Estimating the Class Posterior Probabilities in Protein Secondary Structure Prediction

Yann Guermeur and Fabienne Thomarat

LORIA – Equipe ABC
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy Cedex, France
{Yann.Guermeur,Fabienne.Thomarat}@loria.fr

Abstract. Support vector machines, let them be bi-class or multi-class, have proved efficient for protein secondary structure prediction. They can be used either as sequence-to-structure classifier, structure-to-structure classifier, or both. Compared to the classifier most commonly found in the main prediction methods, the multi-layer perceptron, they exhibit one single drawback: their outputs are not class posterior probability estimates. This paper addresses the problem of post-processing the outputs of multi-class support vector machines used as sequence-to-structure classifiers with a structure-to-structure classifier estimating the class posterior probabilities. The aim of this comparative study is to obtain improved performance with respect to both criteria: prediction accuracy and quality of the estimates.

Keywords: protein secondary structure prediction, multi-class support vector machines, class membership probabilities.

1 Introduction

With the multiplication of genome sequencing projects, the gap between the number of known protein sequences and the number of experimentally determined protein (tertiary) structures is widening rapidly. This raises a central problem, since knowledge of the structure of a protein is a key in understanding its detailed function. The prediction of protein structure from amino acid sequence, i.e., *ab initio*, has thus become a hot topic in molecular biology. Due to its intrinsic difficulty, it is ordinarily tackled through a divide and conquer approach in which a critical first step is the prediction of the secondary structure, the local, regular structure defined by hydrogen bonds. Considered from the point of view of pattern recognition, this prediction is a three-category discrimination task consisting in assigning a conformational state α-helix, β-strand or aperiodic (coil), to each residue (amino acid) of a sequence.

For almost half a century, many methods have been developed for protein secondary structure prediction. Since the pioneering work of Qian and Sejnowski [1], state-of-the-art methods are machine learning ones [2,3,4,5]. Furthermore, a majority of them shares the original architecture implemented by Qian and

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 260–271, 2011.

Sejnowski. Two classifiers are used in cascade. The first one, named sequence-to-structure, takes in input the content of a window sliding on the sequence, or the coding of a multiple alignment, to produce an initial prediction. The second one, named structure-to-structure, takes in input the content of a second window sliding on the initial prediction. Making use of the fact that the conformational states of consecutive residues are correlated, it mainly acts as a filter, increasing the biological plausibility of the prediction. Until the end of the nineties, the classifiers at the basis of most of the prediction methods implementing the cascade treatment were neural networks [6], either feed-forward, like the multilayer perceptron (MLP) [1,2,4] or recurrent [3]. During the last decade, they were gradually replaced with bi-class support vector machines (SVMs) [7,5] and multi-class SVMs (M-SVMs) [8,9,10]. This resulted in a slight increase of the prediction accuracy. On the other and, an advantage of the neural networks over the SVMs rests in the fact that under mild hypotheses regarding the loss function and the activation function of the output units, they estimate the class posterior probabilities (see for instance [11,12]). This is a useful property in the framework of interest, for two main reasons. The first one is obvious: such estimates provide the most accurate reliability indices for the prediction (see [13,4] for indices based on them, and [7] for an index based on the outputs of bi-class SVMs). The second one is of higher importance, since it deals with the future of protein secondary structure prediction. It is commonly admitted that the main limiting factor for the prediction accuracy of any prediction method based on the standard cascade architecture is the fact that local information is not enough to specify utterly the structure. This limitation is only partly overcome by using recurrent neural networks. Several works [8,9,13] have considered a more ambitious alternative, consisting in the implementation of hybrid architectures combining discriminant models and hidden Markov models (HMMs) [14]. In short, the discriminant models are used to compute class posterior probability estimates from which the emission probabilities of the HMMs are derived, by application of Bayes' formula. This approach widens the context used for the prediction, and makes it possible to incorporate some pieces of information provided by the biologist, such as syntactic rules. It appears highly promising, although it still calls for significant developments in order to bear its fruits. It is this observation that motivated the present study. Our thesis is that in the framework of such hybrid architectures, an efficient implementation of the cascade architecture could result from using as sequence-to-structure classifiers M-SVMs endowed with a dedicated kernel, provided that the structure-to-structure classifier is chosen appropriately. In this article, we thus address the problem of identifying the optimal structure-to-structure classifier when the classifiers performing the sequence-to-structure prediction are dedicated M-SVMs, and the final outputs must be class membership probability estimates. In practice, we want to take benefit of the high recognition rate of the M-SVMs without suffering their drawback. Our contribution, the first of this kind, focuses on the use of tools from nonparametric statistics.

