SALIENCY MAP ON CNNS FOR PROTEIN SECONDARY STRUCTURE PREDICTION

Guillermo Romero Moreno *Mahesan Niranjan* *Adam Prugel-Bennett*

School of Electronics and Computer Science
University of Southampton, Southampton, UK
Email: grm1g17@soton.ac.uk, m.niranjan@southampton.ac.uk, apb@ecs.soton.ac.uk

ABSTRACT

Deep learning, a powerful methodology for data-driven modelling, has been shown to be useful in tackling several problems in the biomedical domain. However, deep neural architectures lack interpretability of how predictions from them are made on any test input. While several approaches to "opening the black box" are being developed, their application to biological and medical data is very much as its infancy. Here, we consider the specific problem of protein secondary structure prediction using the techniques of saliency maps to explain decisions of a deep neural network. The analysis leads to two important observations: (a) one-hot-encoded amino-acids are irrelevant in the presence of PSSM values as extra features; and (b) in predicting α-helices at any position, amino-acids to the right are far more important than those to the left. The latter observation may have a biological basis relating to the synthesis of proteins by ribosome movement from left to right, sequentially adding amino-acids.

Index Terms— Interpretability, Saliency Maps, Protein Secondary Structure Prediction, Convolutional Neural Networks.

1. INTRODUCTION

Secondary structure prediction is a long-time studied problem in bioinformatics. The 3D structure of a protein determines the function it is going to adopt in the cell. However, the protein structure cannot be easily measured without being too costly, so computational tools can be an alternative by predictions based on amino-acid sequences —easy to obtain through DNA sequencing— and proteins with known structure. Since direct prediction of the 3D structure is still a hard problem, tackling the prediction of the secondary structure can be seen as an easier middle step. Protein secondary structure prediction is a sequence structural tagging problem: each element (amino-acid) of the protein sequence has to be assigned a class (secondary structure). There are 21 different types of amino-acids and eight possible goal classes, composed of 3 types of helices (H, G, I), two types of β-sheets (B, E), and three types of coils (T, S, L) [1]. Along with the amino-acid themselves, other relevant features can also be added as inputs, such as Position Specific Substitution Matrices (PSSMs) [2]. PSSMs encode the evolutionary probability of finding substitutions in each element of the amino-acid chain, and brought a significant performance improvement in the classficiation task. A common input sequence would have length l (variable from protein to protein) and width 42: 21 from one-hot encoded amino-acids and 21 from the PSSM values, normalized to a range between zero and one [3]. A new generation of deep learning approaches started recently with [4], who implemented a Generative Stochastic Network fed by a 1D Convolutional Neural Network (CNN) architecture. Later works already included 1D CNNs with five or more layers [5, 6, 7] or recurrent neural networks [8, 9], which are deep in the sense of signals being processed for many time-steps.

Saliency maps (also known as *attribution techniques* [10]) are a visualisation technique that aims to reveal which parts of an input sample are mainly responsible for the output decision made by a classification system. They can be regarded as a type of *sensitivity analysis*, applied in many other fields of research. A saliency map has the same dimensions as the input and contains their importance values, i.e. their contribution to the output. Depending on their calculation method, saliency maps can be broadly grouped into *perturbation-based approaches* (making modifications on the input and assessing changes in the output) or *backpropagation-based approaches* (obtaining the importance information from the gradient of the output respect to the input) [11]. Although the first group is intuitive and useful for small input spaces, it becomes quickly intractable when the input size grows, as all possible combinations of inputs should be examined for a complete analysis. The second group allows the computation of importance scores in a single gradient computation, with a reduced computational complexity that makes it preferable for bigger input spaces.

Back-propagation approaches can be thought as a linear approximation of the classification function around a sample input point x_0 by applying a first-order Taylor expansion, as introduced in [12]:

$$f(x) \approx w^T x + b \,, \qquad w = \left.\frac{\partial f}{\partial x}\right|_{x_0} .$$

In their simplest form, the saliency maps of back-propagation methods are equivalent to the gradient value on the input [12]. A second approach [13] would multiply the gradient by the input values to leverage out the gradients that don't carry relevant information. A last wave of methods proposes including a reference point and hence more closely resembling the Taylor approximation. The main examples of this trend are integrated gradients [14], deep Taylor decomposition [15] and DeepLIFT [11]. They overcome problems of previous methods such as saturation or discontinuities in the gradient, although they bring the extra difficulty of choosing an appropriate reference point.

