Predicting Secondary Protein Structures given the Sequence

William Mayhew, wm4g21
University of Southampton

1. Introduction

Accurately predicting the secondary structure of proteins from their amino acid sequence is a fundamental challenge in Computational Biology. Typical methods classify each residue into one of three structural states: α-helix, β-strand, and coil. Early work by Qian & Sejnowski in 1988 used neural networks for secondary structure prediction with reasonable accuracy [1]. This report re-implements and evaluates this neural network approach, while also exploring potential improvements utilising different preprocessing and more modern techniques such as Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs). Testing was conducted on two different datasets; the by Qian & Sejnowski dataset, and the RS126 dataset.

2. Background

In 1988, Qian & Sejnowski proposed one of the earliest neural network-based methods for predicting protein secondary structures from sequences [1]. They employed two feed-forward neural networks with multiple interconnected layers. The first network's input layer consisted of a sliding window of 13 amino acid residues plus a spacer, each represented by a unique binary code. The network aimed to predict the secondary structure class (α-helix, β-strand, or coil) of the central residue in the window. The second network utilised the results of the first network as its input, also operating on a sliding window approach with an output layer representing the three secondary structure classes. In both networks, specific weights were utilised dependent on the position of the amino acid and secondary structure in the window. Output biases were also added to the outputs of both networks.

However, Qian & Sejnowski's work encountered limitations in achieving optimal accuracy [1]. Despite various attempts to improve performance, such as modifications to input representations and network architecture, they achieved a Q3 accuracy (the percentage of residues correctly predicted into the three states) of 64.3%, falling short of the theoretical limit of 70%. This shortfall was attributed to potential information loss during training, indicating the need for more sophisticated approaches.

In 1993, Rost & Sander introduced several key enhancements to neural network-based methods, significantly improving prediction accuracy [2] . These improvements included:

- **Multiple Sequence Alignment (MSA):** Instead of using single protein sequences as inputs, Rost & Sander employed MSA representing amino acid profiles of related proteins. This approach captured additional structural information and patterns.
- **Balanced training:** Unlike traditional methods focused solely on overall accuracy, their neural network was trained using equal proportions (33%) of each secondary structure class. This approach particularly improved the prediction of the challenging β-strand class without compromising others.
- **Structure context training:** Rost & Sander introduced a second "structure-to-structure" neural network, which took the output of the first "sequence-to-structure" network as input. This network was trained to recognise and correct patterns in the predicted secondary structure.
- **Jury of networks:** To further enhance accuracy, they combined predictions from 12 different neural networks through a majority voting scheme.

As a result of these enhancements, the final "jury of networks" achieved an overall prediction Q3 accuracy of 69.7%, a substantial improvement over earlier methods which ranged around 62-66%.

Further explorations in secondary structure prediction were made by utilising Support Vector Machines (SVMs). These methods saw comparable but improved accuracy over traditional neural networks.

Sujun Hua and Zhirong Sun introduced a new method of secondary structure prediction based on SVMs [3]. Trained with multiple sequence alignments, six binary classifiers from a One-vs-Rest (OvR) and One-vs-One (OvO) strategy for each structural class were used to produce a tertiary classifier. Each binary classifier utilised different optimised window sizes. This tertiary classifier used several different methods on different sets of binary classifiers to get a final prediction. Results from these classifiers were compiled by a 'jury' to combine all the results, achieving a Q3 accuracy of 73.5%.

Similarly, Ward et al. [4] recognised the promising results from the use of SVMs and utilised a similar strategy. Instead of one-hot encoded MSAs, position-specific scoring matrices derived from PSI-BLAST searches were used to provide more informative evolutionary information. Different optimisations were also made, such as the use of a quadratic function kernel instead of a radial basis function kernel used in Hua & Sun's implementations [3]. This saw

an overall improvement with a Q3 accuracy of 77.07%.