The organization of the paper is as follows. Section 2 proposes a general introduction to the M-SVMs. The four M-SVMs involved in the experiments are then characterized in this framework, and their implementation for sequence-to-structure classification is detailed. Section 3 is devoted to the description of the different models considered for the structure-to-structure prediction. Experimental results are gathered in Section 4. At last, we draw conclusions and outline our ongoing research in Section 5.

2 Multi-class Support Vector Machines

2.1 General Introduction

The theoretical framework of M-SVMs is the one of large margin multi-category classifiers [15]. Formally, it deals with Q-category pattern recognition problems with $3 \leqslant Q < +\infty$. Each object is represented by its description $x \in \mathcal{X}$ and the set $\mathcal{Y}$ of the categories y can be identified with the set of indices of the categories: $[\![1, Q]\!]$. The assignment of the descriptions to the categories is performed by means of a *classifier*, i.e., a function on $\mathcal{X}$ taking values in $\mathbb{R}^Q$. For such a function g, the corresponding *decision rule* f is defined as follows:

$$\forall x \in \mathcal{X}, \quad \begin{cases} \text{if } \exists k \in [\![1, Q]\!] : g_k(x) > \max_{l \neq k} g_l(x), \text{ then } f(x) = k \\ \text{else } f(x) = * \end{cases}$$

where $*$ denotes a dummy category introduced to deal with the cases of ex æquo. Like all the SVMs, all the M-SVMs published so far belong to the family of *kernel machines* [16,17], which implies that they operate on a class of functions induced by a positive type function/kernel [18]. For a given kernel, this class, hereafter denoted by $\mathcal{H}$, is the same for all the models (it only depends on the kernel and Q). In what follows, κ designates a real-valued kernel on $\mathcal{X}^2$ and $\left(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}\right)$ the corresponding reproducing kernel Hilbert space (RKHS) [18]. The RKHS of $\mathbb{R}^Q$-valued functions [19] at the basis of a Q-category M-SVM whose kernel is κ can be defined simply as a function of κ.

Definition 1 (RKHS $\bar{\mathcal{H}}$). *Let κ be a real-valued positive type function on $\mathcal{X}^2$. Then, the RKHS of $\mathbb{R}^Q$-valued functions at the basis of a Q-category M-SVM whose kernel is κ is $\bar{\mathcal{H}} = \mathbf{H}_\kappa^Q$. Furthermore, the inner product of $\bar{\mathcal{H}}$ can be expressed as a function of the inner product of $\mathbf{H}_\kappa$ as follows:*

$$\forall \left(\bar{h}, \bar{h}'\right) \in \bar{\mathcal{H}}^2, \ \bar{h} = \left(\bar{h}_k\right)_{1 \leqslant k \leqslant Q}, \ \bar{h}' = \left(\bar{h}'_k\right)_{1 \leqslant k \leqslant Q}, \ \langle \bar{h}, \bar{h}' \rangle_{\bar{\mathcal{H}}} = \sum_{k=1}^{Q} \langle \bar{h}_k, \bar{h}'_k \rangle_{\mathbf{H}_\kappa} \ .$$

Definition 2 (Class of functions $\mathcal{H}$). *Let κ be a real-valued positive type function on $\mathcal{X}^2$ and let $\bar{\mathcal{H}}$ be the RKHS of $\mathbb{R}^Q$-valued functions derived from κ according to Definition 1. Let $\{1\}$ be the one-dimensional space of real-valued constant functions on $\mathcal{X}$. The class of functions at the basis of a Q-category M-SVM whose kernel is κ is*

$$\mathcal{H} = \bar{\mathcal{H}} \oplus \{1\}^Q = \left(\mathbf{H}_\kappa \oplus \{1\}\right)^Q \ .$$

Thus, a function h in $\mathcal{H}$ can be written as

$$h(\cdot) = \bar{h}(\cdot) + b = \left(\bar{h}_k(\cdot) + b_k\right)_{1 \leqslant k \leqslant Q}$$

where the function $\bar{h} = \left(\bar{h}_k\right)_{1 \leqslant k \leqslant Q}$ is an element of $\bar{\mathcal{H}}$ and $b = (b_k)_{1 \leqslant k \leqslant Q} \in \mathbb{R}^Q$.

