Protein Secondary Structure Prediction

Francesco Codice'

January 17, 2022

Abstract

Motivation: Protein secondary structures are locally stabilized three-dimensional segments of proteins chains. The availability of reliable in silico models to predict secondary structures can provide information about the function and can be useful in the prediction of the tertiary structure.
Methods:: The GOR method approach will be employed first and then a non linear SVM classifier will be tested.
Results:: As expected the GOR method obtains an overall Q3 accuracy of 0.62 on the blind-set. The SVM reach 0.69 of Q3 accuracy on the blind set.
Contact: francesco.codice3@studio.unibo.com

1 Introduction

1.1 Motivation

Protein secondary structures are locally stabilized three-dimensional segments of proteins chains. The two most common types of secondary structures motifs are alpha-helix and beta-strands. The stabilization of secondary structures elements is due to *hydrogen bonds*: this electrostatic force of attraction occurs between the carboxyl oxygen atoms and the amide-group hydrogen.

The problem of the protein secondary structure prediction was central in the field of bioinformatics, since the prediction of secondary structure is more tractable than the prediction of the tertiary structure that is still problematic.

Adopting a reductionist approach to predict the secondary structure can be useful for the prediction of the whole protein three-dimensional structure.

1.2 Secondary Structure Prediction

The first important attempt to deal with this problem is known as the *Chou-Fasman method*, a model based on the relative frequencies of different amino-acids with respect to the belonging secondary structure. Indeed different amino-acids, due to their physico-chemical properties, have different probabilities to form different secondary structure conformations; for example alanine and glutamic acids are common in helix conformations. This method takes into consideration only the propensity of single amino acids to be part of a particular secondary structures without considering the local environment of the residue. [1]

In 1978 *Garnier et al.* [2] introduced a new method based on *information theory*; the method, named GOR, is based on the conditional probability of a certain amino-acid to be present in a secondary structure conformation given the relative neighbors environment probabilities. With this method the accuracy reached was in the 60%-65% range.

In the 1990s the third generation methods for the prediction of the secondary structure have been introduced; those methods are mainly base on sophisticated machine learning methods that take into consideration evolutionary information such as *multiple sequence alignments*. With this generation of methods the accuracy reached the 80% accuracy threshold.

1.3 Manuscript Approach

The main goal of this manuscript is to test and compare two different secondary structure prediction methods: the *GOR method* and a *Support Vector Machine Method*. The data used to train the methods come from JPred paper [3] ; this data are also used to test the model in a 5-fold cross-

validation. For this manuscript we also generated a blind-set containing 150 proteins with sequence identity lower than 30% with respect to the JPred4 dataset.

The results show how the machine learning based Support Vector Machine method greatly overcome the performances obtained using the GOR method.

2 Material and Methods

2.1 Training Set Characteristics

The training set used in this work is obtained from the *JPred4 Training Set*, [3] as described in *JPred4: JNet training (v.2.3.1) details*. The authors started from SCOP superfamily structure-based representative sequences [4] so that the internal redundancy, in terms of evolutionary relationships, is reduced. They started from 1987 representative sequences and they performed a filtering based on structural resolution ($<$ 2.5 Å) , sequence length (between 30 and 800 residues), missing DSSP information (if $>$9 consecutive residues) and other structural inconsistencies. At this step the authors separated 1357 sequences from the total of 1507 proteins to build the training and blind set. From the 1357 proteins a total of 9 sequences are removed as the produced PSI-Blast output is empty so we end up with a total of *1348 proteins*.

Starting from this dataset we proceed by building sequence profiles using PsiBlast [5] against SwissProt (E-value threshold 0.01), all sequences with empty profile are filtered out (144 sequences). The final training set is composed of 1204 sequences with relative profile and secondary structures.

2.1.1 Sequences lengths

We start by analyzing the sequence length distribution to have an overview of the training-set composition (*Figure 1*). In the *Table 1* we reported the basic values for number, mean, median and standard deviation in terms of sequences lengths.