2.1. Measuring Accuracy

Two common evaluation metrics are utilised, Q3 and Matthew's Correlation Coefficient. The Q3 accuracy measures the percentage of residues correctly predicted into the three structural states (α-helix, β-strand, and coil). The Matthew's correlation coefficient (C_X) provides a balanced measure of prediction quality for each structural class, ranging from -1 to 1, where +1 value corresponds to a perfect classification, and -1 corresponds to a perfectly opposite prediction.

2.2. Comparison of Approaches

Table 1 summarises the key results between the approaches proposed by Qian & Sejnowski in 1988, the enhancements introduced by Rost & Sander in 1993, and the alternative approaches with the use of SVMs by Hua & Sun in 2001 and Ward et al in 2003.

TABLE 1. COMPARISON OF RESULTS [1], [2], [3], [4]

Model	Q3(%)	C_C	C_E	C_H
Qian & Sejnowski (1988) [1]	64.3	0.41	0.31	0.41
Rost & Sander (1993) [2]	69.7	-	-	-
Hua & Sun (2001) [3]	73.5	0.64	0.52	0.51
Ward et al (2003) [4]	77.07	0.585	0.634	0.725

3. Implementations

The model discussed by Qian & Sejnowski [1] was re-implemented. An improved version of this model was also implemented, as well as a model utilising a support vector machine (SVM). The two new models leveraged the different techniques from across the different approaches. The datasets used for training and testing are the same as those used by Qian & Sejnowski [1].

3.1. The Dataset

The dataset used was the same as Qian & Sejnowski [1]. Two datasets were provided, one for training and one for testing. The training dataset consisted of 18105 samples and the testing dataset 3520 samples. As can be seen in 1, the classes in the training data are heavily imbalanced towards coils because these are generally more common. To attempt to achieve a balanced accuracy of all three classes, class weights can be utilised when necessary. This can also prevent overfitting and having a bias towards coils.

3.2. Qian & Sejnowski 1988 Model

The model architecture proposed by Qian & Sejnowski in 1988 [1] was implemented as a cascading sequential neural network model.

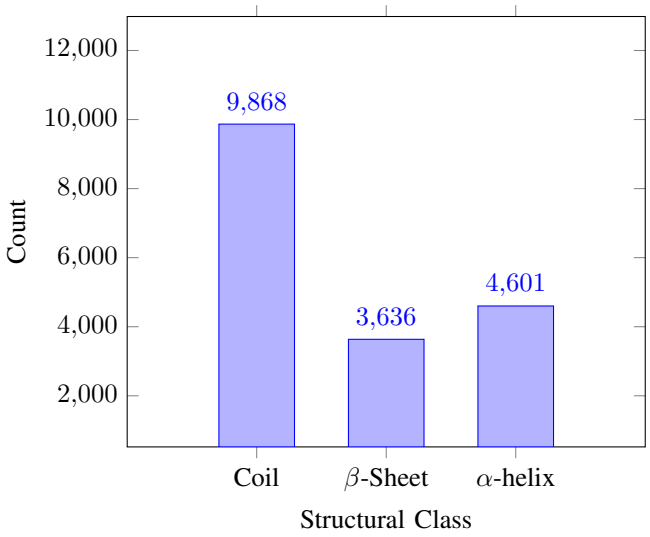

Figure 1. Training Class Distribution of Coils (C), β-sheets (B), and α-helices (H)

3.2.1. Input Representation. Initially, the sequences are translated into 21 bits using one-hot encoding where the amino acid letter is represented as an array with 0's where the corresponding letter is a 1. The secondary structures were also encoded into integers such that C is 0, E is 1, and H is 2. The sequences were then put into windows of size 13 as it was said to reach a performance maximum at this point [1]. The centre residue represented the amino acid to be predicted with the remaining residues (6 on either side) there to provide further structural information.

3.2.2. Network Architecture. The model architecture proposed by Qian & Sejnowski in 1988 [1] consisted of two interconnected feed-forward neural networks. The first network aimed to predict the secondary structure class of the central residue in a given window of amino acids. The input layer of the first network accepted a window of 13 one-hot encoded amino acid residues. This utilised 40 hidden units. The output layer of the first network consisted of three units, corresponding to the three secondary structure classes yielding the predicted probabilities for each class. The output biases were applied to these predictions.