Let $m \in \mathbb{N}^*$. For a given sequence of examples $d_m = ((x_i, y_i))_{1 \leqslant i \leqslant m}$ in $(\mathcal{X} \times [\![1, Q]\!])^m$, we denote $\mathbb{R}^{Qm}(d_m)$ the subspace of $\mathbb{R}^{Qm}$ made up of the vectors $v = (v_t)_{1 \leqslant t \leqslant Qm}$ satisfying:

$$\left(v_{(i-1)Q+y_i}\right)_{1 \leqslant i \leqslant m} = 0_m \ . \tag{1}$$

Furthermore, for the sake of simplicity, the components of the vectors of $\mathbb{R}^{Qm}(d_m)$ are written with two indices, i.e., v_{ik} in place of $v_{(i-1)Q+k}$, for i in $[\![1, m]\!]$ and k in $[\![1, Q]\!]$. As a consequence, (1) simplifies into $(v_{iy_i})_{1 \leqslant i \leqslant m} = 0_m$. For n in $\mathbb{N}^*$, let $\mathcal{M}_{n,n}(\mathbb{R})$ be the algebra of $n \times n$ matrices over $\mathbb{R}$ and $\mathcal{M}_{Qm,Qm}(d_m)$ the subspace of $\mathcal{M}_{Qm,Qm}(\mathbb{R})$ made up of the matrices $M = (m_{tu})_{1 \leqslant t,u \leqslant Qm}$ satisfying:

$$\forall j \in [\![1, m]\!], \ \left(m_{t,(j-1)Q+y_j}\right)_{1 \leqslant t \leqslant Qm} = 0_{Qm} \ .$$

Once more for the sake of simplicity, the components of the matrices of $\mathcal{M}_{Qm,Qm}(d_m)$ are written with four indices, i.e., $m_{ik,jl}$ in place of $m_{(i-1)Q+k,(j-1)Q+l}$, for (i, j) in $[\![1, m]\!]^2$ and (k, l) in $[\![1, Q]\!]^2$. With these definitions and notations at hand, a generic model of M-SVM can be defined as follows.

Definition 3 (Generic model of M-SVM, Definition 4 in [20]). *Let $\mathcal{X}$ be a non empty set and $Q \in \mathbb{N} \setminus [\![0, 2]\!]$. Let κ be a real-valued positive type function on $\mathcal{X}^2$. Let $\bar{\mathcal{H}}$ and $\mathcal{H}$ be two classes of functions induced by κ according to Definitions 1 and 2. Let $P_{\bar{\mathcal{H}}}$ be the orthogonal projection operator from $\mathcal{H}$ onto $\bar{\mathcal{H}}$. For $m \in \mathbb{N}^*$, let $d_m = ((x_i, y_i))_{1 \leqslant i \leqslant m} \in (\mathcal{X} \times [\![1, Q]\!])^m$ and $\xi \in \mathbb{R}^{Qm}(d_m)$. A* Q-category M-SVM *whose kernel is κ is a large margin discriminant model obtained by solving a convex quadratic programming problem of the form*

Problem 1 (M-SVM learning problem, primal formulation).

