BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–5

Predicting the subcellular location of eukaryotic proteins with support vector machines

Ramon Viñas Torné [1]*

[1]University College London. MSc Machine Learning. COMPGI10 Bioinformatics coursework.

23 March 2018

ABSTRACT

Motivation: Within the last few years the complete sequence has been determined for over 3000 genomes. Predicting the function of a protein has proved to be a difficult task where no clear homology to proteins of known function exists. Knowing the subcellular location of such proteins might be a crucial feature to determine their function.

Methods: This work presents an approach for predicting the subcellular location (nuclear, mitochondrial, cytosolic or secreted) of non-homologous proteins. This method extracts several N-terminal and global features, and performs classification using a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel.

Results: Results show that the SVM classifier is able to effectively exploit the engineered features, achieving an accuracy of 68% on an independent test set consisting of 1845 non-homologous amino acid sequences.

Availability: The source code is freely available online on https://github.com/rvinas/predicting_subcellular_location

Contact: ramon.torne.17@ucl.ac.uk

1 INTRODUCTION

Eukaryotic cells have evolved ways to partition off different functions to various locations in the cell. Different organelles play different roles in the cell, and proteins within them perform a vast number of functions that enable the well-functioning of the organism, including DNA replication, catalyzing metabolic reactions and transporting molecules from one location to another. Understanding the role of proteins based on their amino acid residues has centered the effort of the scientific community for several years, and determining where these proteins reside in the cell might be a critical milestone towards this goal.

Numerous efforts have been made to develop methods for predicting protein subcellular location based on the amino acid sequence information. TargetP (Emanuelsson O *et al*., 2000) assigns protein location with a feed-forward neural network, and it is based on the predicted presence of specific N-terminal presequences: chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP) or secretory pathway signal peptide (SP). WoLF PSORT (Horton P *et al*., 2007) converts protein amino acid sequences into numerical localization features based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs, and uses a K-nearest neighbor classifier for prediction. Several other methods exists, such as ProLoc-GO (Huang W-L *et al*., 2008) or KnowPredsite (Lin H-N *et al*., 2009), but they are less effective when dealing with non-homologous proteins because they rely on ontologies or knowledge graphs.

It is well known that most proteins in eukaryotic cells are synthesized in the cytosol, and many need to be further sorted to one or other cellular organelles. This process usually relies on specific signal peptides that are located in the N-terminal, such as the mitochondrial targeting peptide (mtTP) or the secretory pathway signal peptide (SP). These signals are recognized by a translocation machinery, and the proteins are delivered to the corresponding organelle. Once the proteins arrives at the destination, the signal peptide is typically removed by a signal peptidase. In this study, several global and local features are extracted from the amino acid sequence with the aim of capturing these patterns.

This work focuses on the use of Support Vector Machines (SVMs) to predict subcellular location. SVMs perform binary classification by (1) efficiently projecting the input features into a custom feature space through a kernel, and (2) finding an hyperplane that separates the classes with a large margin. In addition, SVMs have been widely used in bioinformatics because (1) a good kernel choice allows to exploit non-linear patterns effectively, and (2) they are able to control overfitting through a regularization hyperparameter that trades off the misclassification of training examples against the simplicity of the decision boundary. Moreover, SVMs are well studied and theoretically grounded, and several bounds on the test error rate exist (Bartlett P and Shawe-Taylor J, 1999; Vapnik V and Chapelle O, 2000).

2 METHODS

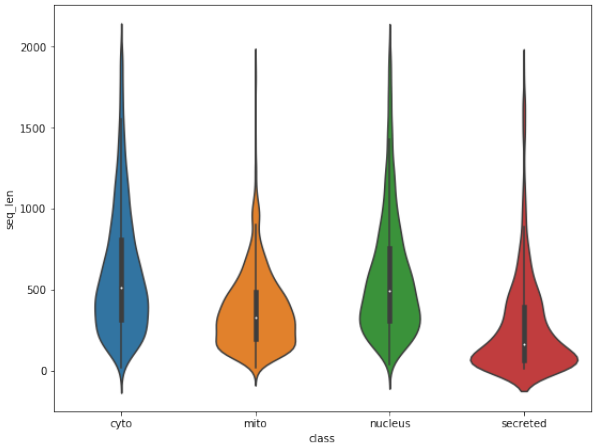
2.1 Data and preprocessing

The data consisted of 9222 amino acid sequences labeled with the subcellular location (cytosolic, nuclear, mitochondrial or secreted), encoded in FASTA format. In addition, a set of 20 unlabeled sequences was provided to check the generalization performance in hindsight, and every sequence in these sets can be assumed to be non-homologous. The labeled data (9222 examples) was split via stratified random sampling into a train and a test sets consisting of 7377 and 1845 examples, respectively.