Table 1: Statistics on protein lengths

Measure	Value
Number of proteins	1348
Arithmetic Mean	162
Median	131
Standard Deviation	104

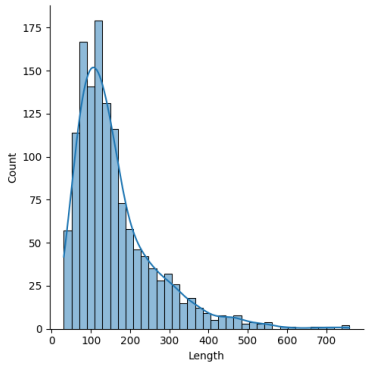

Figure 1: Domain Length Distribution

2.1.2 Secondary Structures Abundance

We proceeded by analyzing the relative abundance of secondary structure conformations in the given dataset. We can observe from *Figure 2* that the most common conformation reported is Coil as the Coil conformation is assigned with any DSSP code that is not H, B or E.

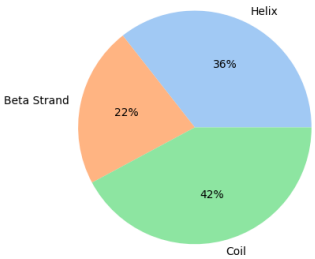

Figure 2: Secondary Structure Conformations Abundance

As our next step we proceeded by analyzing the different relative abundance of SCOP Structural Classes, these classes group together structures with similar secondary structure composition [4] .

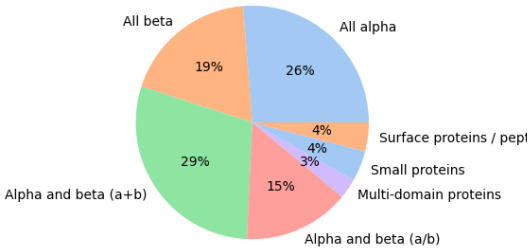

Figure 3: Structural classification SCOP Class - Pie Chart

2.1.3 Comparative amino acid composition

In this step of the analysis we proceed by studying the amino acid composition of the whole dataset (*Figure 3*) and the relative abundance of secondary structure motifs with respect to different amino acids (*Figure 4*). From this latter plot we can observe that the aminoacids Alanine, Leucine, Methionine, Glutamic Acid and Glutamine have n high propensity to be part of an Helix. Aspartic Acid, Proline, Glycine, Asparagine and Threonine show great propensity to be part of a coil. The amino acids that are mainly related to Beta Strands are Valine and Isoleucine.

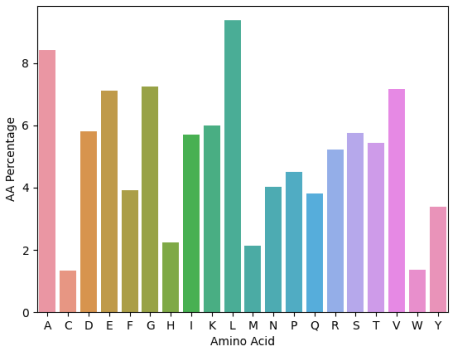

Figure 4: Amino acids frequency

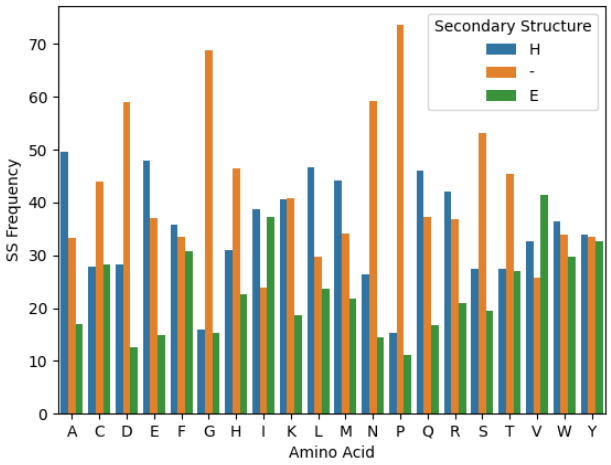

Figure 5: SS Motifs percentages with respect to different amino acids

2.1.4 Taxonomic Classification

A pie chart representing the relative frequencies of the Superkingdoms among the training set has been computed. As we can observe in *Figure 4*. the majority of the proteins belongs to *bacteria* with 51%, followed by *eukaryota* with 36 % and by *archea* with 8 %.