$$\min_{h,\xi} \left\{ \|M\xi\|_p^p + \lambda \|P_{\bar{\mathcal{H}}} h\|_{\bar{\mathcal{H}}}^2 \right\}$$

$$s.t. \begin{cases} \forall i \in [\![1, m]\!], \ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \ \ K_1 h_{y_i}(x_i) - h_k(x_i) \geqslant K_2 - \xi_{ik} \\ \forall i \in [\![1, m]\!], \ \forall (k, l) \in \left([\![1, Q]\!] \setminus \{y_i\}\right)^2, \ \ K_3(\xi_{ik} - \xi_{il}) = 0 \\ \forall i \in [\![1, m]\!], \ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \ \ (2 - p)\xi_{ik} \geqslant 0 \\ (1 - K_1) \sum_{k=1}^{Q} h_k = 0 \end{cases}$$

where $\lambda \in \mathbb{R}_+^*$, $M \in \mathcal{M}_{Qm,Qm}(d_m)$ is a matrix of rank $(Q-1)m$, $p \in \{1, 2\}$, $(K_1, K_3) \in \{0, 1\}^2$, and $K_2 \in \mathbb{R}_+^*$. If $p = 1$, then M is a diagonal matrix.

In the chronological order of their introduction, the four M-SVMs involved in our experiments are those of Weston and Watkins [21], Crammer and Singer [22], Lee and co-authors [23], and the M-SVM^2 [24]. Let $I_{Qm}\,(d_m)$ and $M^{(2)}$ designate two matrices of $\mathcal{M}_{Qm,Qm}\,(d_m)$ whose general terms are respectively

$$m_{ik,jl} = \delta_{i,j}\delta_{k,l}\,(1 - \delta_{y_i,k})$$

and

$$m_{ik,jl} = (1 - \delta_{y_i,k})\,(1 - \delta_{y_j,l})\left(\delta_{k,l} + \frac{\sqrt{Q} - 1}{Q - 1}\right)\delta_{i,j},$$

where δ is the Kronecker symbol. In order to characterize the aforementioned M-SVMs as instances of the generic model, we express the primal formulation of their learning problems as a specification of Problem 1. The corresponding values of the hyperparameters are gathered in Table 1.

Table 1. Specifications of the four M-SVMs used as sequence-to-structure classifier. The first three machines are the ones of Weston and Watkins (WW), Crammer and Singer (CS), and Lee, Lin, and Wahba (LLW).

M-SVM	M	p	K_1	K_2	K_3
WW-M-SVM	$I_{Qm}\,(d_m)$	1	1	1	0
CS-M-SVM	$\frac{1}{Q-1}I_{Qm}\,(d_m)$	1	1	1	1
LLW-M-SVM	$I_{Qm}\,(d_m)$	1	0	$\frac{1}{Q-1}$	0
M-SVM^2	$M^{(2)}$	2	0	$\frac{1}{Q-1}$	0

2.2 Dedication to Sequence-to-Structure Prediction

In our experiments, each protein sequence is represented by a position-specific scoring matrix (PSSM) produced by PSI-BLAST [25]. To generate each PSSM, we ran three iterations of PSI-BLAST against the nr database downloaded in February 2010. The E-value inclusion threshold was set to 0.005 and the default scoring matrix (BLOSUM62) was used. The sliding window of the sequence-to-structure classifiers, i.e., the M-SVMs, is of size 13, and is centered on the residue of interest. The description (vector of predictors) x_i processed by the M-SVMs to predict the conformational state of the i^{th} residue in the data set is thus obtained by appending rows of the PSSM associated with the sequence to which it belongs. The indices of these rows range from $i' - 6$ to $i' + 6$, where i' is the index of the residue of interest in its sequence. Since a PSSM has 20 columns, one per amino acid, this corresponds to 260 predictors. More precisely, $\mathcal{X} \subset \mathbb{Z}^{260}$. The kernel κ is an elliptic Gaussian kernel function applying a weighting on the predictors as a function of their position in the window. This weighting is learned by application of the principle of multi-class kernel target alignment [26] (the training algorithm is a stochastic steepest descent). At last, the programs implementing the different M-SVMs are those of MSVMpack [27].

