Application of the Multi-modal Relevance Vector Machine to the Problem of Protein Secondary Structure Prediction

Nikolay Razin[1], Dmitry Sungurov[1], Vadim Mottl[2], Ivan Torshin[2],
Valentina Sulimova[3], Oleg Seredin[3], and David Windridge[4]

[1] Moscow Institute of Physics and Technology, Moscow, Russia
[2] Computing Center of the Russian Academy of Sciences, Moscow, Russia
[3] Tula State University, Tula, Russia
[4] University of Surrey, Guildford, UK

Abstract. The aim of the paper is to experimentally examine the plausibility of Relevance Vector Machines (RVM) for protein secondary structure prediction. We restrict our attention to detecting strands which represent an especially problematic element of the secondary structure. The commonly adopted local principle of secondary structure prediction is applied, which implies comparison of a sliding window in the given polypeptide chain with a number of reference amino-acid sequences cut out of the training proteins as benchmarks representing the classes of secondary structure. As distinct from the classical RVM, the novel version applied in this paper allows for selective combination of several tentative window comparison modalities. Experiments on the RS126 data set have shown its ability to essentially decrease the number of reference fragments in the resulting decision rule and to select a subset of the most appropriate comparison modalities within the given set of the tentative ones.

Keywords: Protein secondary structure prediction, machine learning, multi-modal relational pattern recognition, Relevance Vector Machine, controlled selectivity of reference objects and object-comparison modalities.

1 Introduction

Within the currently dominant paradigm of the protein science, the primary structure of a protein uniquely determines its spatial structure, which in turn determines the biological roles of the protein. Consequently, one of the main tasks of theoretical protein biology and bioinformatics is the establishment of the laws that govern the relationship between the primary and the spatial protein structure.

The secondary structure represents a projection of the local geometry of the spatial (tertiary) protein structure into a sequence of letters in a certain alphabet, most commonly, H – helix, S – strand, C – coil. The secondary structure prediction is increasingly becoming the work horse for numerous methods aimed at solving the much more challenging problem of predicting the spatial structure [1,2].

T. Shibuya et al. (Eds.): PRIB 2012, LNBI 7632, pp. 153–165, 2012.

The problem of protein secondary structure prediction was first conceived in the early 1960s, when a number of protein structures were determined by X-ray crystallography. A significant increase in the prediction accuracy was achieved once machine learning approaches were applied for solving the problem [3]. Despite an increase in the average accuracy, there is an evident lack of progress in this area in recent decades. For example, experiments within the framework of the conference-tournament CASP (Critical Assessment of the Protein Structure Prediction) [4], which have been carried out since the early 1990s, clearly show the absence of any significant positive trend in the accuracy of protein secondary structure prediction for at least 10 years from 1992 to 2002. It is perhaps for this reason that the problem of protein secondary structure prediction was even removed from the list of problems studied within the CASP framework (see publications of CASP-5 [5]).

The absence of any measurable progress is likely to be the result of numerous auxiliary assumptions of biological sort that underlie the prediction scenarios. It appears appropriate to develop and test algorithms, which should be based on the minimum possible number of additional assumptions drawn from biology and include adequate procedures for the selection of features representing amino acid sequences [6], as well as incorporate adequate training procedures for inferring relationship between the primary and the secondary structure from sufficiently large sets of proteins with the known spatial structure.

The commonly adopted principle of predicting the secondary structure at a position t in the polypeptide chain is its estimation from the local context, i.e., an amino acid window of a fixed length symmetric in relation to the target location t [3]. Given a training set of proteins whose known secondary structures are represented by strings on the three-letter alphabet $\{h,s,c\}$, the problem of inferring the prediction rule is that of pattern recognition.

The Support Vector Machine (SVM) is the most popular method of machine learning in pattern recognition learning [7]. As applied to secondary structure prediction [8], one of its advantages is that it yields a decision rule of classifying amino acid windows in new proteins on the basis of their comparison with a relatively small number of so-called support fragments inferred from the training set as result of training. However, the pay-off for this advantage is the onerous restriction that the comparison function must be a kernel, i.e., must possess the mathematical properties of the inner product in some hypothetical linear space into which the kernel embeds any set of objects. Elements of a linear space are usually called vectors, and this has led to the name of Support Vector Machine.