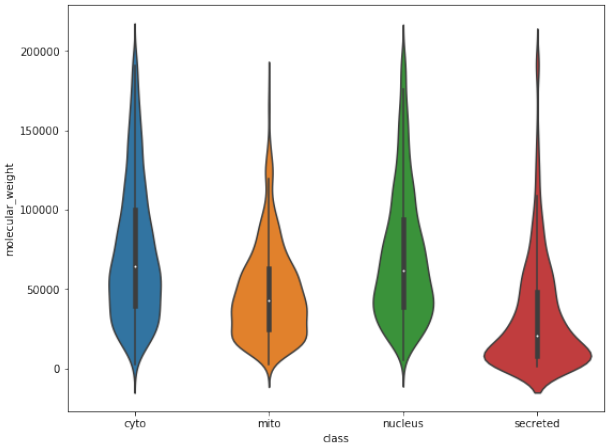
The sequences do not contain gaps nor translation stops, but they occasionally present the code *X* (unknown amino acid code in FASTA format). Whenever it was possible, this code was handled by assigning a uniform uncertainty to each of the 20 amino acids, or ignored otherwise. Similarly, codes *B* (aspartate or asparagine) and *Z* (glutamate or glutamine) were treated with uncertainty over

*to whom correspondence should be addressed

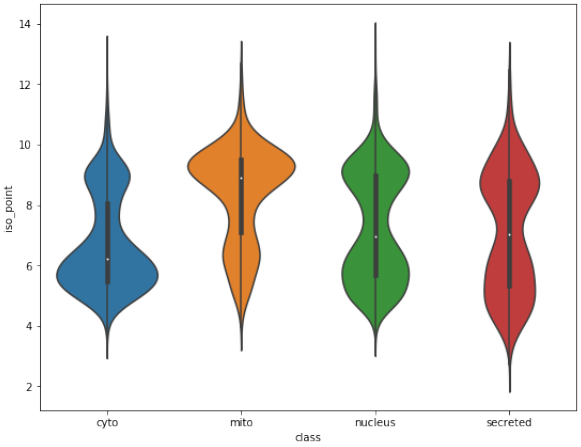
© Oxford University Press 2005.



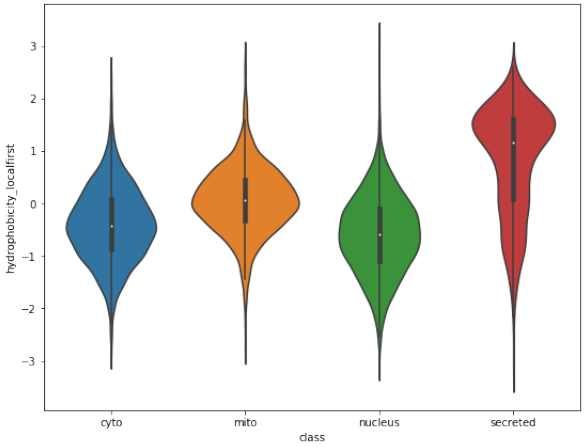
(a) Distribution of the sequence length for each subcellular location.



(b) Distribution of the molecular weight for each subcellular location.



(c) Distribution of the isoelectric point (Bjellqvist B. *et al*., 1993) for each subcellular location.



(d) Distribution of the index of hydrophobicity (Kyte and Doolittle, 1982) for each subcellular location (first 20 amino acids).

Fig. 1: Discriminative power of some engineered features by themselves.

the residues that they represent, but they were arbitrarily replaced by aspartate and glutamate, respectively, whenever this was not possible. The code *U* (selenocysteine) was directly replaced by *C* (cysteine) due to the high structural similarity between the residues that they represent.

