Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Protein annotation

A Support Vector Machine-based Method To Accurately Predict Protein Secondary Structure

Alessandro Lussana [1,*]

[1] M. Sc. Bioinformatics, Alma Mater Studiorum - University Of Bologna, Bologna, Italy

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The capability to accurately impute secondary structure starting from the amino acid sequence is fundamental to the functional annotation of proteins. As the amount of newly sequenced proteins coming from different organisms keeps rising we are enabled to learn the evolutionary variation of the protein domains, and to use this information to better understand the relationship between sequence and structure. It becomes a natural aim to exploit this knowledge for the development of new computational methods for secondary structure prediction.
Results: We developed a classification method that takes sequence context composition and evolutionary information to impute the secondary structure of proteins at the residue level. Our Support Vector Machine-based method was able to classify residues among helix, strand and coil conformations with a multi-class accuracy (Q_3) index above 75% and a Segment Overlap score close to 67% in a 5-fold cross-validation. Similar performance was observed on two blind test sets, outperforming a previous method, based on Bayesian statistics and information theory, and also exploiting evolutionary information.
Availability: `https://github.com/alussana/SVM-II-Str-classifier`
Contact: alessandro.lussana@protonmail.com
Supplementary information: (See Availability)

1 Introduction

Protein secondary structure is defined as the pattern of hydrogen-bonded and geometrical features that characterize segments of the protein chains, as observed from x-ray coordinates (Kabsch and Sander, 1983). Secondary structure is recognized as repeats of the elementary hydrogen-bonding patterns between the amino hydrogen and carboxyl oxygen atoms in the peptide backbone, that give arise to, for instance, α-helices and β-sheets. Geometrically, secondary structure can be defined in terms of torsion angles, that in the Ramachandran plot are observed to range within typical values depending on the secondary structure conformation (Ramachandran *et al.*, 1963). The reliable definition of their tri-dimensional structure is one of the major goals to achieve the functional annotation of proteins. Structural information can provide insights on protein function, drive drug design, help the inference of protein-protein interaction networks and the discovery of protein motifs. Due to the challenges of obtaining experimental molecular structures from X-ray crystallography or nuclear magnetic resonance experiments, for the very large majority of known proteins the corresponding structures have to be predicted starting from the coding sequence: in November 2019 there were still 158549 known protein structures deposited in the Protein Data Bank (Berman HM *et al.*, 2000), compared to over 180 million protein sequences stored in UniProtKB/TrEMBL database (UniProt Consortium, 2019). Homology modeling techniques are computational methods that have been proven to be very effective in imputing a molecular structure with high reliability in cases when some conditions holds, especially when the target protein shares a sequence identity greater than 30% with a template whose structure is known at high resolution. Nevertheless, when a protein's sequence identity with the known structures is too low, no direct homology can be detected with sequence alignment (Rost B,1999), making homology modeling unfeasible. In this case fold recognition methods are a solution to impute the structural features. These "threading" procedures require matching the target sequence with other proteins, known at a three-dimensional level and having a similar secondary structure (Rost, 1997). The rationale behind this is that a limited number of basic folds is found in nature and the amino acids

preferences for structural environment provide a sufficient information to choose the best-fitting fold. Therefore, the capability to predict a secondary structure starting from the amino acid sequence alone is a critical step in finding remote homologs to then achieve a reliable functional annotation. Moreover, secondary structure prediction is fundamental for *ab initio* 3D modeling methods. Previous studies to predict the secondary structure of proteins from their amino acid sequence have clearly shown the position-specific conformation of residues to be dependent from the sequence context composition: the first attempts to translate this knowledge into classification methods were based on the calculation of propensity scales, as a way to encode the biochemical and physical preferences of each amino acid for a specific secondary structure environment, depending on its distance from the assessed position (Chou and Fasman, 1974). This led to the development of the popular GOR method (Garnier *et al.*, 1978), based on Bayesian statistics and Information theory, which only in more recent years has been outperformed by exploiting evolutionary information of protein sequences and state-of-the-art, more computationally demanding, machine learning approaches (Ward *et al.*, 2003). Here we used a modern tool for homology search and the information available from public databases to generate examples and train a classifier based on Support Vector Machines (SVM), a modern and robust machine learning algorithm (Cortes and Vapnik, 1995), in order to address the problem of protein secondary structure prediction.

2 Approach

While sequence context composition has a heavy influence on the secondary structure conformation of protein residues, very different sequences are found in nature having very similar secondary structure, as in the case of distant homologs (Sander and Schneider, 1991), making the relationship between sequence and structure much less easy to be defined. Evolutionary events have produced wide variation in protein families and domains, still maintaining the conditions not to disrupt structures and functions, which are far more conserved. Therefore, considering only the sequence of the target protein allows to take into account a partial information only, namely a particular "instance" of the corresponding functional structure. We considered the sequence of a target protein as only one of the elements belonging to the set of all the possible amino acid chains occurring in nature having the same fold. In this way, we based our method for predicting the secondary structure on the information carried by families of sequences, expected to share a common fold, rather that on single amino acid chains. To do this, we exploited modern methods of database search to efficiently find homologs of representative protein sequences with a known molecular structure and to build sequence profiles corresponding to over a thousand of different and well-established protein domains. Profiles are statistical models of protein sequence consensus and are a way for detecting the conservation of patterns and motifs in protein families. Single-residue context composition profiles were used to train a classifier based on Support Vector Machines (SVM) for the imputation of secondary structure at residue level, among the three canonical conformations "helix", "strand", and "coil". Similarly, the imputation of an unknown structure is done generating a sequence profile via matching the sequence of the target protein with detectable homologs, and submitting the profile to the trained classifier.

