Bioinformatics, 2021, 1–7
doi: 10.1093/bioinformatics
Advance Access Publication Date: 01 02 2021
Manuscript Category

Project Report

Comparative analysis of GOR and Support Vector Machine methods for predicting the secondary structure of proteins

Thomas Isaac[1,*]

[1]Department of Bioinformatics, University of Bologna, Bologna, Italy

Abstract

Motivation: The Garnier-Osguthorpe-Robson (GOR) is a method used for predicting the secondary structure of proteins this is compared to the more recent machine learning method of using support vector machines (SVM) and evaluated. The prediction of secondary structure is vital for functional annotation and with the growing number of known sequences with a lack of determined structure prediction methods grow more useful. GOR uses a combination of information theory and Bayesian statistics to compute the likelihood of a helix, coil or beta strand whereas SVM is a supervised machine learning method able to classify the secondary structure type. Both methods are evaluated against a blind set of 150 sequences and a training set of 1248.

Results: The two methods successfully implemented showed a clear difference in the accuracy of being able to predict the secondary structure of proteins. From the two models produced from GOR and SVM there is a marked improvement in SVM Q3 accuracy and MCC. The accuracy per secondary structure conformation is more variable.

Contact: thomas.isaac@studio.unibo.it

Supplementary information: Supplementary materials are available at: https://github.com/Thomas0197/lb2-2020-project-isaac

1 Introduction

The function of proteins is core to bioinformatics and the understanding biological systems of which structure plays a fundamental role. Knowing the function of proteins can pave the way to fold recognition, new drug design and other important applications. Determining the three-dimensional structure is necessary. The most accurate and reliable way to do this is through experimental methods such as determining the electron density through X-ray crystallography. The problem with these methods is that they are very time consuming and expensive relative the number of proteins discovered. Prediction methods have been developed to help alleviate the problem such as Garnier-Osguthorpe-Robson (GOR) or more recently new machine learning methods. First generation prediction methods first developed between the 1960's and 1970's based on single amino acid properties using simple stereochemical principles or statistics [Chou and Fasman, 1974] to identify the local secondary structure motifs. Secondary generation prediction methods in the 1990's also evaluated the propensities of adjacent residues by using a sliding window between 3-51 residues when identifying the central residue [GOR, 1978]. The accuracy of these methods stalled at 60%. Third generation methods [Jpred4, 2015] are improved by the inclusion of sequence diversion and so take into account sequence conservation when making predictions. This coupled with advances in algorithms and large databases increased the accuracy to over 70%.

The second-generation method of GOR is analysed along with the third-generation technique of Support Vector machines (SVM) to make a comparison of their performance. This was done by training both methods through sequence profiles taken from multiple sequence alignments and testing them against a blind set of 150 sequences. The results show on average a higher performance using SVM most probably due to the limitation of GOR having to assume statistical independence for each

residue in the window when in comparison to SVM which can handle much larger feature spaces.

2 Materials and Methods

2.1 Statistical analysis of the dataset

A basic statistical analysis was done for both the training and blind test set to assess their representation and suitability. In particular interest is the distribution of secondary structure conformations as seen in figures 5 and 6 the difference is minimal with slightly more helixes found in the blind test set. As seen in figures 1 and 2 there is also very little difference in secondary structure conformations per residue and the total proportion of residues in each set (figures 3 and 4). This is to be expected with each amino acid having different atomic compositions and stereochemical properties leading them to each favour a certain distribution of secondary structure conformations of which both sets follow as expected. The training set also covers all main SCOP classes and distribution of animal kingdoms (figure 7) representing the Swiss-Prot database well. The dataset shows no clear signs of bias in taxonomy, length, or proportion of secondary structure conformations.