2.2 Feature engineering

The input features were engineered to capture some informative properties of the amino acid sequences, using the publicly available Biopython software (Cock *et al*, 2009). This set of features include (1) the sequence length; (2^*) the molecular weight; (3^*) the relative frequency of each aminoacid in the chain; (4) the isoelectric point (Bjellqvist B. *et al*., 1993); (5) the aromaticity (Lobry JR. *et al*., 1994); (6) the instability index (Guruprasad K. *et al*., 1990); (7^*) the index of hydrophobicity (Kyte and Doolittle, 1982); (8^*) the index of hydrophilicity (Hopp and Woods, 1981); (9) the flexibility scores (Vihinen M. *et al*, 1994); and (10) the secondary structure scores[1]. Some of these features were also constructed from the local amino acid composition at the N-terminal[2], to account for specific targeting signal patterns that occur at the extremities of some amino acid sequences. For example, it is known that secretory proteins contain a signal sequence of 6 to 12 amino acids with hydrophobic side chains at the N-terminal (Figure 1d). Figure 1 shows the discriminative power of some engineered features by themselves.

[1] Relative frequency of amino acids that tend to be in α-helices (V, I, Y, F, W, L), β-sheets (E, M, A, L) or turns (N, P, G, S).

[2] These features are highlighted with '*'.

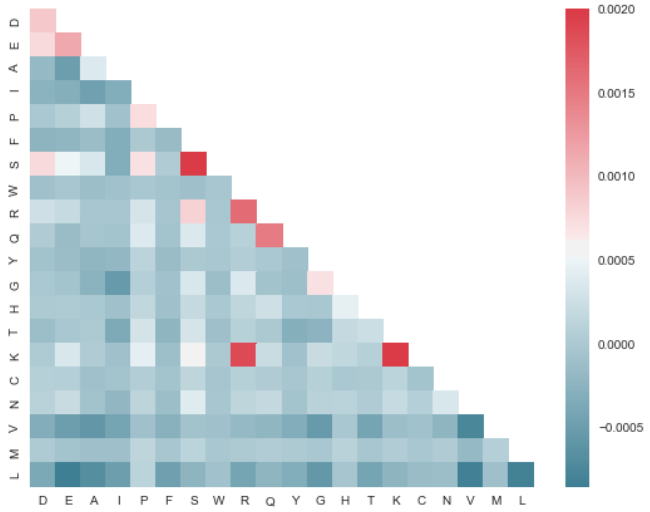


Fig. 2: Amino acid 2-gram correlation in nuclear proteins relative to cytosolic proteins. This plot shows that the frequency of the 2-gram KR (lysine and arginine, positively charged amino acids) is much higher in nuclear proteins than in cytosolic proteins. This informative feature is highly tied to the Nuclear Localization Signal, consisting of a sequence of positively charged amino acids that tag a protein to be moved into the nucleus.

Further analysis was carried out by visual inspection with the aim of finding characteristic patterns of each subcellular location. In particular, a special focus was given to cytosolic and nuclear proteins, because these classes are relatively similar in terms of the univariate feature distributions that describe them. A simple analysis of the relative frequency of the 2-gram amino acid composition (Figure 2) shows that there exist some specific patterns in nuclear proteins that make them distinguishable from cytosolic proteins. Concretely, the frequency of 2-grams involving positively charged amino acids (particularly lysine and arginine) was found to be highly correlated with nuclear proteins. These concrete patterns are likely related to the Nuclear Localization Signal (NLS), a signal consisting of a subsequence of positively charged amino acids that tags a protein to be moved from the cytoplasm to the nucleus through the nuclear pores (Lange A. *et al*., 2007). These targeting signals are recognized by importin, and they differ from protein to protein. Features that loosely capture the presence of a NLS, such as (11) the relative frequency of 2-grams involving lysine and/or arginine, were also added to the set of input features.

The set of input features consisted of 80 distinct features. Further processing was required to ensure that all the features are within a reasonable range. For example, the molecular weight of some sequences may exceed 200000 KDa. To this end, features were normalized by substracting the sample mean and dividing by the standard deviation.