3 Methods

3.1 Dataset

Training Dataset. We downloaded 1348 protein domains, defined by the Jpred4 secondary structure prediction server (Drozdetskiy *et al.*, 2015), consisting of their primary sequence in FASTA format and their secondary structure at single-residue resolution in DSSP format. This dataset was used by the Jpred authors for training the Jnet Naural Network-based predictor running at the Jpred server. It consists of non-redundant sequences chosen among the representatives of the SCOP super-families (Fox NK *et al.*, 2014). Details about the filtering procedure leading to the generation of this dataset are reported on the Jpred website (http://www.compbio.dundee.ac.uk/jpred/).

Starting from the Jpred-defined 1348 protein domains, we computed sequence profiles for each one of the amino acid sequences represented in the training set in order to estimate evolutionary conservation information at residue level. To obtain sequence profiles we run database searches against the UniProt/Swiss-Prot database (release of August, 2019) (UniProt Consortium, 2019), including 560537 sequences, using Psiblast program (version 2.7.1) (Altschul *et al.*, 1997). Sequence profiles in

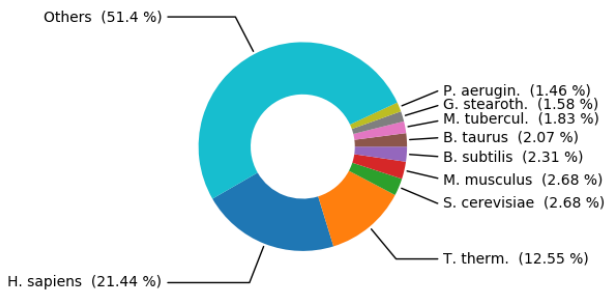

Fig. 2. Most represented organisms in the Training Dataset.

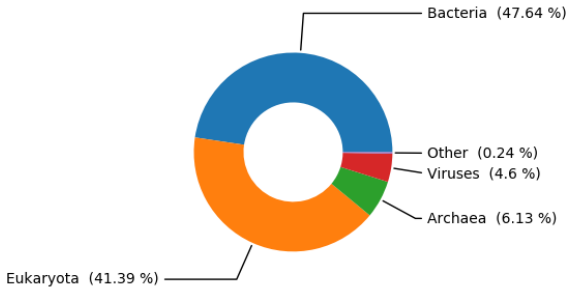

Fig. 1. Taxonomy classification of the Training Dataset

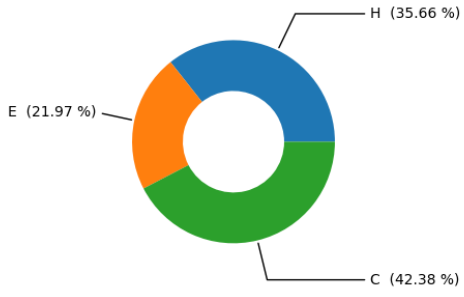

Fig. 3. Percentages of residues by conformational state in the Training Dataset.

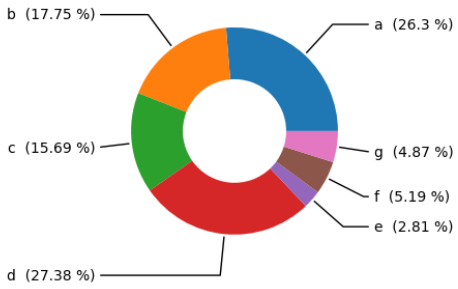

Fig. 4. Percentages of domains by SCOP class as found in the Training Dataset. a: All alpha proteins; b: All beta proteins; c: Alpha and beta proteins (a/b); d: Alpha and beta proteins (a+b); e: Multi-domain proteins (alpha and beta); f: Membrane and cell surface proteins and peptides, g: Small proteins.

PSSM format were successfully obtained for 1026 domains using an E-value threshold equal to 0.01, and three iterations of search rounds. These sequence profiles defined our training dataset. Summary statistics of this set for their composition in taxonomic groups and organism of origin are reported in Fig. 1 and Fig.2. A fairly good representation of all the three states of secondary structure conformations ("helix", "strand" and "coil": i.e. all other states) can be observed in the Training Set: no classes are greatly over- or under- represented, which is important in order to effectively train the classifier, and their relative abundance is observed as expected, with "coil" (C) being the most frequent conformational state, and "helix" (H) being more common than "strand" (E) (Fig. 3).

Analogously, we investigated the relative abundance of SCOP classes (Fox NK *et al.*, 2014) represented by the proteins in the Training Set. No proteins of classes i ('Low resolution protein structures'), j ('Peptides'), k ('Designed proteins'), and l ('Artifacts'), are found, while all other SCOP classes corresponding to "true folds" are present with the exception of class h ('Coiled coil') (Fig. 4).