3 Structure-to-Structure Classifiers

In this section, we make the hypothesis that N M-SVMs are available to perform the sequence-to-structure prediction. The function computed by the j^{th} of these machines is denoted $h^{(j)} = \left(h_k^{(j)}\right)_{1 \leqslant k \leqslant Q}$. The structure-to-structure classifier $g = (g_k)_{1 \leqslant k \leqslant Q}$ uses a sliding window with a left context of size T_l and a right context of size T_r (for a total length of $T = T_l + 1 + T_r$). Thus, the vector of predictors available to estimate the probabilities associated with the i^{th} residue in the data set is $z_i = \left(h_k^{(j)}\left(x_{i+t}\right)\right)_{1 \leqslant j \leqslant N, 1 \leqslant k \leqslant Q, -T_l \leqslant t \leqslant T_r} \in \mathbb{R}^{NQT}$. With slight abuse of notation, we use $g\left(x_i\right)$ in place of $g\left(z_i\right)$ to denote the outputs of the structure-to-structure classifier for this residue. The classifiers we consider to perform the task are now introduced in order of increasing complexity.

3.1 Polytomous Logistic Regression

The most natural way of deriving class posterior probability estimates from the outputs of an M-SVM consists in extending Platt's bi-class solution [28] to the multi-class case. In the framework of our study (implementation of the cascade architecture), this corresponds to applying to the predictors the parametric form of the softmax function such that

$$\forall i \in [\![1, m]\!], \quad g\left(x_i\right) = \left(\frac{\exp\left(\sum_{j=1}^{N} \sum_{t=-T_l}^{T_r} a_{kjt} h_k^{(j)}\left(x_{i+t}\right) + b_k\right)}{\sum_{l=1}^{Q} \exp\left(\sum_{j=1}^{N} \sum_{t=-T_l}^{T_r} a_{ljt} h_l^{(j)}\left(x_{i+t}\right) + b_l\right)} \right)_{1 \leqslant k \leqslant Q}.$$

The values of the corresponding parameters, the coefficients a_{kjt} and the biases b_k, are obtained by maximum likelihood estimation (the training criterion used is cross-entropy), so that the model specified appears as a simplified variant of the polytomous (multinomial) logistic regression model [29]. In practice, the training algorithm that we implemented is a multi-class extension of the one exposed in [30].

3.2 Linear Ensemble Methods

In [31], we studied the class of *linear ensemble methods* (LEMs). Their use requires an additional hypothesis regarding the base classifiers, namely that they take their values in the probability simplex (which is less restrictive than assuming that their outputs are probability estimates). For the sake of simplicity, we present them without taking into account the sliding window. In practice, its introduction is equivalent to the introduction of $N(T-1)$ additional classifiers. For all k in $[\![1, Q]\!]$, let t_k denote the *one of Q coding* of category k, i.e., $t_k = (\delta_{k,l})_{1 \leqslant l \leqslant Q}$. With this notation at hand, an LEM is defined as follows.

Definition 4 (Linear ensemble method). *Let* $\left\{\tilde{h}^{(j)} = \left(\tilde{h}_k^{(j)}\right)_{1 \leqslant k \leqslant Q} : 1 \leqslant j \leqslant N\right\}$ *be a set of N classifiers taking their values in the probability simplex. Let* $\beta =$

$(\beta_k)_{1 \leqslant k \leqslant Q}$, *with the vectors* β_k *belonging to* $\mathbb{R}^{NQ}$. *For all* j *in* $[\![1, N]\!]$ *and all* (k, l) *in* $[\![1, Q]\!]^2$, *let* β_{kjl} *denote the component of vector* β_k *of index* $(j-1)Q+l$. *Then, a* linear ensemble method *is a multivariate linear model of the form*

$$\forall x \in \mathcal{X}, \;\; g(x) = \left(\sum_{j=1}^{N} \sum_{l=1}^{Q} \beta_{kjl} \tilde{h}_l^{(j)}(x) \right)_{1 \leqslant k \leqslant Q}$$

for which the vector β *is obtained by solving a convex programming problem of the type*

Problem 2 (Learning problem of an LEM)

$$\min_{\beta} \sum_{i=1}^{m} \ell_{LEM}\left(t_{y_i}, g\left(x_i\right)\right)$$

$$s.t. \;\; \begin{cases} \beta \in \mathbb{R}_+^{NQ^2} \\ \forall j \in [\![1, N]\!], \; \forall l \in [\![1, Q-1]\!], \;\; \sum_{k=1}^{Q} \left(\beta_{kjl} - \beta_{kjQ}\right) = 0 \\ 1_{NQ^2}^T \beta = Q \end{cases}$$

where the loss function ℓ_{LEM} *is convex.*

Note that a special case of LEM is the standard convex combination obtained by adding the following constraint:

$$\forall (j, k, l) \in [\![1, N]\!] \times [\![1, Q]\!] \times [\![1, Q]\!], \;\; k \neq l \Longrightarrow \beta_{kjl} = 0 \;\; .$$

