A multi-class support vector machine for pattern classification problems in structural biology

Florent Dufay[1], Yann Guermeur[1], Nicolas Sapay[2], and Thérèse Malliavin[3]

[1]LORIA, CNRS UMR 7503
[2]Métropole de Lyon
[3]LPCT, CNRS UMR 7019

1 Introduction

Many problems in structural biology take the form of a pattern classification problem characterized as follows: given a protein sequence, assign to each of its residues a category belonging to a given finite set (identified with $[\![1; C]\!]$ the set of integers ranging from 1 to C). A famous example is provided by the protein secondary structure prediction problem. This three up to eight category classification problem was admittedly solved by Magnan and Baldi (2014). To tackle this kind of task, a hierarchical and modular method was introduced by Guermeur et al. (2004b). This architecture was then developed by Sapay et al. (2006), Guermeur and Lauer (2016) and Guermeur and Malliavin (2024), as it was applied to different problems. The first layer of treatment in the hierarchy combines Multi-Layer Perceptrons (MLPs) (Anthony and Bartlett, 1999) and Multi-class Support Vector Machines (M-SVMs) (Guermeur, 2012; Doğan et al., 2016; Lei et al., 2019) whose outputs are post-processed with a Polytomous Logistic Regression (PLR) model.

This poster deals with the dedication of one of the M-SVMs to the biological problem of interest and its assessment. This is obtained through kernel engineering (Shawe-Taylor and Cristianini, 2004) and model selection (Massart, 2007). The touchstone is provided by two open problems: the prediction of the ω torsion angles in globular proteins and the prediction of amphipathic in-plane membrane anchors in monotopic membrane proteins.

The organization of the paper is as follows. Section 2 describes the M-SVM and its kernel. Section 3 is devoted to the prediction of the ω torsion angles. Section 4 deals with the prediction of the amphipathic in-plane membrane anchors. At last, we draw conclusions in Section 5.

2 Dedicated M-SVM

Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) are binary classifiers obtained as kernelized extensions of the maximum margin hyperplane. The M-SVMs are their multi-class extensions. The model considered here is the initial one, introduced by Weston and Watkins (1998). In order to describe our implementation of this machine, we must first introduce the theoretical context.

We make the hypothesis that the covariates $\mathbf{x}$ live in a domain $\mathcal{X}$. As stated in introduction, the set of categories $\mathcal{Y}$ is identified with the set of indexes of the categories, i.e., $[\![1, C]\!]$ (no structure is assumed on $\mathcal{Y}$). Let $d_m = \{(x_i, y_i) : 1 \leqslant i \leqslant m\}$ be a set of m labelled examples ($d_m \in (\mathcal{X} \times \mathcal{Y})^m$). Let κ be a real-valued positive type function/kernel (Berlinet and Thomas-Agnan, 2004) on $\mathcal{X}^2$ and let $\left(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}\right)$ be its reproducing kernel Hilbert space (RKHS). The architecture considered to perform the discriminant analysis, i.e., the function class $\mathcal{H}$ on which function selection is performed based on d_m, is the set of functions $h = (h_k)_{1 \leqslant k \leqslant C}$ from $\mathcal{X}$ into $\mathbb{R}^C$ given by:

$$\forall \mathbf{x} \in \mathcal{X}, \ h(\mathbf{x}) = \left(\langle \bar{h}_k, \kappa_{\mathbf{x}} \rangle_{\mathbf{H}_\kappa} + b_k\right)_{1 \leqslant k \leqslant C},$$

where $\left(\bar{h}_k\right)_{1 \leqslant k \leqslant C} \in \mathbf{H}_\kappa^C$ and $(b_k)_{1 \leqslant k \leqslant C} \in \mathbb{R}^C$. The corresponding learning problem, a quadratic programming problem, is the following one.

Problem 1

$$\min_{h \in \mathcal{H}, \ \xi \in \mathbb{R}^{Cm}} \left\{ \frac{1}{2} \sum_{k=1}^{C} \|\bar{h}_k\|_{\mathbf{H}_\kappa}^2 + \sum_{i=1}^{m} \sum_{k=1}^{C} C^{(y_i)} \xi_{ik} \right\}$$

$$\textit{subject to:} \begin{cases} \forall i \in [\![1, m]\!], \ \forall k \in [\![1, C]\!], \ \ h_{y_i}(x_i) - h_k(x_i) \geqslant 1 - \delta_{y_i,k} - \xi_{ik} \\ \forall i \in [\![1, m]\!], \ \forall k \in [\![1, C]\!], \ \ \xi_{ik} \geqslant 0 \end{cases} .$$

The difference between this formulation and the original one rests in the use of one soft margin parameter $C^{(k)}$ per category. This idea is borrowed from Ben-Hur and Weston (2010).