This paper is motivated by two intents – first, to remove any restrictions on the manner of comparison between amino acid fragments in contrast to excessively exacting kernels, and, second, to essentially decrease the number of reference fragments in the resulting decision rule. With this purpose, we rest here not on Vapnik's traditional SVM, but on the SVM-based Relevance Vector Machine (RVM) by Bishop and Tipping [9]. Two main advantages of the RVM technique are, first, just tolerance to any kinds of object comparison and, second, usage in the decision rule, instead of relatively few support vectors yielded by SVM, a still smaller number of so-called relevance vectors. In the problem of secondary structure prediction, this means that the structure states at subsequent points in the polypeptide chain of a new protein will be predicted by comparison of the respective windows with only a few reference sequences cut out of the training proteins as some sort of benchmarking windows representing the classes of the secondary structure.

For the window-based prediction of the protein secondary structure, we apply the multi-modal modification of the Relevance Vector Machine described in [10], which, in addition, allows to select a subset of the most appropriate window comparison functions within the given set of tentative ones.

For verification of the proposed technique, we used the RS126 set of protein chains as the source of both training and test sets.

To test the ability of the multimodal RVM to select most relevant comparison functions, two kinds of comparison principles were examined jointly – position-dependence of amino acids in fragments corresponding to the same local secondary structure in a protein [11,12] and a newly developed principle based on Fourier representation of both sequences as functions along the polypeptide axis.

We restrict here our attention to detecting strands in the secondary structure of proteins, which, as practice shows, represent an especially problematic element of the secondary structure. The aim of the paper is rather to explore the performance of the Relevance Vector Machine in the problem of widow-based secondary structure prediction than achieving some record-breaking results. Nevertheless, experiments on the RS126 data set have shown the accuracy of about 75% in detecting strands as especially problematic element of the secondary structure.

2 The Local Machine-Learning Approach to Secondary Structure Prediction – Pattern Recognition in a Sliding Amino Acid Window

Let $\boldsymbol{\omega}=(\alpha_t, t=1,...,M)$ be the finite amino acid sequence which represents the primary structure of a protein of individual length $M{=}M_{\omega}$, where $\alpha_t{\in}\mathbb{A}{=}\{\alpha^1,...,\alpha^m\}$, $m=20$ are symbols corresponding to the alphabet of amino acids. The protein's hidden secondary structure will be completely represented by a symbolic sequence $\mathbf{y}=(y_t, t=1,...,M)$ of the same length $M{=}M_{\omega}$, whose elements $y_t\in\mathbb{Y}=\{h,s,c\}$ are associated with three classes of structure: h – helix, s – sheet, c – unspecified structure usually referred to as coil.

Let, further, the observer be submitted a training set of proteins whose amino acid sequences are labeled by the "correct" assignments of secondary structure:

$$\left\{(\boldsymbol{\omega}_l, \mathbf{y}_l), l{=}1,...,N^0\right\},\ \boldsymbol{\omega}_l=(\alpha_{lt}, t=1,...,M_l),\ \mathbf{y}_l=(y_{lt}, t=1,...,M_l) \tag{1}$$

Given a new amino acid sequence $\boldsymbol{\omega}=(\alpha_t, t=1,...,M_{\omega})$ not represented in the training set, we are required to estimate the secondary structure of the respective protein $\hat{\mathbf{y}}(\boldsymbol{\omega})=\left(\hat{y}_t(\boldsymbol{\omega}), t{=}1,...,M_{\omega}\right)$.

Following [13], in this paper we restrict our consideration to prediction based on the principle of a sliding amino acid window. This means that the decision on the class of secondary structure at position t is made from the symmetric interval

$\omega_t = (\alpha_\tau, t-T \le \tau \le t+T)$ of the entire amino acid chain $\boldsymbol{\omega} = (\alpha_t, t = 1,...,N)$. The odd width $\mathcal{T} = 2T+1$ of the sliding window is thus defined by its half-width T as a parameter to be preset. Estimation of the secondary structure of a protein thus takes place only within its amino acid sequence truncated at both sides by the window's half-width $\hat{\mathbf{y}}(\boldsymbol{\omega}) = \left(\hat{y}_t(\boldsymbol{\omega}), T+1 \le t \le M-T\right) = \left(\hat{y}_t(\omega_t),\ T+1 \le t \le M-T\right)$.

