

MSVMpred User's guide

October 23, 2025

Contents

1	Introduction	3
2	Architecture of the application	3
3	Data	3
3.1	Coding of the data	3
3.2	Data sets provided	4
4	Multi-class support vector machines	4
5	Neural networks	5
6	Tools	5

1 Introduction

This archive gathers the basic components of the MSVMpred application [3], devised to perform protein sequence segmentation tasks. MSVMpred implements a hierarchical approach including all the multi-class support vector machines (M-SVMs) of the literature [2] and neural networks as base classifiers. It has been applied to protein secondary structure prediction [6, 3], the prediction of amphipathic in-plane membrane anchors in monotopic proteins [9] and the prediction of the ω torsion angles in globular proteins [5]. The programs are written in C ANSI or Python, and thus can be used under the various releases of UNIX, Linux, IRIX, etc.

2 Architecture of the application

The architecture of the application is depicted in Figure 7.1 of [3]. The following table lists the modules available, with their location in the archive and date of last update.

Module	Location	Last update
<code>project_matrix</code>	<code>Tools</code>	June 20, 2025
WW-M-SVM	M-SVM/WW-M-SVM	August 12, 2025
MLP	NN/MLP	September 4, 2025
<code>tune_kernel</code>	<code>Tools</code>	September 27, 2025
PLR	NN/PLR	October 15, 2025
LLW-M-SVM	M-SVM/LLW-M-SVM	October 23, 2025

3 Data

The principle of the hierarchical approach is to start with a local prediction and gradually extend the context so as to exploit interactions which are distant in the sequence but close in the structure.

3.1 Coding of the data

For the base classifiers, the prediction is local, based on the content of an analysis window sliding on the primary structure. Precisely, the description associated with a residue to be classified corresponds to the content of the analysis window when it is centered on this residue. The coding of this

peptide is obtained by replacing each residue with the rank (starting from 1) of the corresponding amino acid in the alphabet used. Empty positions in the window, observed in the vicinity of the N and C termini, are associated with the value 0. If the data sets initially available use alternative codings, then the program `process_data` can be used to perform the appropriate changes. It is located in the directory `Tools`.

The files containing the data sets must be text files, with the three first lines corresponding respectively to the number of examples/residues, the size of the analysis window (number of predictors) and the number of categories. The following lines provide the examples (description + category). We illustrate this structure on the training set contained in the file `Data/IPM1.app`.

```
11342 ← number of residues in the protein sequences (number of ex-
amples)
15 ← size of the analysis window (number of components of the vector
describing an example)
2 ← number of categories
0 0 0 0 0 0 0 6 12 8 5 1 12 10 5 1 ← description of the first ex-
ample and label of its category (here 1)
0 0 0 0 0 0 6 12 8 5 1 12 10 5 9 1 ← description of the second
example and label of its category
...
```

3.2 Data sets provided

The files in the directory `Data` correspond to the training and test sets associated with three data sets used in the literature of predictive structural biology. The files `struct.app` and `struct.test` correspond to the set of 1096 globular proteins (with 268575 residues) assembled by G. Pollastri for protein secondary structure prediction [6]. The files `IPM*.app` and `IPM*.test` correspond to the set of 30 proteins gathered by N. Sapay to predict the in-plane membrane (IPM) anchors of monotopic proteins [9]. At last, the files `omega?.app` and `omega?.test` correspond to an HIV-protease data set provided by T. Malliavin to predict the ω torsion angles of globular proteins. This set contains 176 protein sequences made up of 16646 residues.

4 Multi-class support vector machines

The name of the directory containing the M-SVMs is `M-SVM`. Currently, only two such machines are available, the one introduced by Weston and Watkins

[10], in the subdirectory `WW-M-SVM`, and the one introduced by Lee, Lin and Wahba [8], in the subdirectory `LLW-M-SVM`.

5 Neural networks

The name of the directory containing the neural networks is `NN`. Currently, only two such networks are available: the polytomous logistic regression (PLR) [7] and the standard multi-layer perceptron (MLP) [1]. The PLR is not used as a base classifier, but as a post-processing for the outputs of the M-SVMs.

6 Tools

The mains tools available are those used to tune the parameters of the Gaussian kernel of the M-SVMs (see Formula 9.5 in [4]). They can be found in the directory `Tools`. The script `project_matrix` takes in input the substitution matrix `matrix.txt` and outputs the matrix `projected_matrix.txt` of dot product between amino acids. The program `tune_kernel` computes the weighting on the positions of the analysis window.

References

- [1] M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.
- [2] Y. Guermeur. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 6(6):555–577, 2012.
- [3] 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.
- [4] Y. Guermeur, A. Lifchitz, and R. Vert. A kernel for protein secondary structure prediction. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, chapter 9, pages 193–206. The MIT Press, Cambridge, MA, 2004.

- [5] Y. Guermeur and T. Malliavin. Méthode hiérarchique hybride de prédiction des angles de torsion ω des protéines. In *SFC*, 2024.
- [6] 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, 2004.
- [7] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, London, 1989.
- [8] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.
- [9] 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.
- [10] 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.