2. PREVIOUS WORK

Perturbation-based approaches are prevalent, with perturbation as genetic mutations [16, 17], small sliding windows with random genetic code [18] or known motifs [19], among others. Gradient-based approaches have barely been translated to the biological field. Lanchantin et al. [20] include saliency maps with the form of gradient $\times$ input for TF binding site classification. They extracted the window with the highest score from each saliency map and compared them with a database of known motifs, matching almost half of the motifs thus produced. Shrikumar et al. [11] developed the reference-based saliency map technique DeepLIFT and simulated a motif detection task within a genomic sequence to prove its effectiveness. Finnegan and Song [21] utilised Markov chain Monte Carlo methods to withdraw samples from the maximum entropy distribution around a single sequence and assessed the importance scores by looking at the variance of the samples at each position. This method was applied to a previously trained DNA-protein binding CNN and proved to have better results than DeepLIFT.

All these methods address classification problems where there is a single output (classification task) for each sequence. A significant difference between this work and previous papers that make use of saliency maps is that they focus on many-to-one classification problems (one output class per input sequence/image), whereas our classification task is many-to-many (each position of the sequences is assigned a class), producing as many saliency maps as positions in a sequence. To the best of our knowledge, interpretability techniques have not been applied yet to this sort of problems.

3. METHODS

The experiments used the database produced and made public by Zhou and Troyanskaya [4]. It includes two sub-sets (training and test, with 5534 and 514 protein sequences of varying length, respectively) of proteins that come from different sources after removing the proteins that share 25% or more similarity, thus ensuring that the test set is composed of totally new samples. The proteins in the dataset already come in one-hot form, along with their Q8 class in one-hot as well and PSSM values. The dataset is heavily imbalanced, as it can be seen in Table 1

Q8 grouping		**Explanation**	**%**
α-helix	H	Helix with 4 turns	34.54
3_{10}-helix	G	Smaller helix with 3 turns	3.91
pi-helix	I	Bigger helix with 5 turns	0.02
β-bridge	B	Isolated β-bridge	1.03
β-strand	E	Participates in β-ladders	21.78
Turn	T	Turns smaller than a helix	11.28
Bend	S	Curved piece	8.26
Loop	L	Sometimes also as coil (C)	19.19

Table 1. Targets for the secondary structure prediction problem, as defined by Kabsch and Sander [1] in their Dictionary of Secondary Structure or Proteins (DSSP) and their presence on the training set.

The network architecture is composed of three successive convolutional neural networks and a dense layer on top. Each of the convolutional layers contains three sets of filters of size 3, 5 and 7, respectively, with 16 filters per size. There are skip connections at every convolutional layer. The dense layer has 200 neurons and is connected to the softmax output layer. The convolution operations are carried out with padding at each end of the sequence to preserve the length throughout the process. The total window size of the network is 19, meaning that for making a single secondary structure classification the network obtains information from 9 adjacent positions at each side. The network has been built and trained using the open-source code developed by Jurtz et al. [9].

Saliency maps are calculated by the conventional technique of computing the gradient of the output with respect to the inputs and multiplying it by the value of the input (gradient $\times$ input) [22]. Every single position in a sequence produces a saliency map that spans the width of the input vector of size 42 and 9 positions to each side, due to the architecture's window size of 19. Each output class has its independent saliency values, so the total size of a position saliency map is 8x42x19.

The presence of overlapping saliency maps allows for different ways in which to aggregate them to extract meaningful information. If we focus on a sequence of length l and want to obtain a single sequence-specific saliency map, we can add up the overlapping areas to form a saliency map of size 8x42xl. By changing the focus to a broader look on what the network has learnt, the addition of the saliency maps for the positions in all sequences could create a single saliency map of size 8x42x19 that shows an average behaviour of the network. From this map, we can extract general information about a particular class (creating a class-specific saliency map) or about a particular input (PSSM-specific saliency map).