The second network in the architecture took the output of the first network as its input, operating on a sliding window approach similar to the first network with a size of 13. The input layer of the second network accepted the predicted probabilities from the first network for a window of residues. Analogous to the first network, the second network comprised multiple hidden units to further process and refine the predictions from the previous stage. The output layer of the second network also had three units which were adjusted by the output biases representing the final predictions for the three secondary structure classes.

3.2.3. Result. Overall, a Q3 accuracy of 59.46% was achieved. The approach attempt to mimic Qian & Se-

jnowski's implementation [1] and was successfully able to produce similar results. MCC's of $C_C = 0.37$, $C_E = 0.29$ and $C_H = 0.31$ were achieved where only helix was significantly behind to the original model.

3.3. Qian & Sejnowski Improved model

An improved version of Qian & Sejnowski's model was implemented. This utilised more complex neural networks, and further preprocessing of the input data utilising profiles.

3.3.1. Input Representation. Similarly to Rost & Sander's implementation which used windows of amino acid-residue frequency for the input data, a Position-Specific Scoring Matrix was integrated utilising the multiple sequence alignments. Amino acids where assigned a probability in window sizes of 13 dependent on their background frequencies at their specific positions.

3.3.2. Network Architecture. This new model continued the same structure from the previous model, implementing a cascading neural network for structure-to-sequence then sequence-to-sequence. However, further improvements were implemented such as more hidden layers, and regularisation techniques including dropout and kernel regularisers. This helped reduce significant overfitting problems which was being encountered in the first model, improving the generalisability. Early stopping was also utilised to ensure that the models were stopping training at the optimal point where the loss was minimised.

3.3.3. Result. Overall, a Q3 accuracy of 62.53% was achieved. MCC's of $C_C = 0.34$, $C_E = 0.29$ and $C_H = 0.41$ were achieved. This model saw significant improvements in the helix predictions but also encountered a minor reduction in coil predictions. Therefore approach successfully improved upon the reimplementation albeit very minor. Figures 2 and 3 showcase the accuracy and loss graphs of both Qian & Sejnowski's [1] and the improved model. Early stopping was only utilised on the second implementation where training was halted at epoch's 3 and 6.

3.4. SVM Implementation

This implementation closely resembles the method discussed by Sujun Hua & Zhirong Sun [3] while also implementing methods from Rost & Sander [2]. Multiple binary SVM classifiers were utilised to form a 'jury'.

3.4.1. Input Representation. The input sequences were encoded similarly to Qian & Sejnowski's [1] implementation, without the use of MSA. These were put into a variety of window sizes which were found to be more optimal dependent on the binary class found by Hua & Sun [3] (Table 2).