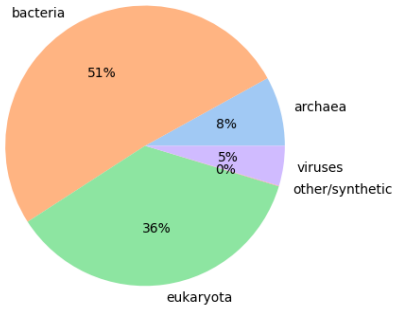

Figure 6: Superkingdom abundance pie chart

2.2 Blind Set Characteristics

The *blind set* is fundamental to assess the quality of the predictions. It is generated by gathering all the PDBs that respect the following properties:

- resolution lower than 2.5 Å
- sequence length between 50 and 800 residues
- deposit date after Jan 2015

Once about 40.000 sequences had been obtained we proceed with the reduction of internal redundancy. Using *MMseqs2* [6] with a greedy set cover approach we reduce the internal redundancy below 30% of sequence identity. Then a reduction of the external redundancy is performed in order to remove any protein that has any significant match with >30% S.I. against the JPred training set. To do that we performed a *blastp* search (0.1 E-value threshold) of our sequences against the JPred training set. Once the filtering is completed the secondary structures are retrieved using DSSP [7] .

In the end we randomly select 150 sequences from the dataset and we compute the sequence profiles using *PsiBlast* [5] (E-value threshold 0.01). In the end we obtain 150 proteins in the blind set of which 17 have an empty sequence profile that is replaced with the one-hot matrix corresponding to the sequence.

2.3 GOR Method Description

The *Garnier-Osguthorpe-Robson* is a method introduced in 1978 for the prediction of the protein secondary structure. [2] , it is based essentially on *information theory* concepts and on Bayesian statistics. As we are doing in the *Chou-Fasman* method we base the prediction of the secondary structure on the amino acid propensities. In this case for each residue we are considering the sequence local context (neighbors residues); given a window's length w for each residue we select $d = \frac{w-1}{2}$ residues both on the right and on the left to be taken into consideration.

The assigned secondary structure conformation is the one with the highest propensity score.

The main goal is to compute the *information function*, in order to evaluate to which extent the presence of the window-residues context influences the probability of having a certain protein's secondary structure conformation.

For each residue R with a sliding window equal to w ($d = \frac{w-1}{2}$) the information function is computed as follows:

$$I(S; R_{-d}, \dots, R_{+d}) = log\frac{P(S|R_{-d}, \dots, R_d)}{P(S)} = \\ = log\frac{P(S, R_{-d}, \dots, R_d)}{P(S)P(R_{-d}, \dots, R_d)} \tag{2.1}$$

Where $P(R)$ and $P(S)$ are the *marginal probabilities* and $P(S, R_{-d}, \dots, R_d)$ is the *joint probability.*

As the computation of the joint probability with respect to the w residues of the windows would need a large database and an high computational cost we assume the statistical independence of the residues in the window. Given this assumption we obtain

$$P(R_{-d}, \dots, R_d) = \prod_{k=-d}^{d} P(R_k) \tag{2.2}$$

The *information function* is computed in the following way

$$I(S; R_{-d}, \dots, R_{+d}) = log\frac{P(S, R_{-d}, \dots, R_d)}{P(S)P(R_{-d}, \dots, R_d)} = \\ = log \prod_{k=-d}^{d} \frac{P(R_k, S)}{P(S)P(R_k)} = \sum_{k=-d}^{d} log\frac{P(R_k, S)}{P(S)P(R_k)} \tag{2.3}$$

The predicted secondary structure is the one with the highest information function with respect to the specific window.

$$S^* = \text{argmax}_S I(S; R_{-d}, \dots, R_d) = \\ = \text{argmax}_S \sum_{k=-d}^{d} I(S; R_k) \tag{2.4}$$

In the specific implementation used in this project the *sequence profile* is taken into consideration in the computation of the information function; evolutionary information improve the quality of the predictions.

2.4 Support Vector Machine

The *support vector machine* is a machine learning method that is widely adopted both for classification and for regression tasks.