In order to assess to which extent our data were a good representation of - and are concordant to - the current state of knowledge about residues conformation in proteins, we explored simple relationships between residue composition and secondary structure. In the histogram displayed in Fig. 5 the frequency of each one of the three conformational states we considered is reported by residue type. The most frequent amino acid to be found in coil regions is Glycine, while the most common in helices are Alanine, Leucine, and Glutamate, which, along with Glycine, are also among the most aboundant amino acids overall. β-strands are characterized by a great presence of Valine, followed by Lysine and Isoleucine.

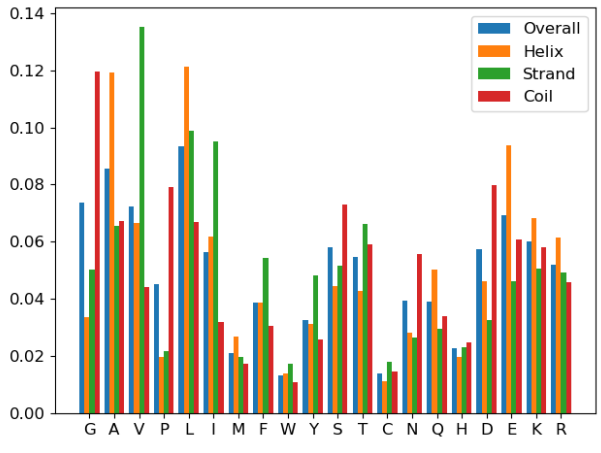

Fig. 5. Residue type frequency by DSSP class, as observed in the Training Dataset.

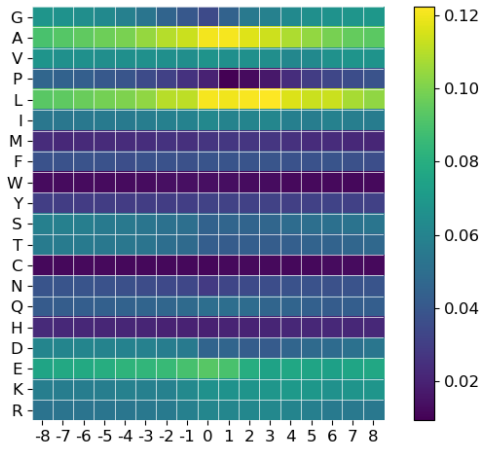

Fig. 6. Amino acids propensity for "helix" state. The position 0 on the horizontal axis marks a residue in "helix" conformation; positions from 1 to 8 and from -1 to -8 extend a window of nearby residues, upstream and downstream position 0. At each position, the relative frequency of each amino acid is color-coded from dark blue (less frequent) to bright yellow (more frequent).

We then extended our analysis to windows of residues to determine context amino acid composition in different secondary structure states. We considered sequence segments of 17 residues, centered on a position with a known secondary structure state, and computed the frequency of each residue type relatively to each position in the window when the central residue was in "helix" conformation (Fig. 6) or in "strand" conformation (Fig. 7). For "Helix", we observed an enrichment of Alanine, Leucine, and Aspartate, with a pattern showing a slowly degrading gradient in frequency from the center to the extremities of the window.

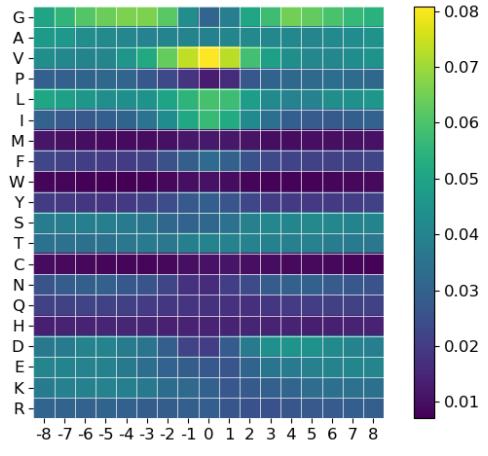

Fig. 7. Amino acids propensity for β-strand. The position 0 on the horizontal axis marks a residue in β-strand conformation; positions from 1 to 8 and from -1 to -8 extend a window of nearby residues, upstream and downstream position 0. At each position, the relative frequency of each amino acid is color-coded from dark blue (less frequent) to bright yellow (more frequent).

Analogously, the β-strand propensity map (Fig. 7) reveals that Valine is highly enriched in the positions displaying a β-strand conformation and in the immediate vicinity, with a frequency gradient rapidly fainting along the window. Interestingly, Glycine shows a peculiar frequency gradient behavior: its frequency follows an oscillation, being lowest at the center of the window, i.e. corresponding to the β-strand state, and highest at a distance of 3-4 positions on both C-terminal and N-terminal sides from the residue in "strand" conformation. This corresponds to Glycine being showily underrepresented in strands but enriched at their extremities, likely due to the minimal steric hindrance of its lateral side chain that allows the amino acid to fit this specific structural environment. In concordance to this, the distance between the two frequency peaks along the window is roughly of 6 positions, that is in fact the most common length for strands (Penel *et al.*, 2003). Finally, Leucine and Isoleucine show a positive propensity towards the strand conformation, while Aspartate shows a negative propensity.