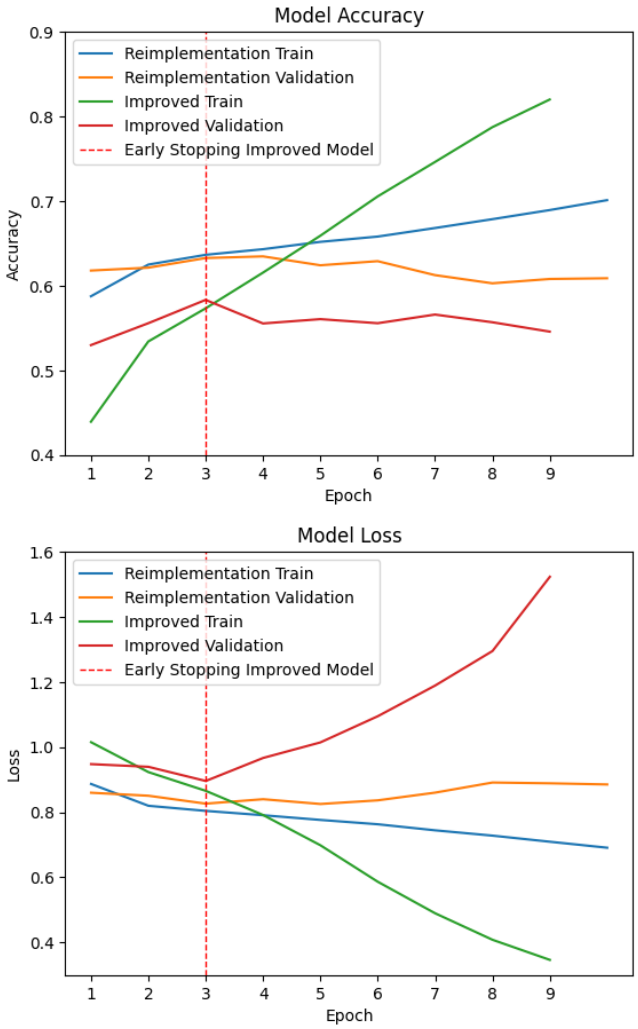

Figure 2. Accuracy and loss of the structure-to-sequence network from the reimplementation of Qian & Sejnowski's [1] and an improved version utilising techniques from Rost & Sander [2].

TABLE 2. BINARY SVM CLASSIFIERS

Classifier	Window Size	Classifies
C/¬C	7	Coil or not Coil
E/¬E	9	Strand or not Strand
H/¬H	11	Helix or not Helix
C/H	9	Coil or Helix
E/C	5	Strand or Coil
H/E	9	Helix or Strand

3.4.2. Model Architecture. 6 binary SVM classifiers were built for the structural classes. These SVM classifiers utilised class weights to ensure that over represented classes were not being biased during training.

5 further tertiary classifiers were utilised which took subsets the binary classifiers outputs to produce a prediction for all 3 classes. A Convolutional Neural Network (CNN) was also used which took the input of all 3 classes with a window size of 13, similarly to the implementation from

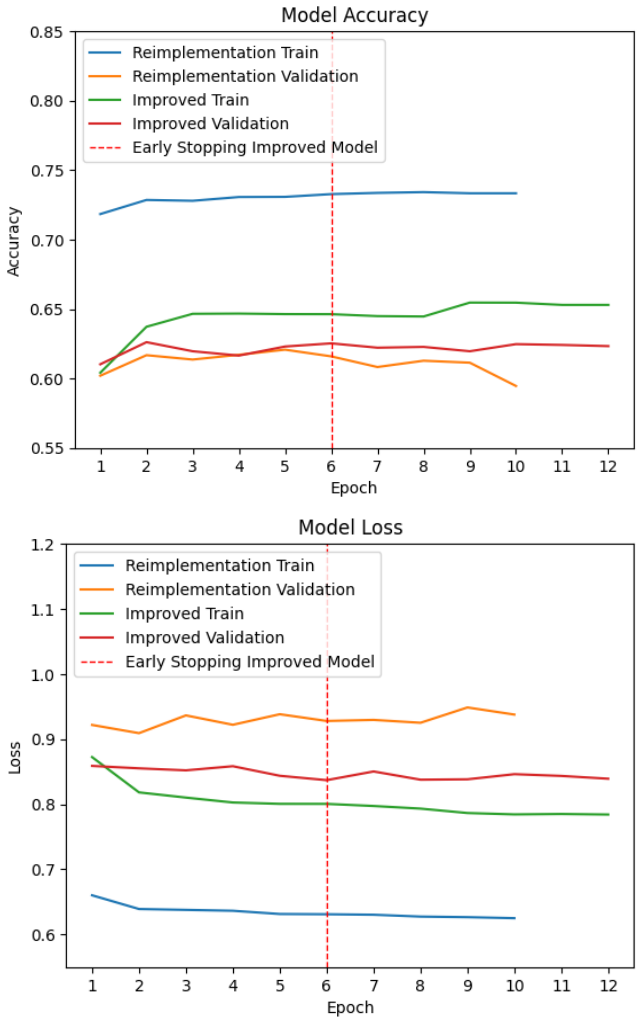

Figure 3. Accuracy and loss of the sequence-to-sequence network from the reimplementation of Qian & Sejnowski's [1] and an improved version utilising techniques from Rost & Sander [2].