The basic principle of the dedication of the machine to the problems of interest is the implementation of a local prediction, involving an analysis window sliding on the protein sequence. Thus, the description $\mathbf{x}$ associated with a residue in a sequence is the window content when the window in centered on the residue. As a consequence, if the index of the residue in the sequence is i and the window size is $2\ell + 1$, then

$$\mathbf{x} = (x_{i+k})_{-\ell \leqslant k \leqslant \ell},$$

where x_i is the amino acid / residue to be classified and the x_{i+k} for $k \neq 0$ are the neighboring amino acids (or empty positions at the N and C termini of the sequence). In the sequel, with a slight abuse of notation, we simply denote $\mathbf{x} = (x_k)_{-\ell \leqslant k \leqslant \ell}$ a description (ignoring when possible the relative position of the corresponding residue in the data sets). With this notation at hand, the kernel $\kappa_{G,\mu,\boldsymbol{\theta}}$ applied to these descriptions, introduced by Guermeur et al. (2004a), is defined as follows:

$$\kappa_{G,\mu,\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\mu \sum_{k=-\ell}^{\ell} \theta_k \left\|x_k - x'_k\right\|_G^2\right) \qquad (1)$$

where $\mu \in \mathbb{R}_+^*$ and $\boldsymbol{\theta} = (\theta_k)_{-\ell \leqslant k \leqslant \ell} \in \mathbb{R}_+^{2\ell+1}$. Formula (1) involves the norm $\|\cdot\|_G$ on the set of amino acids (endowed with a structure of vector space). This norm is the one associated with the Gram matrix G derived from a substitution matrix. The transition from the substitution matrix to the matrix G is the projection operator on the convex cone of symmetric semi-definite matrices. The weighting vector $\boldsymbol{\theta}$ is introduced to take into account the relative importance (for the prediction) of the different positions in the window. As such, the value of one of its components can be set arbitrarily, so that the values of the other ones are derived accordingly. In practice, we chose to set $\theta_0 = 1$. The transition from the relative importance to the absolute one is provided by the "bandwidth" parameter μ.

Problem 1 and Formula (1) define the hyperparameters whose values are to be selected in order to dedicate the M-SVM to the problem of interest. In practice, the model selection procedure depends on the biological problem. This is illustrated in the two following sections.

3 Prediction of the ω torsion angles

We start with a problem whose importance to determine the 3D structure has recently been highlighted by da Rocha et al. (2024).

3.1 Biological problem

Machine learning models, especially the deep learning ones, proved capable of predicting protein structures from the primary sequence (Varadi et al., 2023). However, their capacity to predict statistically significant long-range proximities frequently rests on the availability of a sufficiently large alignment of protein sequences (Radjasandirane and de Brevern, 2023). Unfortunately, the alignments can be too small for small families of proteins (Radjasandirane and de Brevern, 2023), or can be irrelevant in the case of intrinsically disordered regions. This motivates the study of protein structure inference from local geometry information only.

Initial results have shown that variations in stereochemistry must be taken into account to obtain a relevant prediction (da Rocha et al., 2024). In particular, variations in torsion angles ω are essential. Taking our inspiration from (Guermeur and Malliavin, 2024), we address the prediction of these variations as a four-category discrimination problem with the categories being:

$$\begin{cases} \delta\omega \leqslant -3^\circ : \text{class } 1 \\ -3^\circ < \delta\omega \leqslant 0^\circ : \text{class } 2 \\ 0^\circ < \delta\omega \leqslant 4^\circ : \text{class } 3 \\ 4^\circ < \delta\omega : \text{class } 4 \end{cases}$$

with $\delta\omega = sign(\omega) \cdot 180^\circ - \omega$.

3.2 Dedication of the machine

The literature provides many kernel engineering methods (Shawe-Taylor and Cristianini, 2004). We chose to set the values of the hyperparameters G and $\boldsymbol{\theta}$ of the kernel by means of the centered kernel target alignment (Cortes et al., 2012). The initial feasible solution for the matrix G is the projection of the BLOSUM62 matrix (Henikoff and Henikoff, 1989). As for the weighting vector $\boldsymbol{\theta}$, all its components are initialized at 1. At last, the values of the bandwidth μ and the soft margin parameters $C^{(k)}$ are obtained as the result of a search on a grid, through cross-validation.

3.3 Experimental results

The M-SVM and a MLP were applied to the HIV-protease data set. This set contains 176 protein sequences from which 16645 15-residue peptides were extracted (the window size is 15).

A seven-fold cross-validation was implemented, involving a subset of the training set dedicated to model selection (validation set). The quality of the prediction is measured by means of three standard criteria: the recognition rate, the Matthews (1975) correlation coefficients and the cross-entropy (CE). The Matthews' correlation coefficients characterize the quality of the prediction per category. As for the cross-entropy, it provides us with an idea of the feasibility of the post-processing by the higher modules of the hierarchical architecture, especially the Inhomogeneous Hidden Markov Model (IHMM) (Guermeur et al., 2004b). The results obtained are gathered in Table 1.

Classifier	% rec.	C_1	C_2	C_3	C_4	CE
WW-M-SVM + PLR	66.3	0.66	0.29	0.17	0.66	1.04
MLP	64.5	0.67	0.28	0.14	0.64	1.70

Table 1: Relative performance of the M-SVM and a MLP for the prediction of the ω torsion angles.