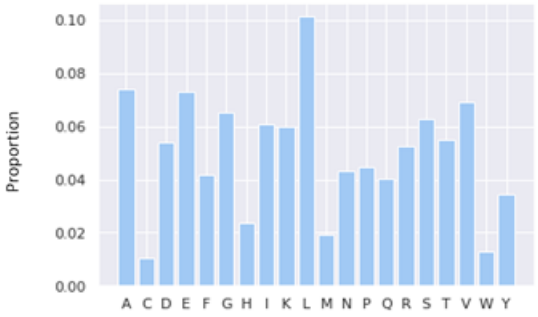

Fig 3. The relative abundance of each type of residue in the blind dataset

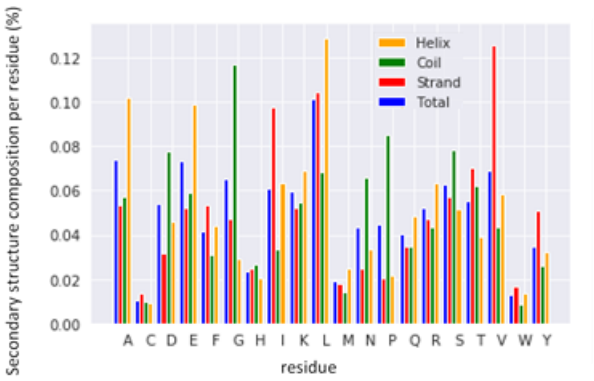

Fig 1. The secondary structure conformation proportions in relation to each residue in the blind dataset.

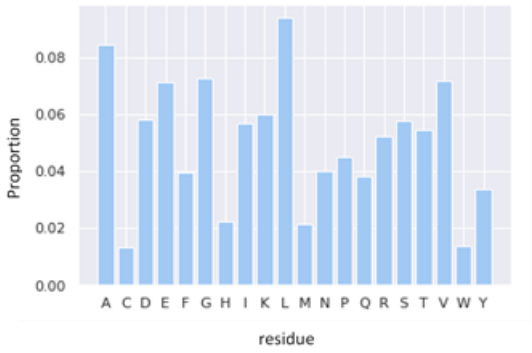

Fig 4. The relative abundance of each type of residue in the jpred4 dataset

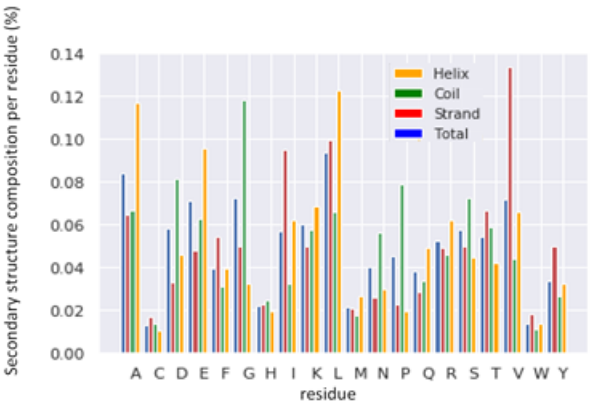

Fig 2. The secondary structure conformation proportions in relation to each residue in the jpred4 dataset.