Rost & Sander [2]. The tertiary classifiers mimic Hua & Sun's [3] implementation, employing those listed in Table 3. The CNN employed three 1D convolutional layers with 32, 64, and 128 filters utilising the ReLU activation function. Batch normalisation is applied as well as dropout layers with a rate of 0.3 to help prevent against overfitting. This is then passed into two dense layers and a final dense layer with 3 units for the multi-class classification prediction.

The predictions from all 6 classifiers were combined by a 'jury' where each classifiers vote was weighted dependent on its total accuracy and accuracy of each individual class.

3.4.3. Result. Overall, a Q3 accuracy of 62.36% was achieved. Although a similar approach to Hua an Sun [3] was taken, the multiple binary SVM classifiers and CNN model was unable to produce results similar. However, the CNN on its own was capable of producing results better results that all the re-implementations achieving an accuracy of 64.78% with correlation coefficients matching or bettering those as well.

4. Discussion

Overall, Qian & Sejnowski's implementation was successfully implemented and was able to produce similar results. This was a very basic model utilising two-cascading neural networks with minor preprocessing of the input data. A small improvement was also gained by utilising techniques discussed by Rost & Sander including class weights, multiple sequence alignments, and profiling. Compared to Rost & Sanders implementation which garnered an accuracy of 69.6% [2] on a different dataset, these improvements fall behind. It was likely that the preprocessing steps and input formatting were performed incorrectly which lead to the less than satisfactory results. The implementation somewhat replicating Hua & Sun's model also fell short of achieving what was expected. An accuracy beating the previous models was achieved but once again only saw minor improvements. However, the CNN on its own was able to achieve an accuracy of about 2.5% higher than the previous. This is likely due to the CNNs ability to utilise local patterns in the data.

In all implementations, C_C and C_H were fairly consistent, with C_E lagging behind. This was discussed as a common issue in early implementations where coils were being predicted correctly significantly more than strands and helices [1], [2]. More modern techniques improved upon this which led to a more balanced final result but this was unachievable in this instance likely due to the inadequate data processing. Balancing the dataset was attempted, as well as implementing a bias towards strands and helices but this unfortunately led to a significantly lower accuracy overall with very minor improvements in the respective accuracies.

4.1. RS126 Dataset Testing

Testing was also conducted on the RS126 dataset. This is a set of 126 non-homologous globular protein chains commonly used in other reports. Using this additional dataset meant that I better idea on the generalisation of these models can be made.

A final comparison of all the models and the implementations of these can be seen on Table 4.

Utilising the RS126 dataset, the results were relatively consistent with the original dataset. A large improvement was only seen in Hua & Sun's reimplementation which saw a increase from 62.1% to 68.55%, achieving the highest Q3 accuracy overall.

5. Conclusion

Accurately predicting the secondary structure of proteins from their amino acid sequence remains an important challenge in computational biology. This report explored two

TABLE 3. TERTIARY CLASSIFIERS

Classifier	Inputs	Accuracy (%)	C_C	C_E	C_H
SVM_MAX_D	C/¬C, E/¬E, H/¬H	59.89	0.37	0.27	0.32
SVM_TREE1	H/¬H, E/¬C	59.01	0.36	0.26	0.36
SVM_TREE2	E/¬E, C/¬H	59.72	0.37	0.28	0.33
SVM_TREE3	C/¬C, H/¬E	59.55	0.37	0.28	0.32
SVM_VOTE	C/¬C, E/¬E, H/¬H, C/¬H, E/¬C, H/¬E	59.69	0.34	0.26	0.32
CNN	MSA with window size 13	64.78	0.42	0.29	0.40
Overall	-	62.1	0.38	0.28	0.36

Classifiers tested on Qian & Sejnowski's dataset [1].