2.3 Model design

The model design was based on a Support Vector Machine (SVM) (Cortes C and Vapnik V, 1995). The publicly available *Scikit-learn* software was used to train the SVM (Pedregosa F *et al*, 2011). A different SVM was trained for each pair of subcellular locations via a 'one-versus-one' approach. This yielded $k(k-1)/2$ binary classifiers (where $k = 4$ is the number of classes), that were trained on the subset of examples belonging to each pair of subcellular locations. Training these SVMs involves the minimization of the following optimization problem for each $j \in [1, k(k-1)/2]$:

$$\frac{1}{2}\mathbf{w}_j^\top \mathbf{w}_j + C\sum_{i=1}^{N_j}\xi_{ij} \tag{1}$$

subject to the constraints:

$$y_i(\mathbf{w}_j^\top \phi(\mathbf{x}_i) + b_j) \geq 1 - \xi_{ij},\ \xi_{ij} \geq 0,\ i \in [1, N_j] \tag{2}$$

where $\mathbf{w}_j$ and b_j are the paramters of the *j*-th SVM; N_j is the number of examples belonging to the *j*-th class pair; C is a regularization hyperparameter that trades off the misclassification of training examples against the simplicity of the decision boundary; $\mathbf{x}_i$ and y_i are the input vectors of features and the binary class label of the i-th pair-wise example, respectively; ξ_{ij} are the slack variables that handle non-separable data; and $\phi(\cdot)$ is the feature map function that allows to enrich the input set of features, ideally by combining them in a non-linear manner.

At prediction time, the outputs were produced through a voting scheme, where each pair-wise classifier casts a vote for one class, and the class with the most votes is selected. One drawback of Support Vector Machines is that they do not provide probability estimates for the posterior probability of class memberships by their own nature. This difficulty was overcome using the pair-wise coupling method proposed by Wu *et al* (2004), which extends Platt scaling (John C. Platt, 1999) to produce probability estimates for the multiclass case.

The performance of several kernel functions was investigated via stratified 4-fold cross-validation, and the Radial Basis Function (RBF) kernel was found to have the lowest validation error:

$$K(\mathbf{x}, \mathbf{t}) = \langle \phi(\mathbf{x}), \phi(\mathbf{t}) \rangle = exp(-\gamma||\mathbf{x} - \mathbf{t}||)^2 \tag{3}$$

where $\mathbf{x}$ and $\mathbf{t}$ are the vectors of input features for two examples, and γ is a regularization parameter that controls how much influence a single training example has.

3 RESULTS AND DISCUSSION

Subcellular location	f1-score	AUC
Cytosolic	0.60	0.80
Mitochondrial	0.71	0.94
Nuclear	0.66	0.84
Secreted	0.88	0.97

Table 1. The prediction performance of the SVM for each subcellular location. This table was created from results obtained using the independent test set. The system obtains a remarkable performance for secreted proteins, but experiences more difficulties to distinguish cytosolic and nuclear proteins.

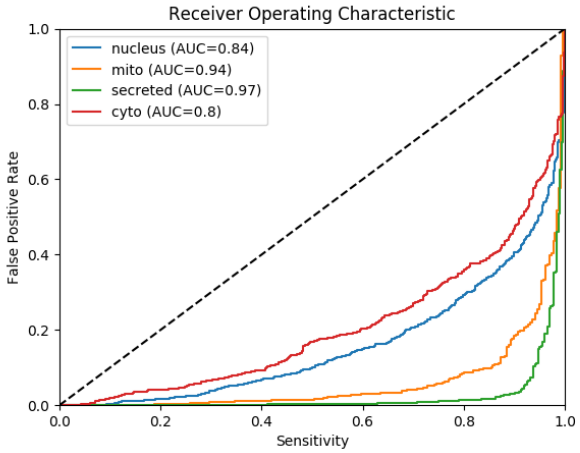


Fig. 3: The predictive performance shown as sensitivity versus false positive rate for each subcellular location. This plot was created from results obtained using the independent test set. The dotted line corresponds to the random performance.