Given a set of samples belonging to two different classes ($y = \pm 1$) the training algorithm guarantees to find the best separating hyperplane $< \vec{w}, \vec{x} > +b = 0$ between the two classes.

The learning of the hyperplane parameters is based on the maximization of the *margin* between the two classes that corresponds to the minimization of the norm of w. Furthermore a constraint on the optimization problem should be set: we need to ensure that the samples are divided by the hyperplane. $y_i(< \vec{w}\vec{x} > +b) \geq 1$. To guarantee the satisfaction of the constraints the *Dual Lagrangian* is used: the optimization will consist in the learning of the lagrange multipliers α_i. The w and the b can be then computed on the basis of the lagrange multipliers selected and the support vectors found.

In our specific case we are going to adopt a *soft margin* approach to introduce a certain degree of tolerance in the classification. The function to be minimized is

$$min\ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \tag{2.5}$$

The constraints to be imposed are

$$y_i(< \vec{w}, \vec{x} > +b) \geq 1 - \xi_i \quad \xi_i \geq 0, \forall i \tag{2.6}$$

where ξ_i is a *slack variable* that can be seen as an upper bound of the classification error for the sample i. The C parameter is a tradeoff parameter between the error and the margin, high C values corresponds to *hard margin* and instead low C values corresponds to *soft margin*.

Implementing the *dual lagrangian* we obtain the following optimization problem to be solved

$$max\ \sum_{i=1}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j < x_i, x_j > \tag{2.7}$$

with as constraints

- $0 \leq \alpha_i \leq C\ \ \forall i$
- $\sum_{i=1}^{n}\alpha_i y_i = 0$

This optimization problem is solved using *quadratic programming* algorithms the final solution can be find using the following formulas

$$w = \sum_s \alpha_s y_s x_s \tag{2.8}$$

$$b = y_k(1 - \xi_k) - \sum_s \alpha_s y_s < x_s, x_k > \tag{2.9}$$

The classification function is

$$f(x) = \sum_s \alpha_s y_s < x_s, x > +b \tag{2.10}$$

- if positive the sample belongs to $y = +1$
- if negative the sample belongs to $y = -1$

The approach described above is able to perform a classification for *linearly separable classes*; to perform non-linear classification we should rely on the *Kernel trick*. We substitute the scalar product in the Dual lagrangian with a function $K(x_i, x_j)$. In that way we map implicitly our samples data into a higher dimensional space where the two classes are separated meaningfully by the best separating hyperplane.

In our specific case the kernel function is *Radial-basis function*; the value of that function depends on the distance between the two points in input.

$$K(x_i, x_j) = \exp(\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{2.11}$$

The hyperparameter σ controls to which degree two points should be considered close one to each other. Related to σ in the scikit-learn SVM model implementation we use γ as hyperparameter: high values of γ corresponds to a "strict" classification while low values of γ corresponds to a more general classification.

In our application we are going to test the *RBF Gaussian kernel*. For the RBF kernel model we are going to perform a grid search to select the best hyperparameters values (for C and γ).

Thinking in terms of *protein secondary structure prediction* each sample x^i will represent a residue in the protein sequence with its sliding windows context extracted from the sequence profile. In this application a window $w = 17$ is used. For each sample we end up having $20 \cdot 17 = 340$ features.

To deal with the *multi-class* classification problem we adopt the *one-vs-rest* approach. We train

for each of the three secondary structure conformations a classifier that gets as input samples belonging to the specific secondary structure conformation labeled as positives and all the other samples (belonging to the other conformations) labeled as negative.

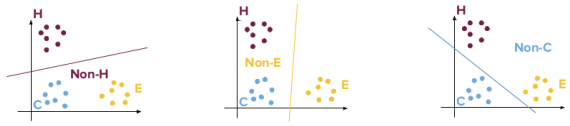

Figure 7: One-vs-rest approach

2.5 Evaluation of the prediction

The evaluation procedure has the goal to assess the quality of the predictions performed by the models that we are testing. To achieve this goal we are going to test the model on the training set using a *5-fold cross-validation* and then the evaluation is performed on a blind-test-set.

