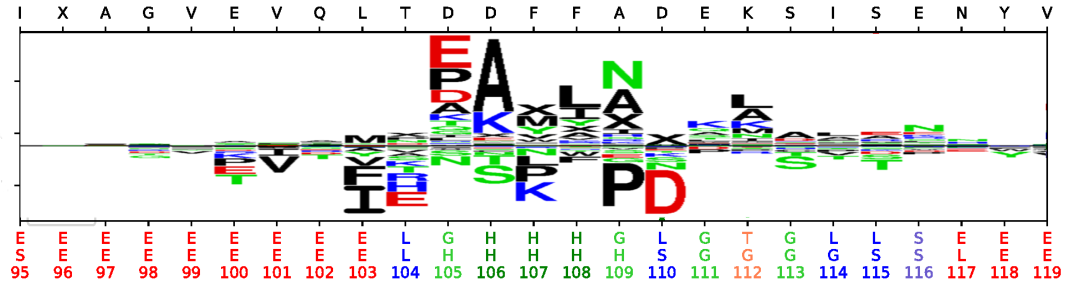


Figure 1.5: *Fragment of sequence map for class H.* The layout of the figure is the same as of Figure 1.4. The image has been generated by *SeqLogo* Thomsen and Nielsen (2012).

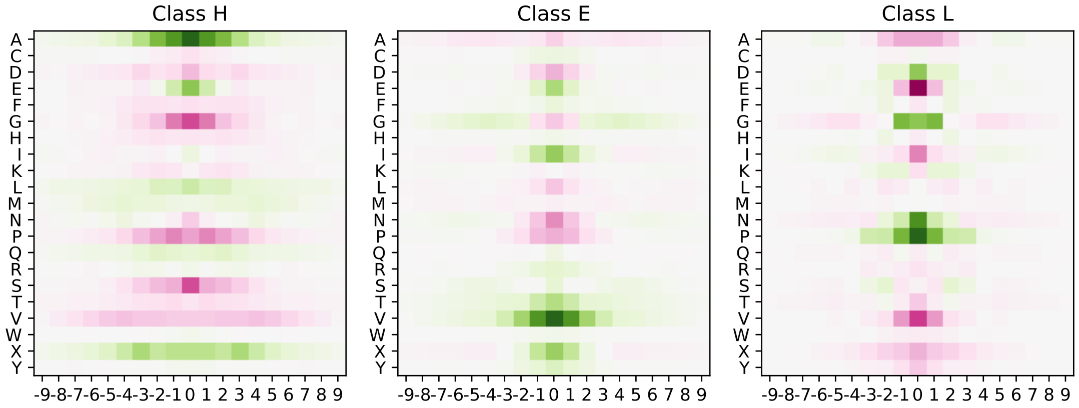


Figure 1.6: *Per-class aggregated saliency maps of the main three classes along the window.* Only the saliency of *pssm* value is shown here. Green means positive saliency values and purple means negative.

Sheer addition

A simple way to aggregate the saliency maps would be adding them up with element-wise addition, thus composing a single saliency map of size 8x42x19. This mega-saliency map should convey rough information of the general features of the network. Since there is no a sense of "left" or "right" in the sequences, the aggregation should be *side-invariant* and therefore an extra *mirroring* operation is applied to the aggregation in order to make the aggregated saliency map symmetric along the centre of the window.

To start with, Figure 1.6 shows the aggregated saliency maps of the three main classes. There are a few details to notice here. First, higher saliency values concentrate around the centre of the window, meaning that most relevant information is located around 5 positions from the predicted one. Second, most amino-acids preserve the same sign along the window of a class, although a few exceptions also occur. Third, as it could be expected, class L generally presents inverse patterns w.r.t. classes H and E. This makes sense since coils appear wherever there are no α-helices or β-strands, they are complementary. Lastly, each class has a "feature amino-acid" that is their best indicator, amino-acid A for class H, V for E, and P for L.

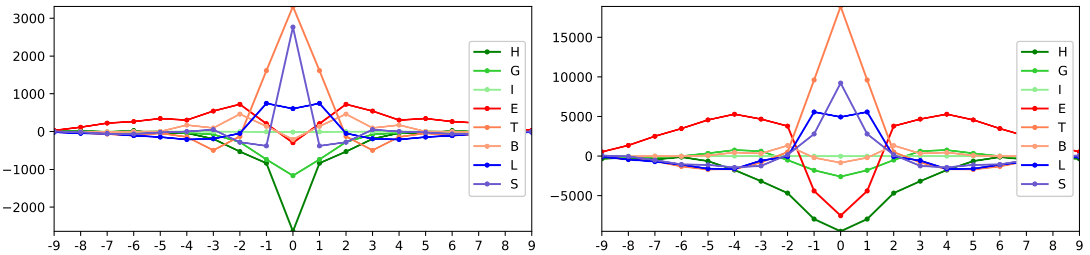


FIGURE 1.7: *Per-class aggregated saliency map of for amino-acid G along the window.* On the left the information from the *one-hot encoded* part, on the right the values corresponding to the *pssm* half of the input; plotted in different scales.