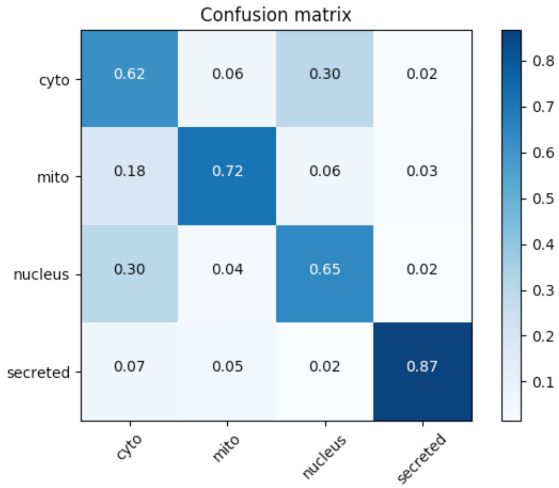


Fig. 4: Error analysis of the SVM shown as a normalized confusion matrix. The plot was constructed from results obtained using the independent test set. The true category and the predicted category are represented in the y-axis and x-axis, respectively.

The final classifier included six 'one-versus-one' SVMs with a RBF kernel. The model was trained on a set with 7377 sequences, and tested on a set with 1845 sequences, split randomly in a stratified manner. The hyperparameters of the model ($C = 40$, $\gamma = 0.1$) were determined using stratified 4-fold cross-validation. The overall accuracy of the final predictor on the test set was 68%.

Table 1 summarizes the performance of the 'one-versus-one' SVM in terms of f1-score and Area Under the Curve (AUC). Results show that secreted proteins are more easily distinguished from the rest. This is the behavior that one might expect given the set of input features, because secretory proteins need to be excreted through the cell membrane, and therefore they tend to be shorter and lighter than the rest (Figures 1a and 1b). In addition, secretory proteins can be usually recognized by a signal sequence in the N-terminal consisting of 6 to 12 amino acids with hydrophobic side chains (Figure 1d).

Figure 3 shows the prediction performance of the SVM in terms of sensitivity and the level of false positives. When the sensitivity is below 40%, the false positive rate is reasonably low for all the classes. In particular, the system excels in detecting secreted proteins, achieving a sensitivity of 90% with a false positive rate below 10%. A normalized confusion matrix is shown in Figure 4. In general, it is difficult to distinguish mitochondrial, nuclear and secreted proteins from cytosolic proteins, mainly because most eukaryotic proteins are synthesized in the cytosol. The method is particularly prone to confuse nuclear proteins by cytosolic proteins (when the true subcellular location is the nucleus, the SVM predicts cytosol 30% of the time), potentially because the set of input features is not rich enough. Attempting to detect the Nuclear Localization Signal (NLS) is probably an interesting approach (Lange A. *et al.*, 2007). Figure 2 shows the discriminative power of certain 2-gram frequencies of positively charged amino acids in order to distinguish nuclear proteins from cytosolic proteins.

Finally, table 2 shows the predictions of the proposed method for the 20 blind sequences.

ID	Subcellular location	Confidence
SEQ677	cytosol	0.48
SEQ231	cytosol	0.48
SEQ871	mitochondrion	0.70
SEQ388	cytosol	0.50
SEQ122	nucleus	0.47
SEQ758	nucleus	0.68
SEQ333	nucleus	0.50
SEQ937	cytosol	0.78
SEQ351	cytosol	0.65
SEQ202	cytosol	0.33
SEQ608	cytosol	0.46
SEQ402	nucleus	0.45
SEQ433	secreted	0.98
SEQ821	secreted	0.85
SEQ322	nucleus	0.77
SEQ982	nucleus	0.89
SEQ951	cytosol	0.47
SEQ173	cytosol	0.71
SEQ862	cytosol	0.45
SEQ224	cytosol	0.72

Table 2. Predictions for the 20 blind sequences. In general, the method is quite confident when it predicts the underrepresented classes (mitochondrion and secreted), but the uncertainty is usually bigger when the predicted class is cytosol.

4 CONCLUSION AND FUTURE WORK

The aim of this work was to provide a method for predicting the subcellular location (nucleus, cytosol, mitochondrion or secretory) of eukaryotic proteins based on their amino acid sequence. The method presented in this work has demonstrated a decent generalization performance on novel sequences, and the eukaryotic

protein properties captured by the proposed set of features have been demonstrated to be informative of the subcellular location. Support Vector Machines with a Radial Basis Function have been shown to be an effective approach for predicting the subcellular location, mainly because of its ability to recognize nonlinear patterns. A 'one-versus-one' strategy was successfully adopted to combine multiple binary classifiers, and estimates of the posterior probability of class membership were effectively produced using a pair-wise coupling technique proposed by Wu *et al* (2004).