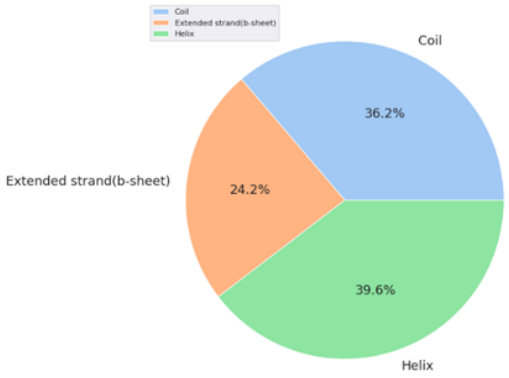

Fig 5. Percentage of secondary structure conformation in the blind dataset

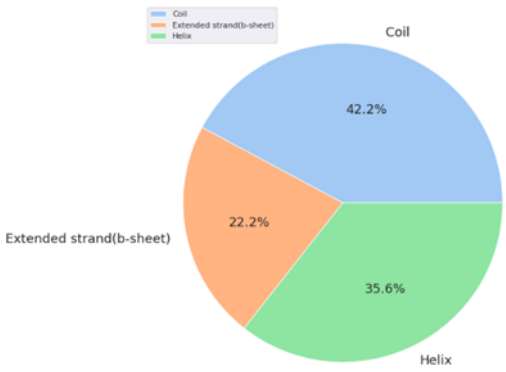

Fig 6. Percentage of secondary structure conformation in the jpred4 dataset

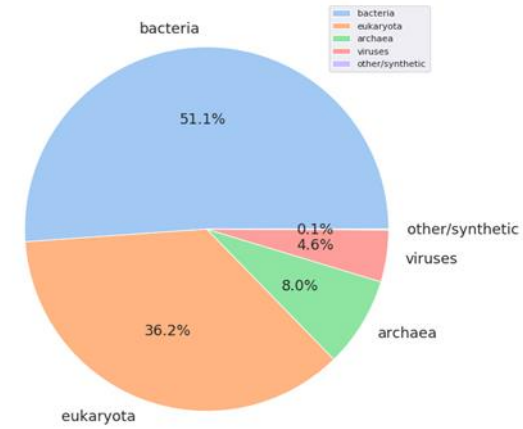

Fig 7. Distribution of the super kingdom in the jpred4 dataset

2.2 Datasets

The dataset used is comprised of 2 sets: the training and the blind test set. These are used to evaluate both GOR and SVM. Both datasets have co-responding DSSP files (Kabsch W et al, 1983) created these are then mapped to their original fasta sequence to evaluate the results.

2.3 Training Set

The training set is comprised of the Jpred4 dataset produced from 1987 fasta sequences containing a single representative of each super family domain in SCOPe. Sequences with a length between 80 and less than 800 with a resolution under 2.5 were kept. Multi chains, fragmented domains and those with missing or partial DSSP files were also removed to ensure the accuracy of the predictions compare with reality. Any structures from the Protein Data Bank (PDB) which showed inconsistencies against the DSSP or ASTRAL were removed for the same reason. This left 1348 structures left in the starting training set.

2.4 Blind Test Set

A blind test set was produced to compare against the training for both GOR and SVM. This was done by taking from the PDB structures which met the following criteria:

- A resolution equal or below 2.5 Å
- Method X-ray crystallography
- Chain length between 50 and 300 residues
- Release date after January 30th, 2015

This returned a possible 21417 structures from which all non-identical chain fasta sequences were downloaded from the PDB. Internal redundancy was then reduced within the set using mmseqs2 (Steinegger M and Soeding, 2018) to cluster all the sequences together with a sequence identity of 30%. One from each cluster with the best resolution was picked leaving 7970 cluster representatives. These were then reduced further by removing sequences showing an external redundancy of above 30% with the Jpred4 training set using an e-threshold of 0.01. This was accomplished using Blastp (Altschul et al, 1997), this required using Blastmakedb to create a database based on the whole Uniprot database (Uniprot consortium, 2018). This left 6904 cluster representatives and 150 sequences were randomly selected to be placed into the blind test set.

2.5 Sequence Profiles

The best way to see whether a residue is structurally or functionally important is to look at multiple sequence alignments. This can be seen in sequence profiles with those structurally important more represented out of a base value total of 1. These were created for both the training and the blind test set using PSIBLAST (Position-Specific Iterative Basic Local Alignment Search Tool) each profile is a representation of a protein family for each specific target protein used. It produces position specific scoring systems (PSSM). Each of the 20 amino acids in each position are measured in their frequency by which they occur in the original alignment. PSIBLAST was run for each file contained in the Training and Blind test set each producing their own PSSM, it was run for 3 iterations against the database, an e value threshold of 0.01 was selected and the number of alignments set to 10000(due to the large database). Sequences detected with a score threshold below 0.01 were deleted, this left 133 PSSM's left in the blind test set and 1252 in the training set. The sequence profile from the PSSM files was extracted by a python program and each value within was normalized by dividing by 100. Upon further analysis some profiles were seen to contain only zeros. Using a python script any sequence profile found to contain only zeros was deleted. This left 129 sequence profiles for the blind test set and 1204 for the training set.