Thus, the original problem of predicting the entire secondary structure of a protein $\hat{\mathbf{y}}(\boldsymbol{\omega})$ is reduced to the series of independent problems $\hat{y}_t(\omega_t) = \hat{y}_t(\alpha_{t-T},...,\alpha_t,...,\alpha_{t+T})$ of estimating the class of secondary structure $\hat{y}_t \in \{h,s,c\}$ for the central amino acid α_t in the respective window.

The window-based approach implies treating the training set as an unordered assembly of all continuous amino acid fragments $\left\{(\omega_j, y_j),\ j=1,...,N\right\}$ cut out of the given set of indexed amino acid sequences $\omega_j = (\alpha_{j\tau}, t-T \le \tau \le t+T)$, $y_j \in \{h,s,c\}$ (1). As a simplification resulting from our restricting the problem to distinguishing between strands and other elements of the secondary structure, we shall train a two-class classifier: $y_j \in \{1,-1\} = \{s,\bar{s}\} = \left\{s, \{h,c\}\right\}$.

3 The Multi-modal Relevance Vector Machine

The mathematical and algorithmic technique we use for window-based prediction of protein secondary structure is that of the multi-modal Relevance Vector Machine outlined in [10] which rests on three well-established principles of pattern-recognition learning.

First of all, we proceed from the featureless approach proposed by Duin et al. [14] under the name of Relational Discriminant analysis, which consists in the idea of representing the pattern recognition objects ω, not by individual feature vectors $\mathbf{x}(\omega) \in \mathbb{R}^k$, but by an arbitrary real-valued measure of pair-wise relation between them. In terms of window-based secondary structure prediction, the idea is to treat the values of this function between an arbitrary amino acid fragment ω and those of the training set $\left\{(\omega_j, y_j),\ j=1,...,N\right\}$ as the vector of secondary features $\left(x_j(\omega) = S(\omega_j, \omega), j=1,...,N\right)$. Then, the standard convex SVM training technique will yield the parameters $(a_1,...,a_N,b)$ of a discriminant hyperplane in the linear space of secondary features $\mathbb{R}^N$

$$d(\omega) = \sum\nolimits_{j=1}^{N} a_j S(\omega_j, \omega) + b \gtrless 0, \tag{2}$$

which can be applied it to any new amino acid fragment

$$\omega = (\alpha_\tau, -T \le \tau \le T). \tag{3}$$

In order to weaken the demand of storing very large numbers of reference amino acid fragments $\{\omega_j, j=1,...,N\}$, we apply Bishop and Tipping's Relevance Vector

Machine (RVM) [9], underpinned by the notion of selecting only a small number of most informative Relevance Objects in the training set:

$$d(\omega)=\sum\nolimits_{j\in\hat{J}} a_j S(\omega_j,\omega)+b\gtrless 0\,,\quad \hat{J}\subset\{1,...,N\}\,. \tag{4}$$

However, the Bayesian principle of selecting secondary features implied by the original RVM results in a non-convex training problem.

The novel aspect of [10] which is immediately applicable in this paper is the assumption that several comparison modalities for pair-wise object representation are available $S_i(\omega',\omega''),\ \ i=1,...,n$. The presence of several object-comparison functions expands the number nN of secondary features for any object $\left(x_{ij}(\omega)=S_i(\omega_j,\omega),\right.$ $\left.i=1,...,n,\, j=1,...,N\right)$. A straightforward generalization of the doubly-regularized SVM [15] has led in [10] to the *multimodal* convex training criterion which we call the multi-modal Relevance Vector Machine and which we shall apply in this paper to training sets of amino acid fragments $\left\{(\omega_j,y_j),\ j=1,...,N\right\}$:

$$\begin{cases}\sum\nolimits_{i=1}^{n}\sum\nolimits_{l=1}^{N}\left[(1-\mu)a_{il}^2+\mu\left|a_{il}\right|\right]+C\sum\nolimits_{j=1}^{N}\delta_j\rightarrow\min(a_{il},b,\delta_j),\\ y_j\left(\sum\nolimits_{i=1}^{n}\sum\nolimits_{l=1}^{N}a_{il}S_i(\omega_l,\omega_j)+b\right)\geq 1-\delta_j,\,\delta_j\geq 0,\, j=1,...,N.\end{cases} \tag{5}$$