Cross-Validation. We used the Training Dataset described above to perform a 5-fold cross-validation of the secondary structure prediction methods assessed here. Subsets of equal size were created randomly from the non-filtered Jpred-defined dataset including 1348 sequences (see Section 3.1, Training Dataset). We then filtered the subsets according to the non-null sequence profiles availability, as reported in the definition of the Training Dataset (see Section 3.1, Training Dataset). The final five folds include 211, 204, 207, 199, and 205 proteins respectively (Fig. 8). Since homology and internal redundancy of this dataset was minimized by construction (Drozdetskiy *et al.*, 2015), the five subsets are expected to be as heterogeneous as possible. This is important for the cross-validation procedure to be reasonably unbiased.

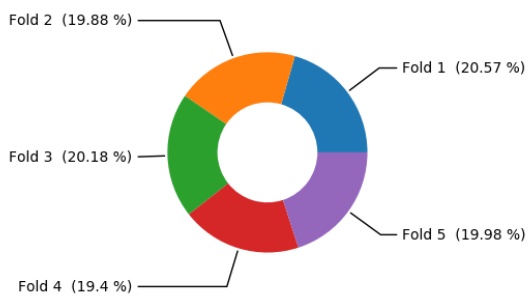

Fig. 8. Subdivision of the Training Dataset into disjoint subsets for performing cross-validation. Percentages with respect to the whole sets are shown for Fold 1 (211), Fold 2 (204), Fold 3 (207), Fold 4 (199), and Fold 5 (205).

Blind Test Datasets. We tested the performance of the Support Vector Machine-based classifier, trained on the complete Training Dataset with optimized hyper-parameters (see Methods), on two blind test sets consisting of sets of proteins completely disjoint from the training set, and that, therefore, were not used in the training phase of any of the model's parameters or hyper-parameters. We retrieved the non-redundant Blind Test Datasets from the Protein Data Bank (Berman HM *et al.*, 2000) applying the filtering procedure described as follows:

First, we selected the PDB-defined representative structures at 30% sequence identity, applying the following criteria: absence of Expression Tag, structure released between 2015-01-01 and 2019-09-12, presence of a protein chain in the structure, resolution of the crystal between 0.0 Å and 2.5 Å, and presence of at least a molecular sequence with length comprised between 50 and 300 residues. This resulted in 2392 entries that included a total of 5679 chains. We retained amino acid chains only and filtered out the single chains whose length was outside the 50-300 interval. 4759 chains passed this filter. We used the program Blastclust (version 2.2.26) from the stand-alone BlastP Suite (Wei *et al.*,2012) to group together chains sharing a sequence identity greater than 30%, removing internal redundancy. We chose only the representative sequences, defined as the set comprising the longest sequence of each cluster; 1885 protein chains were selected from this procedure. We then checked for external redundancy comparing this set against the Training Dataset (see Methods). We used BlastP program (version 2.7.1) to run a database search with each one of the 1885 sequences defined above against the Training Dataset, using an E-value threshold equal to 1. All the chains for which at least one significant match with a sequence identity $\geq 30\%$ was detected were discarded, reducing the blind test set to 942 chains.

We proceeded to download the pdb files to obtain atomic coordinates of the proteins in the blind test set, and we computed the secondary structure in DSSP format using the software Mkdssp (version 2.2.1) (Kabsch *et al.*, 1983). Since atomic coordinates of these structures not always cover all the residues as reported in the canonical sequences, we further filtered the set to exclude sequences for which a secondary structure could not be imputed on a continuous, non-gapped fragment of length between 50 and 300 amino acids, ending up with 605 proteins.

Finally, we computed sequence profiles for the blind test set proteins using Psiblast program (version 2.7.1) (Altschul *et al.*, 1997) to perform a database search on the UniProtKB/Swiss-Prot database (release of August, 2019) (UniProt Consortium, 2019), including 560537 sequences. Sequence profiles were successfully generated for 543 proteins using an E-value threshold equal to 0.01, and three iterations of search rounds. From these set we randomly selected 150 sequences twice, that eventually defined the two final Blind Test Datasets.

3.2 GOR Implementation

The Garnier-Osguthorpe-Robson method (GOR) (Garnier *et al.*, 1978), is a popular Information theory-based method for secondary structure prediction, developed shortly after the Chou-Fasman method. It imputes secondary structure states at residue level among the "helix" (H), "strand" (E), and "Coil" (C) classes, essentially using a Bayesian analysis.

In order to evaluate our SVM-classifier, detailed below, in comparison to a simpler but well-established method, we implemented a grammar-free GOR classifier in Python language using the Numpy Python library (Oliphant, 2006), and we trained it using the same Training Dataset as we did for the SVM-based classifier. We assessed its classification performance using the same cross-validation procedure and Blind Test Datasets.