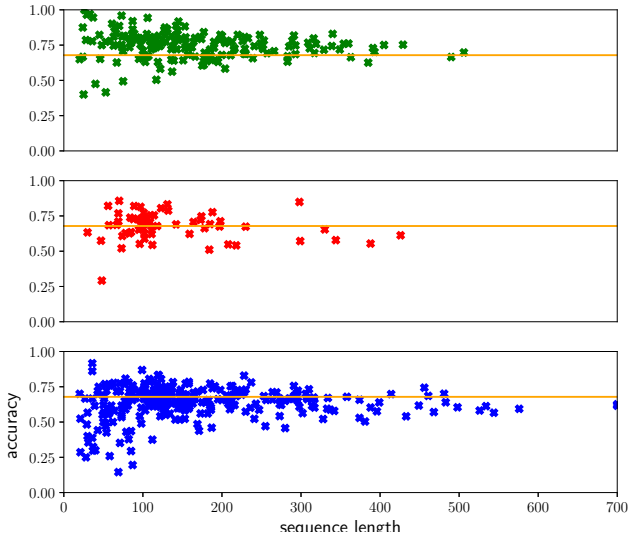

Fig. 1. Mean sequence accuracy against sequence length for the proteins in test set. Each point represents a single protein sequence and the horizontal line marks the total mean accuracy. On top, sequences with majority of helices (H, G, I); β-sheets (E, B) at the middle; and coils (L, S, T) below. Best accuracies are achieved with high percentage of helices and worst accuracies with with majority of coils.

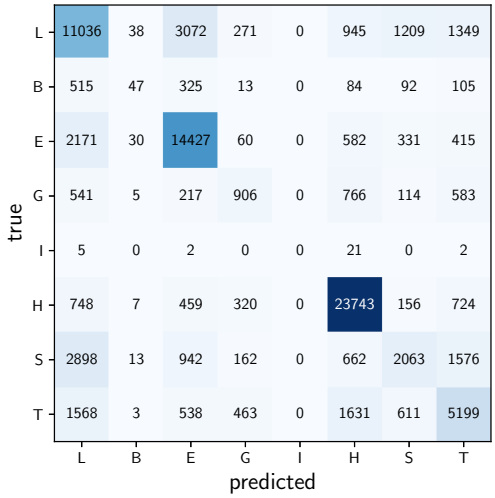

Fig. 2. Confusion matrix of the test set. Classes I and B are scarce in the dataset. Classes G, H and T have relatively high levels of confusion because they are all composed of turns, but with different lengths. Loops (L) are a loose category and easily misclassified.

4. RESULTS

The network described aboved was trained for 400 epochs with regularisation parameter $\lambda = 10^{-4}$, and learning rate $\mu = 10^{-4}$. Five of such networks are trained (with the stopping criteria decided via a validation subset[1]) and form an ensemble that reaches an accuracy of 69.23% on the test set, not far from the 71% reached by the state-of-the-art [3]. The aim of this work is not to outperform the state-of-the-art predictions, but to build a network with a moderately simple structure (to keep the calculation times of saliency maps on reasonable levels) and fair performance. We 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.

Figures 1 and 2 show more information on the network performance. Figure 1 shows the distribution of per-sequence mean accuracy over different sequence lengths. As it could be expected, the variance in accuracy increases with shorter sequences. Higher accuracies can be expected from sequences rich in helices and lower accuracies for sequences high in coils. Figure 2 displays the resulting confusion matrix of the predictions on the test set. It is similar to the ones obtained on other state-of-the-art networks [6]. The lack of π-helix (I) predictions is noticeable; its presence in the dataset is so rare that the machine keeps high accuracy even when ignoring it. β-bridges (B) are also largely misrepresented for the same reasons. Loops present high levels of confusion with other classes, probably due to the arbitrary discretisation into eight classes, while it has been pointed that the transition between structures and coils is not sharp [23].

Saliency maps can be used as additional evidence for theories around protein secondary structure. For instance, one point of concern has been the inclusion of inputs with different nature: a one-hot amino-acid input along a PSSM dense vector. Some authors [8, 7] embedded the one-hot vector into a denser space, reporting a marginal improvement in accuracy (0.5% and 0.4%, respectively). Spencer et al. [24] reported a 2% Q3 improvement by not including the one-hot amino-acids at all. Saliency maps provide information about the importance that different parts of the input have, so direct comparison of both kinds of inputs can be made by looking at their associated saliency values. To do so, each saliency map is split into two halves, corresponding to amino-acid and PSSM, respectively, and all the values of each half are summed up in absolute value to form a single saliency score. The comparison of such scores for all positions in the dataset is made in Figure 3, revealing that the PSSM inputs had four times or more relevance for making the classification decision in the great majority of positions, with around half of them having seven times or more relevance. We further validate these findings by training a second network that uses PSSM inputs but ignores the one-hot amino-acids, all the other things remaining equal. An ensemble of five of these networks reaches an accuracy of 69.36% on the test set, 0.14% points higher than

[1]The validation subset contains 512 sequences, 10% of the training set.