This training criterion differs from the usual SVM by a more complicated regularization term which is a mix of L_2 and L_1 norms of the direction vector with an additional weighting parameter $0\leq\mu<1$ instead of the pure L_2 norm in the classical case.

We shall use the following notations for sets of, respectively, object-comparison modalities, training objects and all secondary features:

$$I=\left\{1,...,n\right\},\, J=\left\{1,...,N\right\},\, F=\left\{ij, i=1,...,n,\, j=1,...,N\right\}=I\times J.$$

The training criterion (5) is both modality-selective and reference-object-selective, therefore, we refer to it as the modality-selective Relevance Vector Machine. The subset of relevant secondary features $\hat{F}=\left\{ij\!:\hat{a}_{ij}\neq 0\right\}\subseteq F$ determines the subsets of relevant modalities $\hat{I}$ and relevant objects $\hat{J}$:

$$\hat{F}=\left\{ij\!:\hat{a}_{ij}\neq 0\right\}\subseteq F\!: \Rightarrow\quad \hat{I}=\left\{i\!:\exists j(a_{ij}\neq 0)\right\}\subseteq I=\{1,...,n\},\quad \hat{J}=\left\{j\!:\exists i(a_{ij}\neq 0)\right\}\subseteq J=\{1,...,N\}. \tag{6}$$

As a result, the optimal discriminant hyperplane, being a generalized analog of (4), takes into account only the relevance modalities of any new object, and is completely determined by the relevance objects of the training set:

$$d(\omega)=\sum\nolimits_{ij\in\hat{F}}a_{ij}S_i(\omega_j,\omega)+b\gtrless 0\,,\quad \hat{F}\subseteq F\,. \tag{7}$$

If $\mu=0$, the method equates to the classical SVM retaining all the secondary features $x_{ij}(\omega)=S_i(\omega_j,\omega)$, namely, the entire training set as the set of reference objects (2) and all the object-comparison modalities expressed by functions $S_i(\omega_j,\omega)$. As the structural parameter grows $0 \rightarrow \mu \rightarrow 1$, the subset of relevance features $\hat{F}$ diminishes, and both subsets of relevance objects $\hat{J}$ and relevance comparison modalities $\hat{I}$ shrink along with it. If $\mu \rightarrow 1$, the criterion becomes extremely selective. Experiments have shown [10] that in the latter case it becomes practically equivalent to the original RVM [9] except for having the favourable feature of being convex.

4 Modalities of Pair-Wise Amino Acid Fragment Comparison for Protein Secondary Structure Prediction

In this paper, we experimentally apply the outlined multimodal RVM technique to protein secondary structure prediction by utilizing several different modalities of amino acid sequence comparison. Two kinds of comparison principles are jointly examined – similarity measures exploiting the position-dependence of amino acids in fragments corresponding to the same local secondary structure in a protein [12], and a newly developed class of similarity measures implied by Fourier representation of both sequences as functions along the polypeptide axis.

In accordance with (3), each comparison function $S_i(\omega',\omega'')$ must be applicable to any two amino acid fragments $\omega'=(\alpha'_\tau, -T\leq\tau\leq T)$ and $\omega''=(\alpha''_\tau, -T\leq\tau\leq T)$ of length $2T+1$ defined by the half-width parameter of the window T. In our experiments, we examined two different half-width parameters:

– $T=6$, i.e., the window length $2T+1=13$, for comparison from the viewpoint of amino acid positions,

– $T=17$, i.e., the window length $2T+1=35$, for Fourier-based comparison; this window length fulfills the goal of exploring long-range dependencies of protein secondary structure on the amino acid sequence.