GOR Training. As in the original GOR method (Garnier *et al.*, 1978), we used the Training Dataset (see Section 3.1, Training Dataset) to compute three different 17×20 matrices, from now on called propensity matrices, for "helix", "strand", and "coil" conformations, respectively. The propensity matrices describe the information carried by each one of the 20 proteinogenic amino acids, from the 17 positions of a window of consecutive residues counted from -8 to +8, about the conformational state of the residue at position 0, i.e. the central residue of the window. The values stored in the propensity matrices are log-odds scores, which are functions of the probability of finding a given secondary structure state in the central position of the window when a given residue type at a given position of the 17-residue segment is found. The definition of the log-odds score is the following:

$$I(S; R_{i,j}) = \log \frac{P(S|R_{i,j})}{P(S)}$$

where i is the position in the window, ranging from -8 to $+8$, j is one of the twenty amino acids, S is the conformation class of the residue at position 0,

and can be "helix" (H), "strand" (E), or "coil" (C), R is the residue of type j at the position i, $P(S|R_{i,j})$ is the conditional probability of observing S given $R_{i,j}$ in the training data, and $P(S)$ is the marginal probability of observing the conformation S in the training data. Assuming that the sequence context composition has an influence on the conformational state of a residue, $I(S; R_{i,j})$ is a measure of that influence derived from an Information theory framework, evaluating to which extent observing the amino acid of type j at position i is informative about the probability of the residue at position 0 to be or not in conformation S.

The actual parameters estimated from the training procedure are from a simple rearrangement of the equation above: considering the Bayes theorem, being $P(S|R_{i,j}) = P(S, R_{i,j})$ / $P(R_{i,j})$, then

$$I(S; R_{i,j}) = \log \frac{P(S, R_{i,j})}{P(S)P(R_{i,j})}$$

where $P(R_{i,j}, S)$ is the joint probability of observing the residue of type j at position i and the residue at position 0 in conformation S, and $P(R_{i,j})$ is the marginal probability of observing the residue of type j at position i.

GOR Testing. Prediction of secondary structure is carried out starting from the Position-specific Scoring Matrix (PSSM) of the target protein sequence, derived from three iterations of a PSI-BLAST search (see Section 3.1). A PSSM is in the form of a $M \times 20$ matrix, being M the sequence length. The secondary structure class at each position is assigned evaluating a "local" $W \times 20$ PSSM, defining a window centered on the target position, being W the window length and $W = 17$. In cases where the window extends beyond the protein termini, "empty" attributes were filled with zeros. The GOR method uses the following information function to assign the predicted secondary structure class (S^*) to the central position of the input sequence window:

$$S^* = \arg\max_S \left(\sum_{j=1}^{20} \sum_{i=-8}^{+8} I(S; R_{i,j}) \times M_{i,j} \right)$$

Where $M_{i,j}$ is the frequency of residue type j at the position i of the sequence profile window M. That is, given a 17-residues profile window the GOR method searches for the conformation class S that gives the maximum information score, computed as the sum of individual single-position information functions, relating residues in the window with the central-residue conformation, weighted for the frequency by which each amino acid at each position occurs in the profile window. These computations are done under the simplifying assumption of the statistical independence of the residues observed in different positions of the window, i.e.

$$P(R_{-8,j}, R_{-7,j}, ..., R_{7,j}, R_{8,j}) = \prod_{i=-8}^{+8} P(R_{i,j})$$

3.3 SVM Implementation

To date, secondary structure prediction has been tackled using different machine learning algorithms, including multi-layer perceptrons and recurrent neural networks (McGuffin and Jones, 2003). Support Vector Machines (SVMs) have shown promising results on several biological pattern classification problems too. For example, they have been successfully applied to recognition of protein translation–initiation sites in DNA sequences (Zien *et al.*, 2000) and functional annotation of genes from expression profiles (Brown *et al.*, 2000).

SVMs perform well compared with other learning algorithms and they are effective in controlling the classifier's capacity along with the associated potential for overfitting. This can be achieved by ensuring that the hyperplane pinpointing the decision boundary separating two classes does so with a large margin. Moreover, SVMs have other desirable properties such as relatively few adjustable parameters and the interchangeable use of kernel functions, which define a mapping of the input vectors to a higher-dimensional feature space (Ward *et al.*, 2003).

We built an SVM-based classifier for protein secondary structure prediction using the svm Python module from the Scikit-Learn library, version 0.22.0 (Pedregosa *et al.*, 2011), and we benchmarked it against the GOR secondary structure prediction method described in Section 3.1.