In the 5-fold cross-validation, the JPred dataset is split into 5 independent non-redundant subsets (provided with the JPred dataset it self). For each of the five cross-validation steps, we selected 4 of the 5 subsets as the training set and the other one as the test set. For each iteration the evaluation metrics are computed and at the end an average between the results is performed.

To perform the evaluation the following metrics are computed for each possible conformation. For each specific conformation by *positive* we refer to predictions belonging to that specific conformation and for negative we refer to predictions relative to the other secondary structure conformations.

- *Precision* : the percentage of correct positive predictions with respect to the number of positive predictions.

$$\text{PPV} = \frac{TP}{TP + FP} \tag{2.12}$$

- *Recall* : the percentage of correct positive predictions with respect to the number of actual positive examples.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{2.13}$$

- *Matthews correlation coefficient* (MCC) : balanced metric used to evaluate skewed classes. The values are in the range of $[-1, +1]$.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{2.14}$$

As general metric for the whole prediction accuracy we adopt the Q_3 metric that is computed as follows

$$Q_3 = \frac{TP_E + TP_H + TP_C}{N} \tag{2.15}$$

3 Results

The model has been tested using both the GOR and the SVM models; as expected non-linear SVM models obtains the best performance.

3.1 Cross Validation Test

As described in *Methods* on the training-set a 5-fold crossvalidation has been performed; during that phase a grid-search on SVM hyperparameters has been done. The best hyperparameters selected are *gamma* equal to 0.5 and C equal to 2. The complete gridsearch results can be found in the appendix of the article.

As we can see in *Table 2* the GOR method as expected reach an overall Q3 accuracy of 0.62. The *GOR* method is pretty stable for each iteration of the crossvalidation.

The SVM models finding the best separating hyperplane is better than the GOR method in modeling this complex inference. The amino acid propensity used by the GOR method is indeed an oversimplification that does not model the actual complexity of the relationship between the sequence profile context and the secondary structure conformation. Support vector machines model are able to take into account more information to compute the predictions.

The SVM model performances are reported in *Table 3* , the model selected by grid-search obtains good performances in terms of MCC on each conformation and an overall Q3 accuracy of *0.71*.

Table 2: Results of the GOR Method in 5-fold crossvalidation

Helix	CV1	CV2	CV3	CV4	CV5	Avg
MCC_H	0.52	0.53	0.52	0.54	0.52	0.52± 0.004
PPV_H	0.62	0.64	0.64	0.63	0.61	0.63± 0.004
TPR_H	0.82	0.80	0.79	0.82	0.81	0.81± 0.007
Strand	CV1	CV2	CV3	CV4	CV5	Avg
MCC_E	0.44	0.42	0.44	0.45	0.45	0.44± 0.006
PPV_E	0.50	0.45	0.47	0.49	0.51	0.49± 0.010
TPR_E	0.69	0.70	0.72	0.73	0.70	0.71± 0.008
Coil	CV1	CV2	CV3	CV4	CV5	Avg
PPV_C	0.41	0.42	0.42	0.42	0.42	0.42± 0.002
MCC_C	0.80	0.81	0.80	0.82	0.81	0.81± 0.004
TPR_C	0.42	0.43	0.44	0.41	0.43	0.43± 0.006
$Q3$	0.62	0.62	0.62	0.63	0.63	0.62± 0.002

Table 3: Results of the SVM model (rbf with gamma=0.5 and C=2) in 5-fold crossvalidation

Helix	CV1	CV2	CV3	CV4	CV5	Avg
MCC_H	0.64	0.64	0.70	0.66	0.68	0.66± 0.013
PPV_H	0.83	0.86	0.85	0.85	0.84	0.85± 0.006
TPR_H	0.67	0.67	0.65	0.70	0.72	0.69± 0.009
Strand	CV1	CV2	CV3	CV4	CV5	Avg
MCC_E	0.50	0.48	0.39	0.49	0.48	0.50± 0.023
PPV_E	0.80	0.76	0.81	0.81	0.80	0.79± 0.011
TPR_E	0.42	0.40	0.49	0.39	0.41	0.41± 0.020
Coil	CV1	CV2	CV3	CV4	CV5	Avg
MCC_C	0.48	0.49	0.49	0.49	0.50	0.49± 0.002
PPV_C	0.61	0.62	0.62	0.62	0.63	0.63± 0.002
TPR_C	0.87	0.88	0.88	0.88	0.88	0.88± 0.002
$Q3$	0.70	0.71	0.71	0.71	0.71	0.71± 0.002

3.2 Blind Set Test

The model has been tested on a unseen blind test set that has been generated as described in 2.2.