On the basis of each of these two comparison principles, we constructed three different comparison functions. So, we consider all in all $n=6$ functions of pair-wise amino acid fragment comparison.

4.1 Amino-Acid-Position-Based Comparison

This form of comparison implements and generalizes the method of [12]. Let $\mathcal{A}=\{\alpha^1,...,\alpha^{20}\}$ be the alphabet of amino acids. For each position $-T\leq\tau\leq T$ in the window $\omega=(\alpha_\tau, -T\leq\tau\leq T)$ and each of 20 amino acids, a binary feature is defined $z_{\tau k}(\omega)=1$ if $\alpha_\tau=\alpha^k$ and $z_{\tau k}(\omega)=0$ if $\alpha_\tau\neq\alpha^k$. All the features jointly make the binary $20(2T+1)$ -dimensional feature vector $\mathbf{z}(\omega)=\left(z_{\tau k}(\omega), -T\leq\tau\leq T, k=1,...,20\right)$. We examined three fragment comparison functions based on such features:

$$\begin{aligned} &S_1(\omega', \omega'') = \sum\nolimits_{\tau=-T}^{T} \sum\nolimits_{k=1}^{20} z_{\tau k}(\omega') z_{\tau k}(\omega''), \\ &S_2(\omega', \omega'') = \exp\left\{ -\gamma \left[z_{\tau k}(\omega') - z_{\tau k}(\omega'') \right]^2 \right\}, \\ &S_3(\omega', \omega'') = \sum\nolimits_{\tau=-T}^{T} \sum\nolimits_{k=1}^{20} \left| z_{\tau k}(\omega') - z_{\tau k}(\omega'') \right|. \end{aligned} \tag{8}$$

Two former comparison functions were examined separately in [12]; both of them are kernels on the set of amino acid fragments, but this fact is out of significance in our approach.

It is shown in [12] that the amino-acid-position-based principle of comparison is more adequate to relatively short windows, therefore, we use it with the recommended window length $2T+1=13$.

4.2 Fourier-Transform-Based Comparison

This method is proposed here for the first time. It rests on the fact that both PAM and BLOSUM amino acid substitution matrices result from the same PAM evolutionary model [16], namely, an assumed ergodic and reversible Markov chain, and the main difference between them lies in the different initial data for estimating unknown transition probabilities [17]. Moreover, it is shown in [17] that that all PAM and BLOSUM substitution matrices express probabilities of the existence of a common ancestor for each pair of amino acids and are, by their nature, positive semidefinite matrices. This innate positive semi-definiteness is absent in published matrices only because of traditional logarithmic representation and rounding down to whole numbers.

The initial positive definite PAM matrices for any evolutionary distance can be easily computed from the estimated transition probabilities PAM1 available in [16] via the algorithm outlined in [17].

For the Fourier representation of amino acid fragments $\omega = (\alpha_\tau, -T \le \tau \le T)$, we use the positive definite PAM250 matrix, which we denote as $\mathbf{M} = \big(\mu(\alpha^k, \alpha^l),\ k,l = 1,...,20\big)$. Its positive eigenvalues $\eta^q > 0$ and eigenvectors $\mathbf{M}\mathbf{h}^q = \eta^q \mathbf{h}^q$, $\mathbf{h}^q = (h_1^q \cdots h_{20}^q) \in \mathbb{R}^{20}$, $q=1,...,20$, satisfy the equality $\mathbf{M} = \sum\nolimits_{q=1}^{20} \eta^q \mathbf{h}^q (\mathbf{h}^q)^T$, i.e., $\mu(\alpha^k, \alpha^l) = \sum\nolimits_{q=1}^{20} \eta^q h_k^q h_l^q$. It follows from this equality that all the amino acids α^k may be represented by vectors $\mathbf{a}^k = (a_1^k \cdots a_{20}^k)^T = \big((\eta^1)^{1/2} h_1^k \cdots (\eta^{20})^{1/2} h_{20}^k\big)^T \in \mathbb{R}^{20}$, whose inner products completely coincide with elements of the substitution matrix $\mu(\alpha^k, \alpha^l) = (\mathbf{a}^k)^T \mathbf{a}^l$.