The constraints of Problem 2 are sufficient to ensure that g takes its values in the probability simplex. Furthermore, if ℓ_{LEM} is the quadratic loss (Brier score) or the cross-entropy loss, then the outputs are actually estimates of the class posterior probabilities (see [31] for the proof). In order to make comparison with the polytomous logistic regression straightforward, we selected the second option for our experiments. Since the M-SVMs do not take their values in the probability simplex, their outputs must be post-processed prior to being combined by an LEM. This can be done by means of the polytomous logistic regression model. In that case, the flowchart of the cascade architecture is the one depicted in Figure 1 (right).

3.3 Multi-layer Perceptron

As pointed out in the introduction, the MLP is the standard structure-to-structure classifier. In our experiments, we implemented it with the softmax activation function for the output units (a sigmoid for the hidden units) and the cross-entropy loss function. In that way, it could be seen as a complexified extension of the polytomous logistic regression model described in Section 3.1, making once more the comparison of performance straightforward.

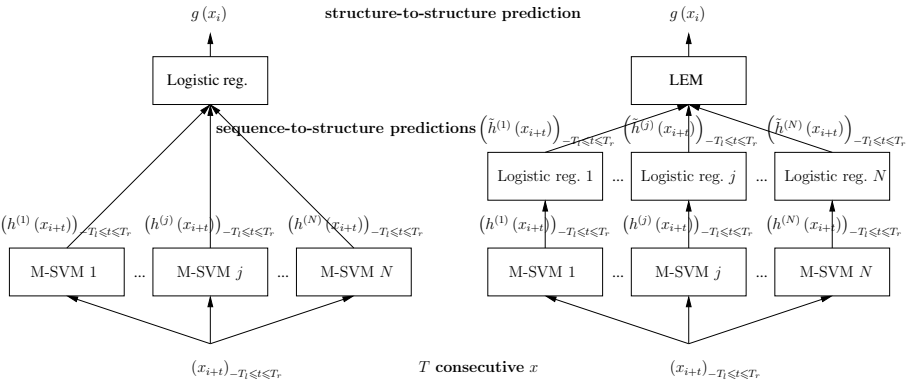

Fig. 1. Flowchart of the computation of the class posterior probabilities using a polytomous logistic regression model (left) and an LEM (right) as structure-to-structure classifier

4 Experimental Results

To assess the different structure-to-structure classifiers described in Section 3, we used the CB513 data set [32]. The 513 sequences of this set are made up of 84119 residues. The derivation of the descriptions x_i of these residues (inputs of the sequence-to-structure classifiers) has been detailed in Section 2.2. As for their labels y_i, the initial secondary structure assignment was performed by the DSSP program [33], with the reduction from 8 to 3 conformational states following the CASP method, i.e., H+G $\rightarrow$ H (α-helix), E+B $\rightarrow$ E (β-strand), and all the other states in C (coil). In all cases, the two sliding windows used were centered on the residue of interest, and their respective sizes were 13 (sequence-to-structure) and 15 (structure-to-structure, $T_l = T_r = 7$). The performance of reference was provided by a cascade architecture involving two MLPs. The respective sizes of their hidden layers are 16 (sequence-to-structure) and 6 (structure-to-structure). The structure-to-structure network is precisely the one described in Section 3.3.

A secondary structure prediction method must fulfill different requirements in order to be useful for the biologist. Thus, several standard measures giving complementary indications must be used to assess the prediction accuracy [34]. We computed the three most popular ones: the recognition rate Q_3, Pearson-Matthews correlation coefficients $C_{\alpha/\beta/\mathrm{coil}}$, and the segment overlap measure (Sov) in its most recent version (Sov'99). The quality of the probability estimates was measured by means of the (averaged) cross-entropy (CE). To train the two levels of the cascade and assess performance, a seven-fold cross-validation procedure was implemented. At each step, two thirds of the training set were used to train the sequence-to-structure classifiers, and one third to train the structure-to-structure classifier. The experimental results obtained are gathered in Table 2.