SVM Training. Position-Specific Scoring Matrices (PSSM) defined in the Training Dataset appear, for each protein, in the form of a $M \times 20$ matrix built from three iterations of a PSI-BLAST search (see Section 3.1, Training Dataset), where M is the length of the protein sequence. From these data the input vectors for training the SVMs were derived according to the following procedure: for each residue belonging to the training data, a new, smaller PSSM, representing the local context profile composition in a window centered on that residue, was considered. The $W \times 20$ "local" PSSMs, being W the window length and $W = 17$, were used in this way as the input to the SVMs. In cases where the window extends beyond the protein termini, "empty" attributes were filled with zeros. 168435 non-null input vectors associated to the respective secondary structure classes were used in training our models. The SVM-based classifiers we trained are composed by three different binary SVMs, each of which is responsible to perform a discrimination between two classes: Helix vs Strand, Strand vs Coil, and Helix vs Coil. Their outputs are integrated to impute the predicted class for the evaluated vector according to the "one-against-one" approach. The "one-against-one" method construct classifiers for the $\binom{2}{k} = k(k-1)/2$ possible class pairings, with each classifier trained on the subset of the examples belonging to the two classes. Different SVM-based classifiers were trained and evaluated in order to optimize the choice of hyperparameters via cross-validation. All the Support Vector Machines we built use the Radial Basis Function kernel (RBF) and an exhaustive search was performed to choose the better-performing combination of hyperparameters. The RBF kernel is defined as:

$$K(x, x') = \exp\left(- \frac{||x - x'||^2}{2\sigma^2} \right)$$

Where x and x' are two feature vectors of the same space, $||x - x'||^2$ is the definition of the squared Euclidean distance between the two feature vectors, and σ is a free parameter. The same kernel can be written in the form:

$$K(x, x') = \exp\left(- \gamma ||x - x'||^2 \right)$$

being gamma (γ) one of the two hyperparameters we adjusted during training via cross-validation.

The other tunable hyperparameter is the "trade-off" parameter C. The training algorithm involved in SVM-based classification requires solving a constrained optimization problem of the function:

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

being ξ_i "slack" variables that have to be minimized. The implementation with Scikit-learn of our SVM-based classifiers relies on LIBSVM (Chang and Lin, 2011), a popular open source machine learning library for solving the quadratic programming (QP) problem that arises during the training of support-vector machines. The quadratic program solver at the core of LIBSVM is used to maximize a Dual Lagrangian function:

$$L(\alpha_i) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j)$$

subject to:

$$\sum_i y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C \quad \forall i$$

in order to find the optimal decision boundary in the form of the hyperplane separating two classes, dependent from the vector w:

$$w = \sum_i y_i x_i \alpha_i$$

Given that, and the fact that minimizing $||w||$ implies maximizing the margin between the classes in the feature space, C is in fact recognizable as a variable controlling the trade-off between the classification error and the margins when learning the decision boundary. We adjusted the hyper-parameters via trial-and-error during cross-validation, evaluating the classifiers with all the possible combinations for the values of C and γ, being $C = 1.0$ or $C = 2.0$ or $C = 4.0$, and $\gamma = 0.3$ or $\gamma = 0.5$ or $\gamma = 2.0$.

SVM Testing. Prediction of secondary structure is carried out starting from the Position-specific Scoring Matrix (PSSM) of the target protein sequence, derived from three iterations of a PSI-BLAST search (see Section 3.1). Analogously to the classification with GOR method, the secondary structure class at each position of the target protein is assigned evaluating a "local" $W \times 20$ PSSM, defining a window centered on the target position, being W the window length and W = 17. The $W \times 20$ PSSM windows define the input vectors of our classifiers. In cases where the window extends beyond the protein termini, "empty" attributes were filled with zeros.

Imputation of secondary structure class by the SVM-based classifiers is done according to the "one-against-one" approach: the outputs from the three SVMs are combined by casting a vote for the 'winner' of each pair-wise comparison (Helix vs Strand, Strand vs Coil, and Helix vs Coil) and assigning the input vector to the class with the most votes, with Coil class used as tie breaker. The class assignment from the binary SVMs for any input vector x is done computing the following decision function:

$$f(x) = \text{sign}\left(\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b \right)$$

where SV is the set of Support Vectors, x_i is the ith Support Vector, and y_i is the class of the ith Support Vector.

3.4 Mesurement of performance

The secondary structure predictions are compared with DSSP (Kabsch and Sander, 1983) assignments of secondary structure derived from x-ray coordinates. Although DSSP defines eight different structural elements, these eight states are commonly translated into three secondary structure states: α-helix, β-sheet and coil. This translation is usually performed, as we did, in the following manner:

- H (α-helix), G (3-helix (3_{10} helix)), I (5 helix (π-helix)) $\implies$ H "helix"
- B (residue in isolated β-bridge), E (extended strand, participates in β ladder) $\implies$ E "strand"
- T (hydrogen bonded turn), S (bend), "" (unassigned) $\implies$ C "coil"

For both the GOR method and the SVM-based classifiers we assessed the secondary structure prediction performance in a 5-fold cross-validation and on two Blind, never-seen-before Datasets (see Section 3.1). In each test we evaluated the performance of the methods computing the canonical scoring measures, which include:

- Positive Predictive Value (Precision):

$$PPV = \frac{TP}{TP + FP} \times 100$$

- True Positive Rate (Recall):

$$TPR = \frac{TP}{TP + FN} \times 100$$

- Matthews correlation coefficient:

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

- Three-class Accuracy:

$$Q3 = \frac{TP_H + TP_E + TP_C}{N} \times 100 = \frac{T}{N} \times 100$$

where: TP = "True Positives"; TN = "True Negative"; FP = "False Positives"; FN = "False Negatives"; T is the total number of True Positives of any class; N is the total number of observations of any class. To perform these computations we used the functions defined in the Scikit-learn python package, version 0.22.0 (Pedregosa *et al.*, 2011).