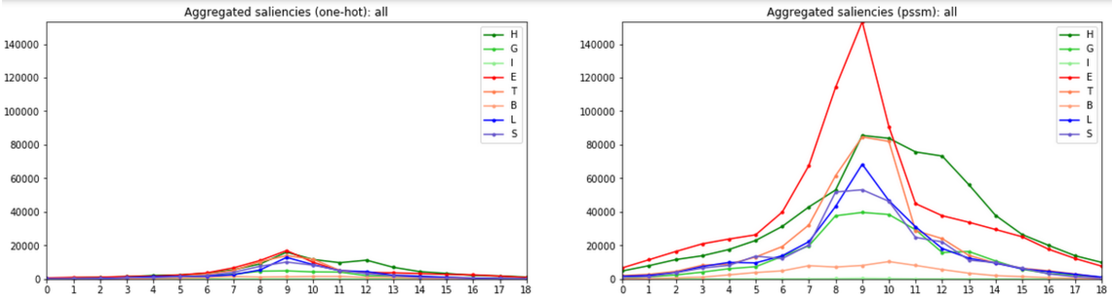


FIGURE 1.8:

Figure 1.7 shows the same information from the mega-saliency map but from a different perspective: it focuses on how a single amino-acid and shows how its position affects its influence on the the eight classes. As it has been discussed above, the *one-hot encoded* amino-acid half of the input appears to be less significant than the *pssm* half (about one fifth in this case). Both graphs show similar patterns, although it differs more for some other amino-acids (not shown). The figure shows that amino-acid G favours β-strands on its sides and *turns* and coils on its spot.

Per-aminoacid and class aggregations

Clustering techniques

aggregate all individual saliency maps

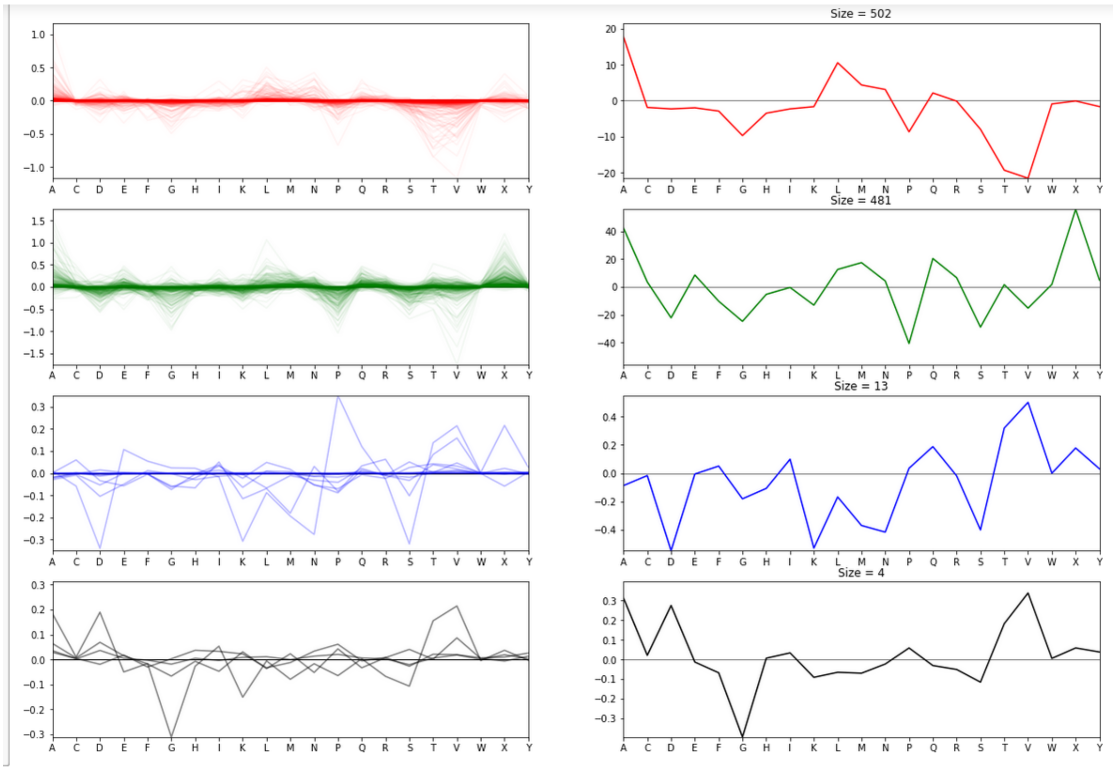


FIGURE 1.9:

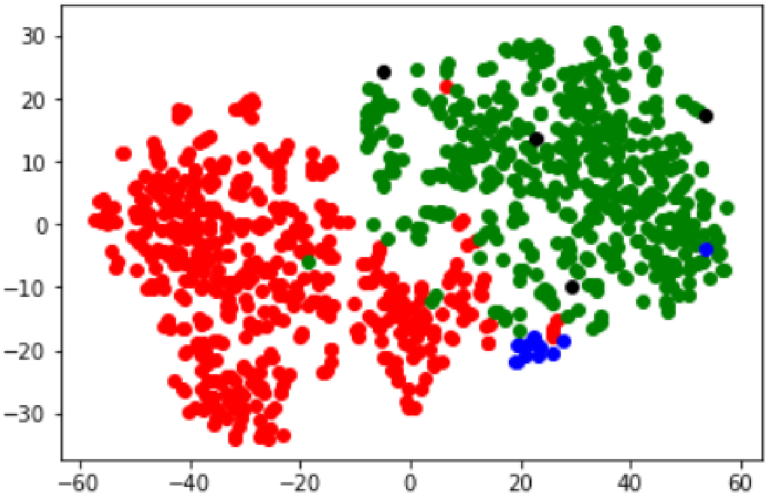


FIGURE 1.10:

Bibliography

Vanessa Isabell Jurtz, Alexander Rosenberg Johansen, Morten Nielsen, Jose Juan Almagro Armenteros, Henrik Nielsen, Casper Kaae Sønderby, Ole Winther, and Søren Kaae Sønderby. An introduction to deep learning on biological sequence data: Examples and solutions. *Bioinformatics*, 33(22):3685–3690, 2017. ISSN 14602059.

Martin Christen Frolund Thomsen and Morten Nielsen. Seq2Logo: A method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Research*, 40(W1), 2012. ISSN 03051048.