2.6 GOR

The Garnier-Osguthorpe-Robson (GOR) method was created in 1978 and uses a combination of information theory and Bayesian statistics to try and predict the secondary structure of proteins. Since its creation, several improvements (Rost et al, 1994) and refinements have been implemented mainly benefiting from the inclusion of evolution information in the sequence profiles and larger databases. The GOR method is based on the idea of the information function:

$$I(S;R) = \log \frac{P(S|R)}{P(S)}$$

- S is one of the three secondary structural conformations
- R is one of the 20 amino acid residues
- P(S|R) is the conditional probability for observing the structural conformation S when a residue R is present
- P(S) is the probability of observing the secondary structure S

If we take the independence assumption that states the statistical independence of the nearby residues to that of a central residue, it allows us to implement a sliding window. The best window size was found to be 17 so 8 residues before the central residue and 8 after. The formula followed to then create a window based GOR method becomes:

$$I(S; R_{-d}, \ldots, R_d) = \log \frac{P(S,R_{-d},\ldots,R_d)}{P(S)P(R_{-d},\ldots,R_d)}$$

$$= \log \prod_{k=-d}^{d} \frac{P(R_k,S)}{P(S)P(R_k)}$$

$$= \sum_{k=-d}^{d} I(S; R_k)$$

Where d in the equation is the window size, so 8 in the version of GOR implemented.

- P(Rk, S) : frequency of residues of type R observed at position k in windows where the central residue R0 is in conformation S
- P(Rk) : frequency of observed residues at position k
- P(S) : frequency of each conformation S

GOR calculates 3 different information assumptions, one for each type of secondary structure (helix, coil and beta strand). The sum for each calculation for that specific central residue position is calculated and the highest is chosen as the predicted secondary structure. The implementation of GOR was done via two python scripts, one to train the model and another to calculate the predictions based on the model created. The training script took as input sequence profiles and their corresponding DSSP files to calculate 4 models one for each type of secondary structure and a final for the overall frequencies for the windows. The prediction script takes as input the 4 models and the sequence profiles of those to be predicted, using the equation above it calculates the predicted secondary structure conformation

2.7 Super Vector Machines

Support Vector machines (SVM) are supervised learning models that can be used for classification and regression models. Hyper-planes separate the classes by the largest margin (most accurate), to be able to identify the maximum hyper-plane we need to find the best supporting vectors (closest points to the hyper-plane) these being the most important training points. To find the linear separator (or hyper-plane), we use the following equation:

$$\mathbf{w}^\top \mathbf{x} + b = 0$$

We can then say from this that :

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq \rho/2$$

Where p is the margin to be maximised (the distance between the two closest points of a different class). As there are many possible linear separators, to find the optimum solution only those with support vectors that the inequality in the equation above becomes an equality are considered. Rescaling of w and b by $\rho/2$ we obtain that the distance between a support vector and the hyperplane by the following:

$$r = \frac{y_i(\mathbf{w}^\top \mathbf{x}_s + b)}{||\mathbf{w}||} = \frac{1}{||\mathbf{w}||}$$

$$\rho = 2r = \frac{2}{||\mathbf{w}||}$$

Now it is possible to formulate a quadratic optimization problem to minimize the function $\frac{1}{2}||w||^2$ under the constraint $y_i(w^T x_i + b) \geq 1(\forall i)$. This solution involves constructing a dual problem where a Lagrange multiplier αI is associated with every inequality constraint.

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

Subject to:

$$\alpha_i \geq 0$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

With the solution to w and b being:

$$\mathbf{w} = \sum_{s} \alpha_s y_s \mathbf{x}_s$$
$$b = y_k - \sum_{s} \alpha_s y_s \langle \mathbf{x}_s, \mathbf{x}_k \rangle$$

The formula described above is suitable only for linearly separable problems for non-linear separable problems, use of the soft margin classification can be done. Slack variables ξi can be added to allow a degree of misclassification for difficult examples. We can modify the quadratic optimization function to:

$$\text{minimize} \qquad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\xi_i \geq 0, \forall i$$

The hyper-parameter C can be viewed as a way to control overfitting, a high C value will allow a narrow margin and a low amount of misclassification with a low C value the opposite.