Thus, from the viewpoint of a specified substitution matrix, any initially discrete symbolic fragment of the amino acid chain $\omega = (\alpha_\tau \in \mathbb{A}, -T \le \tau \le T)$ may be considered as a real-valued 20-dimensional signal $(\mathbf{a}_\tau = (a_{1\tau} \cdots a_{20\tau})^T \in \mathbb{R}^{20}, -T \le \tau \le T)$. The idea is then to represent each scalar component $(a_{k\tau} \in \mathbb{R}, -T \le \tau \le T)$ of this vector signal $i = 1,...,20$ in the form of the vector of Fourier coefficients with respect to the pairs of orthogonal basic harmonic signals, $\cos\left(i(\pi/T)\tau\right)$ and $\sin\left(i(\pi/T)\tau\right)$, of incrementing frequency $\left\{i(\pi/T), i=0,1,...,T\right\}$ in the interval $-T\leq\tau\leq T$.

Let $(a_\tau\in\mathbb{R}, -T\leq\tau\leq T)$ be a scalar signal. Its cosine and sine spectra are expressed by the following formulas:

$$\begin{cases} u_0=\frac{1}{2T+1}\sum_{\tau=-T}^{T}a_\tau, i=0, \quad u_l=\frac{1}{2T+1}\sum_{\tau=-T}^{T}a_\tau\cos\left(l(\pi/T)\tau\right), l=1,...,T, \\ v_l=\frac{1}{2T+1}\sum_{\tau=-T}^{T}a_\tau\sin\left(l(\pi/T)\tau\right), l=1,...,T. \end{cases} \tag{9}$$

To partially dampen the dependence of the Fourier expansion on the shift of the sliding window along the polypeptide axis, we take into account only $T+1$ elements of the amplitude spectrum and ignore the phase of the Fourier transform:

$$f_0=u_0, l=0, \quad f_l=(u_l^2+v_l^2)^{1/2}, \ l=1,...,T. \tag{10}$$

An amino acid fragment $\omega=(\alpha_\tau\in\mathbb{A}, -T\leq\tau\leq T)$ will yield a vector signal $\left(\mathbf{a}_\tau=(a_{1\tau}\cdots a_{20\tau})^T, -T\leq\tau\leq T\right)$ and, respectively, 20 spectra represented by the $T+1$ 20-dimensional vectors $l=0,1,...,T$ corresponding to the series of increasing frequencies in accordance with (9) and (10). In this work, we exploit four first harmonics along with the zero-frequency constant:

$$\begin{aligned} &\mathbf{f}(\omega)=\left[\mathbf{f}_0(\omega), \mathbf{f}_l(\omega), l=1,...,4\right]\in\mathbb{R}^{20\times5}=\mathbb{R}^{100}, \\ &\mathbf{f}_0(\omega)=\left(f_{k0}(\omega), k=1,...,20\right), \mathbf{f}_l(\omega)=\left(f_{kl}(\omega), k=1,...,20\right). \end{aligned} \tag{11}$$

The essence of the Fourier-transform-based comparison of amino acid fragments (ω',ω'') is thus exploitation of the feature vector (11) within a single comparison modality $S(\omega',\omega'')=S\left(\mathbf{f}(\omega'),\mathbf{f}(\omega'')\right)$. We examine here three comparison functions numbered as continuation of (8):

$$\begin{aligned} &S_4(\omega',\omega'')=\mathbf{f}^T(\omega')\mathbf{f}(\omega'')=\sum_{k=1}^{20}f_{k0}(\omega')f_{k0}(\omega'')+\sum_{k=1}^{20}\sum_{l=1}^{4}f_{kl}(\omega')f_{kl}(\omega''), \\ &S_5(\omega',\omega'')=\exp\left\{-\gamma\|\mathbf{f}(\omega')-\mathbf{f}(\omega'')\|^2\right\}= \\ &\qquad \exp\left\{-\gamma\left[\sum_{k=1}^{20}\left(f_{k0}(\omega')-f_{k0}(\omega'')\right)^2+\sum_{k=1}^{20}\sum_{l=1}^{4}\left(f_{kl}(\omega')-f_{kl}(\omega'')\right)^2\right]\right\}, \\ &S_6(\omega',\omega'')=\sum_{k=1}^{20}\left|f_{k0}(\omega')-f_{k0}(\omega'')\right|+\sum_{k=1}^{20}\sum_{l=1}^{4}\left|f_{kl}(\omega')-f_{kl}(\omega'')\right|. \end{aligned} \tag{12}$$

This class of fragment comparison functions is meant to be appropriate for exploring long-range dependencies in protein secondary structure prediction. With this purpose, we use relatively large window length $2T+1=35$.