This work has given a special focus to distinguishing nuclear from cytosolic proteins, since most features (such as the sequence length or the molecular weight) are often not discriminative enough. In particular, this study introduced a method to detect general patterns that relate to the Nuclear Localization Signal via 2-gram frequency features of positively charged amino acids. This set of features has significatively increased the ability of the model to discriminate nuclear and cytosolic proteins, but they are often not specific enough. Further approaches may benefit from attempting to capture richer patterns, potentially by matching profiles of signaling sequences or searching for known protein localization signals in manually curated databases such as LocSigDB (Negi S *et al*, 2015).

REFERENCES

Bartlett, Peter and Shawe-Taylor, John (1999). Generalization Performance of Support Vector Machines and Other Pattern Classifiers. *Advances in Kernel Methods. MIT Press*: 43-54.

Bjellqvist, Bengt and Hughes, Graham J. and Pasquali, Christian and Paquet, Nicole and Ravier, Florence and Sanchez, Jean-Charles and Frutiger, Séverine and Hochstrasser, Denis (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences, *Wiley Subscription Services, Inc., A Wiley Company*, **14.1**: 1023-1031.

Cock, Peter J. A. and Antao, Tiago and Chang, Jeffrey T. and Chapman, Brad A. and Cox, Cymon J. and Dalke, Andrew and Friedberg, Iddo and Hamelryck, Thomas and Kauff, Frank and Wilczynski, Bartek and de Hoon, Michiel J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, **25.11**: 1422-1423.

Cortes, Corinna and Vapnik, Vladimir (1995). Support-vector networks, *Machine Learning*, **20.3**: 273-297.

Emanuelsson Olof and Nielsen Henrik and Brunak Søren and von Heijne Gunnar (2000). Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence, *Journal of Molecular Biology*, **300.4**: 1005-1016.

Guruprasad K, Reddy BV, Pandit MW. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence, *Protein Engineering*, **4.2**: 155-161.

Hopp TP, Woods KR (1981). Prediction of protein antigenic determinants from amino acid sequences, *Proceedings of the National Academy of Sciences of the United States of America*, **78.6**: 3824-3828.

Horton, Paul and Park, Keun-Joon and Obayashi, Takeshi and Fujita, Naoya and Harada, Hajime and Adams-Collier, C.J. and Nakai, Kenta (2007). WoLF PSORT: protein localization predictor, *Nucleic Acids Research*, **35**: W585-W587.

Huang W-L, Tung C-W, Ho S-W, Hwang S-F, Ho S-Y (2008). ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics*. **9.80**.

Kyte Jack, Doolittle F. Russell (1982). A simple method for displaying the hydropathic character of a protein, *Journal of Molecular Biology*, **157.1**: 105-132.

Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH (2007). Classical Nuclear Localization Signals: Definition, Function, and Interaction with Importin , *The Journal of biological chemistry*, **282.8**: 5101–5105.

Lin H-N, Chen C-T, Sung T-Y, Ho S-Y, Hsu W-L (2009). Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinformatics*, **10**(Suppl 15): S8.

Lobry JR, Gautier C (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes, *Nucleic Acids Research*, **22.15**: 3174-3180.

Negi S, Pandey S, Srinivasan SM, Mohammed A, Guda C (2015). LocSigDB: a database of protein localization signals, *Database: The Journal of Biological Databases and Curation*: bav003.

Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python, *Machine Learning*, **12**: 2825-2830.

Platt John C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, *Advances in large margin classifiers. MIT Press*: 61-74.

Vapnik V and Chapelle O (2000). Bounds on Error Expectation for Support Vector Machines. *Neural Computation*. **12.9**: 2013-2036.

Vihinen M, Torkkila E, Riikonen P (1994). Accuracy of protein flexibility predictions, *Proteins*, **19**: 141–149.

Wu, Ting-Fan and Lin, Chih-Jen and Weng, Ruby C. (2004). Probability Estimates for Multi-class Classification by Pairwise Coupling, *The Journal of Machine Learning Research*, **5**: 975-1005.