Table 2. Prediction accuracy and quality of the probability estimates for the different implementations of the cascade architecture considered

Cascade architecture	Q_3 (%)	C_α	C_β	C_{coil}	Sov'99 (%)	CE
MLP + MLP	74.6	0.69	0.59	0.54	71.1	0.615
M-SVMs + logistic reg.	76.5	0.71	0.62	0.57	73.0	0.576
M-SVMs + logistic reg. + LEM	76.5	0.71	0.63	0.57	**73.1**	0.576
M-SVMs + MLP	**76.7**	0.72	0.63	0.57	71.9	**0.572**

Using the two sample proportion test (the one for large samples), the gain in prediction accuracy resulting from using dedicated M-SVMs in place of an MLP for the sequence-to-structure prediction of the cascade architecture appears always (i.e., irrespective of the choice of the structure-to-structure classifier) statistically significant with confidence exceeding 0.95. The value of the cross-entropy confirms this superiority. If we restrict to the architecture we advocate, then the recognition rates of the three structure-to-structure classifiers are almost identical. The value of the cross-entropy does not really help to break the tie. However, this goal can be achieved by resorting to the Sov. In that case, the logistic regression and the LEM appear almost equivalent, and significantly superior to the MLP. The reason for this asymmetry is still to be highlighted.

Thus, the main conclusion that can be drawn regarding the choice of the structure-to-structure classifier is that the prediction accuracy does not benefit significantly from an increase in the complexity. From this point of view, the small size of the hidden layer of the MLP used for this task is telling. This phenomenon can be explained by a well-known fact in secondary structure prediction: the main limiting factor when applying a cascade architecture is overfitting (see for instance [35]). On the one hand, the sequence-to-structure classifiers must be complex enough to cope with the complexity of the task, but on the other hand, the classifier at the second level must be of far lower capacity, otherwise its recognition rate in test will be disconnected with its recognition rate on the training set. With this restriction in mind, the behavior of the LEM appears promising, since its accuracy with respect to both major criteria (Q_3 and cross-entropy) is similar to the one of the other combiners although it requires an additional level of training, resulting from the need for a post-processing of the outputs of the M-SVMs prior to their combination (see Figure 1). Given the experimental protocol described above, we used the same training set to train both the M-SVMs and their post-processing, a strategy which is prone to overfitting. As a consequence, we conjecture that the prediction accuracy of the LEM is the one which should benefit most from the availability of additional training data.

To sum up, these experiments back our thesis that M-SVMs endowed with a dedicated kernel should be used as sequence-to-structure classifiers in a cascade architecture, even in the case when the final outputs must be class posterior probability estimates. In that case, several options are available for the structure-to-structure classifier. So far, it appears that the difference between them is not

significant. The levelling of performance primarily highlights the problem of overfitting.

5 Conclusions and Ongoing Research

This article has addressed the problem of optimizing the cascade architecture commonly used for protein secondary structure prediction with respect to two criteria: prediction accuracy and quality of the class posterior probability estimates. The main result is that using dedicated M-SVMs as sequence-to-structure classifiers, it is possible to outperform the standard solution, consisting in using two MLPs in sequence, according to both criteria. Making the best of the new approach should require a precise choice for the structure-to-structure classifier. From that point of view, capacity control appears to play a major part, since overfitting is a strong limiting factor. Obviously, a touchstone for the architecture we advocate is a comparison with solutions based on *pairwise coupling* (Bradley-Terry model) [36]. To that end, we are currently conducting additional comparative experiments with architectures differing from ours at the sequence-to-structure level: the M-SVMs are replaced with decomposition schemes involving bi-class machines.

This contribution paves the way for the specification of new hybrid prediction methods combining discriminant models and HMMs. By the way, the principle of these methods could be extended to many other fields of bioinformatics, including alternative splicing prediction.

Acknowledgements. This work was supported by the MBI project of the PRST MISN. The authors would like to thank the anonymous reviewers for their comments.