According to the two sample proportion test, the superiority of the M-SVM over the MLP is statistically significant with confidence exceeding 0.95.

4 Prediction of amphipathic in-plane membrane anchors in monotopic proteins

With this second problem, we switch from globular proteins to membrane ones.

4.1 Biological problem

Many prediction methods have been devised to detect transmembrane segments with a α-helical conformation in membrane proteins. However, membrane proteins can also be monotopic, i.e., bound to the membrane interface and thus in contact with only one of the compartments defined by the membrane. The goal here is to detect the membrane anchor parallel to the membrane plane, so-called in-plane membrane (IPM) anchor, of these proteins. This is thus a binary pattern classification problem. It is noteworthy that the two categories are highly unbalanced, with only 7.7% the residues belonging to an IPM anchor.

4.2 Dedication of the machine

This dedication exhibits one single difference with the former one: the choice of the substitution matrix. In line with what was done by Sapay et al. (2006), the PHAT matrix (Ng et al., 2000) is preferred to the BLOSUM62 matrix.

4.3 Experimental results

The M-SVM and a MLP were applied to the AmphipaSeeK data set. This set contains 30 experimentally characterized protein sequences made up of 12105 residues. Here again, the window size was set to 15, and a seven-fold cross-validation was implemented, involving a subset of the training set dedicated to model selection. The quality of the prediction is measured with the same criteria as previously. The results obtained are gathered in Table 2.

Classifier	% rec.	C_{IPM}	CE
WW-M-SVM + PLR	92.0	0.38	1.06
MLP	88.5	0.28	1.79

Table 2: Relative performance of the M-SVM and a MLP for the prediction of IPM anchors.

The use of two soft margin parameters proved decisive to establish the superiority of the M-SVM. Indeed, without this degree of freedom, its Matthews' correlation coefficient C_{IPM} would have been inferior to the one of the MLP. On average, the value of the soft margin coefficient associated with the category "IPM anchor" is ten times larger than the other one.

5 Conclusions and ongoing research

A multi-class support vector machine dedicated to pattern classification problems in structural biology has been introduced and assessed on two open problems of central importance. The experimental results establish its superiority over the standard neural network, the multi-layer perceptron.

Our ongoing research consists in applying to the two problems the whole hierarchical and modular method. Once this is done, a comparison will be made with the latest models of deep learning.

References

M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

A. Ben-Hur and J. Weston. A user's guide to support vector machines. *Methods in molecular biology (Clifton, N.J.)*, 609:223–39, 2010.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, 2004.

C. Cortes and V.N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

W. da Rocha, L. Liberti, A. Mucherino, and T. Malliavin. Influence of stereochemistry in a local approach for calculating protein conformations. *J Chem Inf Model*, 64(23): :8999–9008, 2024.

U. Doğan, T. Glasmachers, and C. Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research*, 17(45):1–32, 2016.

Y. Guermeur. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 6(6):555–577, 2012.

Y. Guermeur and F. Lauer. A generic approach to biological sequence segmentation problems: application to protein secondary structure prediction. In M. Elloumi, C.S. Iliopoulos, J.T.L. Wang, and A.Y. Zomaya, editors, *Pattern Recognition in Computational Molecular Biology: Techniques and Approaches*, chapter 7, pages 114–128. Wiley, 2016.

Y. Guermeur and T. Malliavin. Méthode hiérarchique hybride de prédiction des angles de torsion ω des protéines. In *SFC*, 2024.

Y. Guermeur, A. Lifchitz, and R. Vert. A Kernel for Protein Secondary Structure Prediction. In *Kernel Methods in Computational Biology*, Chap. 9 - ISBN 0-262-19509-7, pages 193–206. The MIT Press, Cambridge, Massachussetts, 2004a.

Y. Guermeur, G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy, and P. Baldi. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing*, 56:305–327, 2004b.

S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919, 1989.

Y. Lei, U. Doğan, D.-X. Zhou, and M. Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

C. Magnan and P. Baldi. Spro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. bioinformatics. *BMC Bioinformatics*, 30(18), 2014.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Springer-Verlag, Berlin, 2007.

B.W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442–451, 1975.

P.C. Ng, J.G. Henikoff, and S. Henikoff. Phat: a transmembrane-specific substitution matrix. *Bioinformatics*, 16(9):760–766, 2000.

R. Radjasandirane and A. de Brevern. Structural and dynamic differences between calreticulin mutants associated with essential thrombocythemia. *Biomolecules*, 13(3):509, 2023.

N. Sapay, Y. Guermeur, and G. Deléage. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. *BMC Bioinformatics*, 7(255), 2006.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, MA. Laydon, A. Žídek, H. Tomlinson, D. Hariharan, J. Abrahamson, T. Green, J. Jumper, E. Birney, M. Steinegger, D. Hassabis, and S. Velankar. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *J Chem Inf Model*, 52(D1):D368–D375, 2023.

J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.