TABLE 4. COMPARISON OF RE-IMPLEMENTATION RESULTS

Model	Q3(%)	C_C	C_E	C_H	$Q3^{RS126}$(%)	C_C^{RS126}	C_E^{RS126}	C_H^{RS126}
Qian & Sejnowski (1988) [1]	64.3	0.41	0.31	0.41	-	-	-	-
Reimplementation	**59.46**	0.37	0.28	0.31	**62.12**	0.39	0.36	0.43
Improved Model	**62.53**	0.34	0.29	0.41	**57.33**	0.32	0.28	0.34
Rost & Sander (1993) [2]	-	-	-	-	69.7	-	-	-
Hua & Sun (2001) [3]	-	-	-	-	73.5	0.64	0.52	0.51
Modified Reimplementation	**62.1**	0.38	0.28	0.36	**68.55**	0.49	0.47	0.54
CNN Implementation	**64.78**	0.42	0.29	0.40	**62.18**	0.39	0.36	0.43
Ward et al (2003) [4]	77.07	0.585	0.634	0.725	-	-	-	-

Results in the first 4 columns from [1], [2], [3], [4] utilised the datasets from their respective research. The reimplementations here utilised the dataset from Qian & Sejnowski. The last 4 columns utilised the dataset from RS126.

influential early approaches proposed by Qian & Sejnowski in 1988 and Rost & Sander in 1993, reimplementing the first neural network model and building upon that which numerous techniques. Additionally, a support vector machine combined with a convolutional neural network approach inspired by later works was implemented.

The reimplementation of the Qian & Sejnowski model achieved a Q3 accuracy of 59.46%, not far off the reported 64.3% in their 1988 paper. However, the more advanced techniques introduced by Rost & Sander, such as multiple sequence alignments and profiling proved challenging to replicate accurately. The implementation of these techniques yielded a slightly higher accuracy compared to Qian & Sejnowski's with a Q3 accuracy of 62.52%.

The support vector machine implementation, incorporating elements from Hua & Sun's 2001 work, achieved a Q3 accuracy of 62.36%. The CNN used within this model was capable of predicting to a Q3 accuracy of 64.78% which was a significant improvement over all the models.

It is important to remember than different datasets were used for the reports excluding [1], thus it is expected than results may not be similar. But nevertheless, an improvement over the initial implementation was managed to be achieved leading to the possible conclusion that utilising convolutional neural networks may be the preferred model out of the ones observed.

It is important to remember the use of different datasets within in each report thus is can be expected that the results may not be similar. Due to this, additional testing utilising the RS126 dataset was used in conjunction because of its frequent usage in newer reports. With this dataset, results stayed relatively consistent aside from the reimplementation from Hua & Sun seeing a strong improvement. This may indicate that while neural network implementations are relatively consistent in generalising for various datasets, SVMs may be preferred for this specific task.

Furthermore, improvements in balancing results are needed. Although balanced results were achieved using SVMs with the RS126 dataset, the same cannot be said for the others. In the end, similar results for coils C_C and helices C_H were able to be achieved wile strands C_B remained relatively unchanged throughout, staying around 0.29 in each implementation. Later works could include further profiling techniques to identify strong patters and/or introduce calculated biases.

Overall, this exploration highlights the difficulties in reproducing and replicating complex machine learning models, even when following published methodologies. Differences in datasets, data preprocessing, architectural details, and training procedures can significantly impact performance. As the field continues to advance, it is crucial to establish standardised benchmarks, datasets, and evaluation metrics to facilitate fair comparisons and reproducibility across different techniques for protein secondary structure prediction.

References

[1] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, no. 4, pp. 865–884, Aug. 1988.

[2] B. Rost and C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 16, pp. 7558–7562, Aug. 1993.

[3] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach1," *Journal of Molecular Biology*, vol. 308, no. 2, p. 397–407, Apr. 2001.

[4] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Secondary structure prediction with support vector machines," *Bioinformatics*, vol. 19, no. 13, p. 1650–1655, Sep. 2003.