Segment Overlap Score. Along with the canonical scoring indexes, we evaluated the classifiers' performance computing the Segment OVerlap score (SOV). The SOV is an assessment index to evaluate the meaningfulness of predicted protein secondary structures using biology-oriented criteria. The SOV score is based on secondary structure segments rather than individual residues and was originally developed as a problem-specific metric in order to assess the biological relevance of a secondary structure imputation in a more reliable way than Q3 (Zemla *et al.*, 1999). We implemented the computation of the SOV in-house, using the Numpy Pyhton library (Oliphant, 2006).

4 Results

We developed a protein secondary structure prediction method that takes sequence context composition and evolutionary information to impute the secondary structure conformation at residue level. Our Support Vector Machine-based predictor was able to classify residues among the "helix", "strand", and "coil" classes with a multi-class accuracy score (Q3) above 75% in a 5-fold cross-validation. Similar performance was observed on two blind test datasets, and in all cases outperforming the popular,

Table 1. Cross-Validation Performance

Cross-Validation		GOR	SVM
PPV	H	65.01 ± 0.42	82.25 ± 0.42
	E	50.04 ± 0.57	76.64 ± 0.36
	C	79.72 ± 0.18	70.53 ± 0.17
TPR	H	79.96 ± 0.28	80.24 ± 0.36
	E	69.76 ± 0.29	53.90 ± 0.50
	C	48.11 ± 0.22	82.87 ± 0.11
MCC	H	0.541 ± 0.002	0.71 ± 0.003
	E	0.451 ± 0.003	0.565 ± 0.003
	C	0.443 ± 0.001	0.567 ± 0.002
SOV	H	67.11 ± 0.28	73.00 ± 0.20
	E	63.57 ± 0.45	53.07 ± 0.63
	C	48.74 ± 0.36	68.60 ± 0.24
Q_3		64.21 ± 0.09	75.56 ± 0.18
SOV_3		60.01 ± 0.16	66.94 ± 0.15

Information theory-based GOR method, also exploiting evolutionary information. We determined, via trial-and-error during cross-validation, the best-performing hyperparameters (see Section 3.3) for the SVM-based classifier to be $\gamma = 0.3$ and $C = 2.0$. We reported the performance, according to the metrics described in Section 3.4, of the GOR-based and the SVM-based secondary structure predictors both in cross-validation (Table 1) and in one of the blind test datasets (Table 2). Both tests on blind datasets gave similar performances (see Supplementary information), which are completely comparable to that observed in cross-validation, also indicating a very low redundancy and high quality of the Training Dataset, and a very scarce bias in the cross-validation procedure. Being the training data completely equivalent for the GOR and the SVM-based methods, our results clearly show the benefits of an advanced machine learning approach with respect to a simpler strategy based on Bayesian analysis and Information theory in addressing this classification problem. Very importantly, a substantial improvement was observed also when assessing the Segment Overlap score (SOV), a biology-oriented metric (see Section 3.4) to evaluate the goodness of secondary structure predictions, indicating the SVM-based predictors to be responsible of more meaningful secondary structure assignments with respect to the GOR method.

Table 2. Blind Test Performance

Blind Test Dataset #1		GOR	SVM
PPV	H	66.05	84.28
	E	55.14	79.92
	C	73.20	65.55
TPR	H	74.60	72.89
	E	71.71	61.11
	C	50.30	83.56
MCC	H	0.530	0.683
	E	0.474	0.612
	C	0.422	0.538
SOV	H	60.37 ± 2.15	62.53 ± 2.32
	E	69.60 ± 1.63	63.21 ± 2.23
	C	53.17 ± 1.13	69.24 ± 1.13
Q_3		64.27	74.09
SOV_3		62.90 ± 1.00	67.32 ± 1.24

5 Discussion

We successfully trained and tested a Support Vector Machine-based predictor for protein secondary structure that exploits sequence context composition and sequences evolutionary information, and that outperformed the popular, Information theory-based GOR method when evaluated with the canonical classification performance metrics and a problem-specific assessment index. The implementation of an advance machine learning method such as the Support Vector Machine (SVM) clearly proved to be very effective in this task. Arguably, the main advantage here resides in the capability not to assume the statistical independence of the residues observed in different positions of the input profile windows, as it is done in the GOR method, and to catch those synergies between positions in the profile that have a weight in pattern learning. This is possible for the SVM with the remapping of the feature vectors to a higher-dimensional space through the kernel function. Nevertheless, despite the achievement of a multi-class accuracy similar to that reached by complex, state-of-the-art methods like PSIPRED (Jones DT, 1999) -, some issues are worth to be considered. The number of support vectors for the final predictor is extremely high, being 40451, 30996, and 55704 for the three binary SVMs at the core of the final classifier, compared to 140989 total training examples. In general, this is undesirable because the fraction of the training examples that become support vectors places an upper bound on the capability of the predictor to be general enough to perform well on new data. It also affects the time complexity, which scales linearly with the number of support vectors, in using the classifier to make new predictions. Causes for a high number of support vectors include the noise in the evolutionary profiles and some ambiguity in the structure assignments. This high error leads to a large fraction of the data set becoming bounded support vectors (Ward *et al.*, 2003). Despite this, a good performance was achieved on both a 5-fold cross-validation, performed on a carefully selected and non-redundant dataset, and on two blind test datasets, indicating no evident signs of overfitting.