The SVM was performed using the LIBSVM package [Chang and Lin, 2011] one of the most currently popular SVM implementations supporting different SVM formulations and multi-class classifications. It consists of two main subroutines SVM training and SVM prediction. SVM training will take as input a training file (the Jpred4 dataset) and output a model. The training file first must be formatted correctly to be read by the LIBSVM package; this was accomplished with a python script. The script would state whether it was a helix, coil or stand into numerical classes of 1, 2 and 3 and place the whole window profile for each central residue onto a single line. The SVM training has several different parameters and kernel functions available to choose from. A grid search was applied to find the most suitable cost (C) and gamma (γ) parameters and a cross validation performed to select the best performing model. The results of which can be seen in table (1) showing a C 1.5 and gamma 0.25 to be the most suitable parameters. The parameters run for the SVM train were:

- t 2(the default kernel type- radial basis function: exp(-gamma*|u-v|^2))
- C 1.5
- G 0.25

This was run against the whole training set to create a model to use against the blind test set. SVM prediction requires the model outputted from SVM training and a test file formatted as before, this returns a file with the predictions to be evaluated.

	Gamma (γ)			
Cost		0.25	0.5	2
	1.5	0.616	0.549	n/a
	2	n/a	0.545	0.166
	4	n/a	0.549	0.157

Table 1. Grid search of the Cost(C) and Gamma(γ) parameters and their resulting MCC value

2.8 Cross Validation

Cross Validation was performed on the training set for GOR and SVM methods to help evaluate the model's ability to predict new structures and the data's quality over the 5 sets. After creating the training sequence profiles 1204 were left and these were split into 4 equal sets of 241 with 1 set containing 240 to total 5 sets overall. The sequences were first randomly ordered, their ID's taken and placed into a file (can be seen in the supplementary materials). A python script would then read in the list of ID's to select which profiles would be used to create the model. When performing cross validation 1 set would be used as the test and the remaining 4 as the training to produce the model.

2.8 Evaluation

The evaluation of the results consists of comparing the real data (DSSP) against the predicted results of both GOR and SVM methods. This is a multi-class problem due to the possible 3 states the secondary structure can be classed into. A 3-class confusion matrix was created via a python script and a Q3 overall accuracy calculated.

$$Q_3 = \frac{p_{HH}+p_{EE}+p_{CC}}{N}$$

Q3 calculates the overall accuracy by summing the total correct (true positive) predictions and dividing by the total. Binary matrices for each type of secondary structure can then be made from the 3-class matrix to calculate:

Sensitivity (SEN) – a measure of the proportion of true positives that are correct.

Positive predictive value (PPV) – a measure of the proportion of positive and negative results.

Matthews correlation coefficient (MCC) – a way of describing the true and false positives.

The formula's displayed below:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Sen = \frac{TP}{TP+FN}$$

$$PPV = \frac{TP}{TP+FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$$

3 Results

As was seen previously in table (1) the best SVM parameters based off cross validation were C 1.5 and gamma 0.25 and so this was the chosen model to perform the GOR prediction for the blind test set on. The results for both GOR and SVM show no real difference in the standard deviation (table 2) between the split sets with a maximum of 0.3 for the SVM accuracy and so shows an even spread of the non-biased dataset. As shown in the cross validation (table 2), SVM reports better predictions in terms of overall accuracy and MCC this is reflected in the blind test set as well. The blind test set (table 3) reports back a near identical accuracy for both GOR and SVM in relation to the cross validation, the MCC for GOR is however slightly lower than expected by 2%. The slight variation from the blind test set to the cross validation is most probably due to the small sample set of 129. This could also be a cause for the accuracy being higher in GOR for strands then coils but having an equal MCC.

Cross Validation	GOR	SVM
H	ACC = 0.800 ± 0.009 MCC = 0.555 ± 0.017 SEN = 0.665 ± 0.020 PPV = 0.745 ± 0.006	ACC = 0.868 ± 0.005 MCC = 0.709 ± 0.007 SEN = 0.800 ± 0.010 PPV = 0.822 ± 0.008
E	ACC = 0.786 ± 0.005 MCC = 0.455 ± 0.013 SEN = 0.686 ± 0.013 PPV = 0.512 ± 0.017	ACC = 0.863 ± 0.006 MCC = 0.571 ± 0.013 SEN = 0.559 ± 0.023 PPV = 0.759 ± 0.012
C	ACC = 0.746 ± 0.004 MCC = 0.476 ± 0.008 SEN = 0.658 ± 0.007 PPV = 0.720 ± 0.014	ACC = 0.782 ± 0.003 MCC = 0.568 ± 0.005 SEN = 0.823 ± 0.006 PPV = 0.710 ± 0.012
Average	ACC = 0.778 ± 0.005 MCC = 0.495 ± 0.011 Q3 = 0.651	ACC = 83.758 ± 0.307 MCC = 0.616 ± 0.006 Q3 = 0.751