The blind set contains 150 proteins of which 17 do have an empty profile represented as the one hot encoding of the sequence. The category of sequences with empty profile have a low level of information.

The GOR method being less sensible to the features information has stable performance both on the full blindset and on the proteins without a meaningful sequence profile.
The SVM model instead is more sensible to the sequence profiles missing information, indeed the performance on this subset are not satisfactory obtaining a Q_3 equal to 0.38 .

As expected on the whole blind-set the SVM model (C=2, gamma=0.5) obtains the best performances in terms of MCC on all secondary structure conformations ($MCC_H = 0.61$, $MCC_E = 0.54$, $MCC_C = 0.47$) and also in the overall Q3 accuracy ($Q_3 = 0.69$).

All the results for the full blind-set are reported in *Table 4*, the performances on the sub-set with missing profiles are reported in *Table 5*.
The test performed on the subset containing only proteins with missing profiles (replace by one-hot encoding) is performed on 17 examples and it is approximative, having few elements leads to not-reliable metrics values.

Table 4: Performances on full blind-set. Comparison of GOR method with SVM model (gamma=0.5, C=2). N=150

Gor Method	Helix	Strand	Coil	Overall
MCC	0.48	0.42	0.40	
PPV	0.63	0.48	0.75	
TPR	0.76	0.69	0.43	
Q3				0.62
SVM Method	Helix	Strand	Coil	Overall
MCC	0.61	0.54	0.47	
TPR	0.61	0.47	0.57	
PPV	0.87	0.81	0.89	
Q3				0.69

Table 5: Performances on blind-set elements with empty profile. Comparison of GOR method with SVM model (gamma=0.5, C=2). N=17

Gor Method	Helix	Strand	Coil	Overall
MCC	0.51	0.40	0.39	
PPV	0.66	0.47	0.75	
TPR	0.82	0.66	0.38	
Q3				0.62

SVM Method	Helix	Strand	Coil	Overall
MCC	0.08	0.0	0.04	
TPR	0.88	0.0	0.38	
PPV	0.01	0.0	0.99	
Q3				0.38

4 Conclusion

The prediction of the secondary structure conformation can be done with a high accuracy, with the model based on SVM a Q_3 accuracy of 0.69 on the blind set is achieved. As expected the GOR method obtain lower overall Q_3 accuracy reaching 0.62 , however this model show a certain degree of robustness when dealing with empty profile proteins.

The SVM model strongly relies on the evolutionary information contained in the multiple sequence alignment and it is able to obtain satisfactory predictions in terms of Q_3 accuracy, MCC , precision and recall on the whole blind-set and on cross-validation.

We can consider the in-silico analysis of the secondary structure conformation satisfying enough to be used when experimental information is not available.

References

[1] P Y Chou and G D Fasman. Empirical predictions of protein conformation. *Annual Review of Biochemistry*, 47(1):251–276, 1978. PMID: 354496.

[2] J. Garnier, D.J. Osguthorpe, and B. Robson. 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, 1978.

[3] Alexey Drozdetskiy, Christian Cole, James Procter, and Geoffrey J. Barton. JPred4: a protein secondary structure prediction server. *Nucleic Acids Research*, 43(W1):W389–W394, 04 2015.

[4] John-Marc Chandonia, Naomi K Fox, and Steven E Brenner. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Research*, 47(D1):D475–D481, 11 2018.

[5] Stephen Altschul, T.L. Madden, A Shaffer, Jiezhang Zhang, and Zhengh Zhang. Gapped blast and psi-blast:a new generation of protein database search programs. *Nucl. Acids. Res.*, 25:3389–3402, 11 1996.

[6] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35, 10 2017.

[7] W Kabsch and C Sander. Dssp: definition of secondary structure of proteins given a set of 3d coordinates. *Biopolymers*, 22:2577–2637, 1983.