References

1. Qian, N., Sejnowski, T.J.: Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology 202, 865–884 (1988)
2. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292, 195–202 (1999)
3. Pollastri, G., Przybylski, D., Rost, B., Baldi, P.: Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47, 228–235 (2002)
4. Cole, C., Barber, J.D., Barton, G.J.: The Jpred 3 secondary structure prediction server. Nucleic Acids Research 36, W197–W201 (2008)
5. Kountouris, P., Hirst, J.D.: Prediction of backbone dihedral angles and protein secondary structure using support vector machines. BMC Bioinformatics 10, 437 (2009)
6. Anthony, M., Bartlett, P.L.: Neural Network Learning: Theoretical Foundations. Cambridge University Press, Cambridge (1999)
7. Hua, S., Sun, Z.: A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. Journal of Molecular Biology 308, 397–407 (2001)

8. Guermeur, Y.: Combining discriminant models with new multi-class SVMs. Pattern Analysis and Applications 5, 168–179 (2002)
9. Guermeur, Y., Pollastri, G., Elisseeff, A., Zelus, D., Paugam-Moisy, H., Baldi, P.: Combining protein secondary structure prediction models with ensemble methods of optimal complexity. Neurocomputing 56, 305–327 (2004)
10. Nguyen, M.N., Rajapakse, J.C.: Two-stage multi-class support vector machines to protein secondary structure prediction. In: 10th Pacific Symposium on Biocomputing, pp. 346–357 (2005)
11. Richard, M.D., Lippmann, R.P.: Neural network classifiers estimate Bayesian *a posteriori* probabilities. Neural Computation 3, 461–483 (1991)
12. Rojas, R.: A short proof of the posterior probability property of classifier neural networks. Neural Computation 8, 41–43 (1996)
13. Lin, K., Simossis, V.A., Taylor, W.R., Heringa, J.: A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21, 152–159 (2005)
14. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 257–286 (1989)
15. Guermeur, Y.: VC theory of large margin multi-category classifiers. Journal of Machine Learning Research 8, 2551–2594 (2007)
16. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge (2002)
17. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
18. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers, Boston (2004)
19. Wahba, G.: Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In: Casdagli, M., Eubank, S. (eds.) Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, vol. XII, pp. 95–112. Addison-Wesley (1992)
20. Guermeur, Y.: A generic model of multi-class support vector machine. International Journal of Intelligent Information and Database Systems (accepted)
21. Weston, J., Watkins, C.: Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science (1998)
22. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research 2, 265–292 (2001)
23. Lee, Y., Lin, Y., Wahba, G.: Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association 99, 67–81 (2004)
24. Guermeur, Y., Monfrini, E.: A quadratic loss multi-class SVM for which a radius-margin bound applies. Informatica 22, 73–96 (2011)
25. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402 (1997)
26. Guermeur, Y., Lifchitz, A., Vert, R.: A kernel for protein secondary structure prediction. In: Schölkopf, B., Tsuda, K., Vert, J.-P. (eds.) Kernel Methods in Computational Biology, pp. 193–206. The MIT Press, Cambridge (2004)
27. Lauer, F., Guermeur, Y.: MSVMpack: a multi-class support vector machine package. Journal of Machine Learning Research 12, 2293–2296 (2011)

28. Platt, J.C.: Probabilities for SV machines. In: Smola, A.J., Bartlett, P.L., Schölkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers, pp. 61–73. The MIT Press, Cambridge (2000)
29. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. Wiley, London (1989)
30. Lin, H.-T., Lin, C.-J., Weng, R.C.: A note on Platt's probabilistic outputs for support vector machines. Machine Learning 68, 267–276 (2007)
31. Guermeur, Y.: Combining multi-class SVMs with linear ensemble methods that estimate the class posterior probabilities. Communications in Statistics (submitted)
32. Cuff, J.A., Barton, G.J.: Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 34, 508–519 (1999)
33. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983)
34. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424 (2000)
35. Riis, S.K., Krogh, A.: Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. Journal of Computational Biology 3, 163–183 (1996)
36. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. The Annals of Statistics 26, 451–471 (1998)