Table 2. The results of GOR and SVM on the 5-fold cross validation with all the scoring indexes evaluated. ± signifying the standard deviation.

Blind test set	GOR	SVM
H	ACC = 0.768 MCC = 0.511 SEN = 0.613 PPV = 0.767	ACC = 0.855 MCC = 0.697 SEN = 0.772 PPV = 0.856
E	ACC = 0.789 MCC = 0.457 SEN = 0.661 PPV = 0.539	ACC = 0.870 MCC = 0.607 SEN = 0.565 PPV = 8187
C	ACC = 0.746 MCC = 0.460 SEN = 0.689 PPV = 0.638	ACC = 0.775 MCC = 0.557 SEN = 0.843 PPV = 0.644
Average	ACC = 0.768 MCC = 0.476 Q3 = 0.651	ACC = 83.304 MCC = 0.620 Q3 = 0.750

Table 3. The results of GOR and SVM on the blind test set with all the scoring indexes evaluated.

4 Discussion

There are clear differences in the level of accuracy between the two methods with SVM predictions being 10% better in Q3(75%). The statistical analysis is a fair representation of the protein space with low internal redundancy and high variation. As expected, the sensitivity of strands is much lower than that of Helix's and Coils due to the fact that stands happen over a long range of residues for their interaction. This is a limitation within the prediction methods as we assign the secondary structure over a smaller local window, this appears to be less of a problem for the GOR blind test set (table 3).

GOR's worse performance is most likely due to the assumption of statistical independence and may be too much of a simplification of the problem in comparison to SVM. The prediction of the blind test set could be improved by using a larger a larger dataset as each set in the cross validation contained more sequences than the actual blind test set and probably the reason for the larger variation in MCC from the cross validation in GOR. The dataset being small due to computational reasons. To further try and improve upon the predicted results you can try to add known evolutionary information about secondary structure such as the minimum size required to form a helix is four residues and 2 to form a strand. This could help to add context for the methods when attempting a prediction to increase the accuracy further.

5 Conclusion

Prediction of structure is important due to the ever-increasing number of protein sequences being found and so the number of sequences without known structure to be able to infer functionality. Both GOR and SVM can predict the secondary structure to a high level with a Q3 of 65% for

the blind test in GOR and 75% in SVM. The SVM method is however better at performing a prediction on nearly all levels though this method is more computationally intensive and time consuming.

References

Nucleic Acids Research, 2018. UniProt: a worldwide hub of protein knowledge. 47(D1), pp.D506-D515.

The Protein Data Bank H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) Nucleic Acids Research, 28: 235-242. doi:10.1093/nar/28.1.235

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389–3402

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Steinegger M and Soeding J. Clustering huge protein sequence sets in linear time. Nature Communications, doi: 10.1038/s41467-018-04964-5 (2018).

Garnier, J., Osguthorpe, D. J., and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of molecular biology, 120(1), 97–120.

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015). Jpred4: a protein secondary structure prediction server. Nucleic acids research, 43(W1), W389–W394.

Prevelige, P. and Fasman, G. D. (1989). Chou-fasman prediction of the secondary structure of proteins. In Prediction of protein structure and the principles of protein conformation, pages 391–416. Springer

Kabsch W, Sander C (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". Biopolymers. 22 (12): 2577–637

James A. Cuff Geoffrey J. Barton.(1999). “Evaluation and improvement of multiple sequence methods for protein secondary structure prediction”

Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. Proteins: Structure, Function, and Bioinformatics, 19(1), 55–72

Cuff, J.A. and Barton, G.J. 2000Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40502–511