Both the GOR and our SVM-based methods impute secondary structure at single residue level relying on local context information. While the strong effect of local context on the secondary structure is well established, the influences of more distant portions of the protein can also play a role. For instance, segments far away in the primary sequence, and whose composition seem to show a high propensity for a given secondary structure class, may be located in close spatial proximity due to the folding process and reciprocally influence their secondary structure conformation towards a different class. This can be favoured by the presence of non-local effects, such as hydrogen bonds or disulfur bridges, that may allow for a lower free energy in a local context of residues having a high propensity for another conformation (Russel and Barton, 1993). Since longer range interactions are not encoded by local windows, this places an upper limit in secondary structure prediction accuracy for methods accounting for the local context only.

Another issue that affects the assessment of secondary structure prediction is associated with conformational variation observed at secondary structure segment ends. Even for homologous protein pairs elements of secondary structure frequently differ in the exact position of their ends. This issue may decrease the performance of the methods, but it is at least partially taken into consideration when evaluating the prediction performance with the Segment Overlap score (SOV) (see Section 3.4). Moreover, the existence of protein-chameleons, and the ambiguity in the position of segment ends due to differences of approach in secondary structure classification (Zemla *et al.*, 1999) may also be relevant in further decreasing the maximum performance.

Possible improvements for the secondary structure prediction approach presented here include embedding the SVM-based predictor in a larger framework that takes in input the classifier's results to further process them according to the present knowledge of the properties of α-helices and β-strands. This would allow to refine the secondary structure assignments based on local context information introducing a grammar for secondary structure segments, for instance disallowing β-strands shorter than two residues or α-helices shorter than three residues. This improvement could in particular affect the SOV performance and may be implemented despite, as it is evident with "one against one" approach for multi-class classification, the acknowledged deficiency of SVMs which do not provide estimates of the posterior probability of class membership, in contrast to many neural networks.

A final refinement of the predictor may consist in improving the quality of the training examples, namely the Position-Specific Scoring Matrices (PSSM), in terms of reduction of noise and enhancement of the information content of the sequence profiles. Possibly, Hidden Markov Model homology searches (Karplus *et al.*, 1998), which recover information from more remote homologous sequences (Rost B, 2001), can be used to replace PSIBLAST for the PSSM generation both in training and in testing phases.

6 Conclusion

In this article we presented a Support Vector Machine-based approach to impute protein secondary structure starting from the primary sequence, and benchmarked it against the popular GOR method. We showed that the implementation of an advanced machine learning framework allowed to achieve a sharply higher performance with respect to a simpler bayesian approach based on Information theory in this context-dependent problem. We reached high multi-class accuracy and Segment Overlap score ($> 66\%$) and we discussed the possible improvements that may be implemented to further enhance the performance. These refinements include a post-processing of the predictions in order to include a grammar of secondary structure segments, and the possibility to generate better training example using Hidden Markov Model-based profile generation.

Funding

This work was supported by the Bologna Biocomputing Group (`http://www.biocomp.unibo.it/`) that provided the computational infrastructure.

References

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

Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963;7:95-9.

Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42.

UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506-D515.

Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974;13(2):222-45.

Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12(2):85-94.

Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol. 1997;270(3):471-80.

Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991;9(1):56-68.

Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015;43(W1):W389-94.

Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.

Fox NK, Brenner SE, Chandonia JM. 2014. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research 42:D304-309. doi: 10.1093/nar/gkt1240

Wei D, Jiang Q, Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics. 2012;13:174.

Penel S, Morrison RG, Dobson PD, Mortishire-smith RJ, Doig AJ. Length preferences and periodicity in beta-strands. Antiparallel edge beta-sheets are more likely to finish in non-hydrogen bonded rings. Protein Eng. 2003;16(12):957-61.

Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol. 1978;120(1):97-120.

Mcguffin LJ, Jones DT. Benchmarking secondary structure prediction for fold recognition. Proteins. 2003;52(2):166-75.

Zien A, Rätsch G, Mika S, Schölkopf B, Lengauer T, Müller KR. Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics. 2000;16(9):799-807.

Brown MP, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA. 2000;97(1):262-7.

Ward JJ, Mcguffin LJ, Buxton BF, Jones DT. Secondary structure prediction with support vector machines. Bioinformatics. 2003;19(13):1650-5.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

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

Travis E. Oliphant. A guide to NumPy, USA: Trelgol Publishing, (2006).

Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins. 1999;34(2):220-3.

Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292(2):195-202.

Russell RB, Barton GJ. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J Mol Biol. 1993;234(4):951-7.

Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics. 1998;14(10):846-56.

Rost B. Review: protein secondary structure prediction continues to rise. J Struct Biol. 2001;134(2-3):204-18.

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

Cortes C, Vapnik V. Mach Learn (1995) 20: 273.