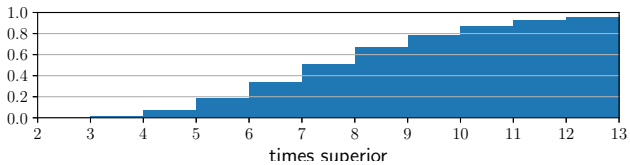

Fig. 3. Cumulative histogram with relative strength of PSSM saliency scores as compared to amino-acid saliency scores.

Labels	Architecture	Accuracy
Q8	Original	69.23%
	PSSM-only	**69.36%**
H / non-H	Right positions only	**86.60%**
	Left positions only	80.81%

Table 2. Accuracies on the test set for different network configurations. Ensembles of networks performed better by ignoring one-hot amino-acids. Networks that look on a window of size nine at the right side predict class H better than when they regard the nine left positions.

the original configuration (see Table 2).

Another useful application of saliency maps can be the study of spatial importance around the predicted position, i.e. where the network is looking at when classifying. Figure 4 includes the average saliency spatial profiles for each class. The construction of each profile has been made as follows. For each correctly predicted position of the class, the corresponding 42x19 slice of the saliency map is extracted and summed up over the feature dimensions, leading to a profile vector of size 19. Figure 4 shows the average of all of such profiles grouped by class, with profiles coming from both the training and test set. The lines reveal which positions were more relevant for predicting the class. For instance, the α-helices (H) hold some periodicity at the right side and large asymmetry. The periodicity goes in line with the structure of helices, which are a succession of turns. The asymmetry can point to a strong dependency on posterior amino-acids when the protein chain is formed. To validate this finding, we retrain two new networks with exclusively binary classification for the class H. The networks have an identical structure to the previous ones except that one of them ignores the input from the left positions and the other the ones from the right positions[2]. While the network making use of the left positions obtains an accuracy of 80.81% on the test set, the one utilising the right positions achieves 86.60%, supporting the idea that future positions are more significant in the formation of an α-helix (Table 2).

[2]This is achieved by zeroing out different halves of the parameters on the convolutional filters.

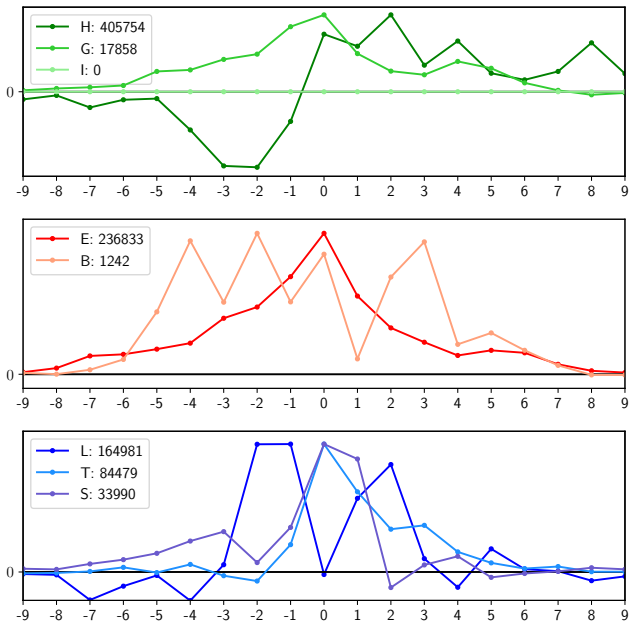

Fig. 4. Average saliency profiles for the eight classes. On top, the three helix classes; in the middle, the two β classes; below, the three coil classes. The legends show the number of profiles averaged over for each class. The focus is on the distribution of each class over the window positions, so each line has been independently normalized for a clearer visualisation of the shapes, and scales only share the zero reference.