5 Experiments

To determine the performance of the multimodal Relevance Vector Machine in the context of protein secondary structure prediction at different levels of relevance-selection for amino acid fragments and fragment comparison functions, we used the RS126 data set that contains 126 proteins having less than 25% sequence identity for lengths greater than 80 amino acids.

All in all, the proteins in RS126 produce the set Ω of $|\Omega|=19075$ amino acid windows $\omega\in\Omega$ of length $2T+1=35$, each labelled by an index of the structural state at the center; $y=\pm 1$, i.e., strand/not-strand. We performed four experiments with this data set.

In each experiment, we independently partitioned the set of all amino acid windows into the training set $\Omega_{tr}\subset\Omega$ of size $N=|\Omega_{tr}|=1600$ randomly drawn from Ω, and the rest $\Omega_{test}=\Omega\backslash\Omega_{tr}$ of size $|\Omega_{test}|=17475$ which served as the source of test sets.

The set of six competing and concurrent fragment comparison functions remained the same, being those derived via functions (8) and (12), $n=6$. The Fourier-transform-based comparison of amino acid windows (12) utilizes the full length of the windows $2T+1=35$, whereas the amino-acid-position-based comparison (8), in accordance with the accepted strategy, is to be applied to shorter windows $2T+1=13$ obtained from the initial ones by ignoring the 11 amino acids at both ends.

Each of the four experiments consisted in training the multi-modal Relevance Vector Machine (5) seven times from the same training set Ω_{tr}, $N=1600$, with seven incrementing values of the selectivity parameter: $\mu_1=0$, $\mu_2=0.3$, $\mu_3=0.5$ $\mu_4=0.6$, $\mu_5=0.8$, $\mu_6=0.9999$, and $\mu_7=0.99999$ $(\mu\to 1)$. Thus, the Relevance Vector Machine was run $4\times 7=28$ times.

The immediate result of each run of the training algorithm with a heuristic initial value of the selectivity parameter μ is the subset of relevant secondary features $\hat{F}(\mu)=\left\{ij: \hat{a}_{ij}(\mu)\neq 0\right\}\subseteq F$ and parameter values $\left(a_{ij}(\mu)\in\hat{F}, b(\mu)\right)$ of the discriminant hyperplane (7). Of particular importance are the resulting subsets of relevant objects (amino acid fragments of the training set), $\hat{J}(\mu)=\left\{j: \exists i(a_{ij}\neq 0)\right\}\subseteq\{1,...,N\}$, and relevant comparison modalities $\hat{I}(\mu)=$ $\left\{i: \exists j(a_{ij}\neq 0)\right\}\subseteq I=\{1,...,n\}$. Their numbers are denoted, respectively, as $\hat{N}(\mu)=$ $|\hat{J}(\mu)|\leq N$ and $\hat{n}(\mu)=|\hat{I}(\mu)|\leq n$.

We then randomly partitioned the remaining set $\Omega_{test}=\Omega\backslash\Omega_{tr}$ of 19075 amino acid windows into 10 test sets of approximately 1900 windows each, and computed the accuracy of recognition of secondary structure states $\{s,\bar{s}\}$, i.e., {strand} versus {not strand}, as the respective percentage values. The overall percentage accuracy in all the test sets for each selectivity μ_k, $k=1,...,7$, was assessed by the average value $Acc(\mu)$ and root-mean-square scatter $\sigma(\mu)$. Finally, the confidence interval was computed for each average percentage as $Acc(\mu)\pm 2\sigma(\mu)$.