5. CONCLUSIONS

This work demonstrates that saliency maps could help to explain black box decisions made by deep neural networks on the biological inference problem of predicting protein secondary structures. While these methods have been developed in the field of computer vision, their application to biological sequence analysis is novel. By this application, we reach two conclusions. Firstly, one-hot amino-acid inputs contain far less useful information than Position Specific Substitution Matrices, to the point of not losing significative performance with the omission of the earlier. Secondly, the prediction of α-helices relies more on amino-acids to the right of the predicted position than to the left, which may be a consequence of the biological processes in protein formation and folding. The benefits of the presented techniques can be of double value: helping biologist understand to get a deeper understanding of the underlying biological processes, and providing machine learning researchers with diagnostic tools for spotting flaws in their systems. Future work should focus on more advanced saliency maps techniques, as well as further methods for analysing and extracting information from them. A wider window width could also be explored, as it has been suggested that β-sheets rely on longer interactions [7, 25].

6. REFERENCES

[1] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[2] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, and Y. Zhou, "Sixty-five years of the long march in protein secondary structure prediction: The final stretch?," *Briefings in Bioinformatics*, vol. 19(3), pp. 482–494, 2018.

[3] A. Busia and N. Jaitly, "Next-Step Conditioned Deep Convolutional Neural Networks Improve Protein Secondary Structure Prediction," *arXiv:1702.03865v1*, 2017.

[4] J. Zhou and O. G. Troyanskaya, "Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction," *Proceedings of the* 31^{st} *International Conference on Machine Learning. JMLR: W&CP*, vol. 32, 2014.

[5] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields," *Scientific Reports*, vol. 6, no. 18962, 2016.

[6] C. Fang, Y. Shang, and D. Xu, "MUFold-SS: Protein Secondary Structure Prediction Using Deep Inception-Inside-Inception Networks," *arXiv:1790.06165*, 2017.

[7] J. Zhou, H. Wang, Z. Zhao, R. Xu, and Q. Lu, "CNNH_PSS: Protein 8-class secondary structure prediction by convolutional neural network with highway," *BMC Bioinformatics*, pp. 99–109, 2018.

[8] Z. Li and Y. Yu, "Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks," arXiv:1604.07176v1, 2016.

[9] V. I. Jurtz, A. R. Johansen, M. Nielsen, J. J. Almagro Armenteros, H. Nielsen, C. K. Sønderby, O. Winther, and S. K. Sønderby, "An introduction to deep learning on biological sequence data: Examples and solutions," *Bioinformatics*, vol. 33, no. 22, pp. 3685–3690, 2017.

[10] C. Olah, A. Mordvintsev, and L. Schubert, "Feature Visualization," *Distill*, https://distill.pub/2017/feature-visualization/, 2017.

[11] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," *arXiv:1704.02685*, apr 2017.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv:1312.6034*, 2014.

[13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE 10(7): e0130140*, 2015.

[14] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *arXiv:1703.01365*, 2017.

[15] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *arXiv:1512.02479*, 2015.

[16] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[17] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nature Methods*, vol. 12, no. 10, 931–4, 2015.

[18] R. K. Umarov and V. V. Solovyev, "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks," *PLoS ONE, 12(2): e0171410*, 2017.

[19] D. R. Kelley, J. Snoek, and J. L. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome Research*, vol. 6, no. 7, pp. 990–999, 2016.

[20] J. Lanchantin, R. Singh, B. Wang, and Y. Qi, "Deep Motif Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks," *arXiv:1608.03644*, 2016.

[21] A. Finnegan and J. S. Song, "Maximum entropy methods for extracting the learned features of deep neural networks," *PLoS Comput Biol 13(10): e1005836*, 2017.

[22] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences," *arXiv:1605.01713*, 2016.

[23] B. Rost, "Review: Protein secondary structure prediction continues to rise," J Struct Biol 2001 May - 134(2-3): 204-18.

[24] M. Spencer, J. Eickholt, and J. Cheng, "A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction," *Ieee/Acm Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 1, pp. 103–112, 2015.

[25] Y. Ni and M. Niranjan, "Exploiting Long-Range Dependencies in Protein β-Sheet Secondary Structure Prediction," in *Pattern Recognition in Bioinformatics*. 2010, pp. 349–357, Springer Berlin Heidelberg.