Figure 1 visually displays the dependence of the accuracy percentage $Acc(\mu)$ and the number of relevant amino acid fragments participating in the final decision rule $\hat{N}(\mu)$ at selectivity level μ. All the results are represented in Table 1.

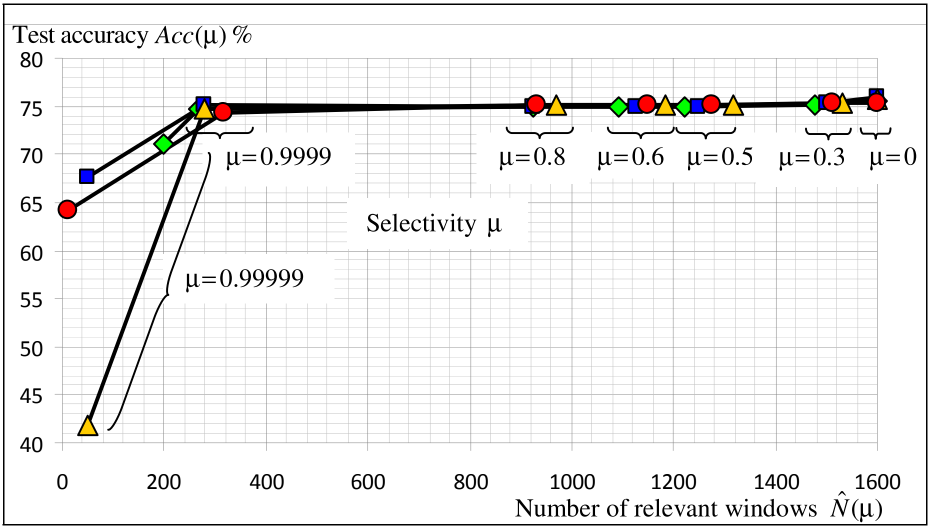

Fig. 1. Experimental dependence of the number of relevant amino acid fragments $\hat{N}$ and the test-set accuracy of detecting strands Acc on the level of secondary feature selectivity μ

It is evident from Table 1 and Figure 1 that, in all experiments, the best average accuracy of approximately 75.5% is achieved with zero selectivity $\mu=0$, when all the 1600 amino acid fragments constituting the training set and all the 6 comparison functions participate in the discriminant hyperplane (7) (which is thus defined in the $nN=9600$-dimensional space of secondary features of a single amino acid window $\omega=(\alpha_\tau\in\mathbb{A}, -T\leq\tau\leq T)$). What is especially interesting is that no traces of overfitting are evident in the determination of the discriminant hyperplane in the linear space of secondary feature vectors, $\mathbf{x}(\omega)=\left(x_{ij}(\omega), i=1,...,6, j=1600\right)$, whose dimension, $\mathbf{x}(\omega)=\mathbb{R}^{9600}$, exceeds, by six times, the size of the training set.

The growth of μ thus diminishes both the number, $\hat{N}(\mu)$, of relevant training-set fragments and the number, $\hat{n}(\mu)$, of relevant fragment-comparison modalities forming the secondary features of current amino acid windows, and initially results in a minor decrease of the test-set accuracy. However, it is worth noting that the accuracy percentage remains practically the same in all independent experiments up to the selectivity level $\mu=0.9999$, when about 300 relevant amino acid fragments of the initial number of 1600 remain in the decision rule for strand detection, and only $\hat{n}=3$ comparison functions are required to classify new windows in the test set. The respective drop of accuracy relative to the absence of any selectivity $\mu=0$ does not exceed 1%.

Table 1. Results of four independent experiments (markers as in Figure 1)

			Accuracy of detecting strands $Acc(\mu)$	Number of relevant windows $\hat{N}(\mu)$	Number $\hat{n}(\mu)$ and list of relevant comparison functions
Experiment 1 ◆	Selectivity μ	0	75.63 ± 1.78%	1600	$6, \hat{I}=\{1,2,3,4,5,6\}$
		0.3	75.04 ± 1.75%	1476	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.5	74.95 ± 1.74%	1222	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.6	74.96 ± 1.74%	1094	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.8	74.96 ± 1.72%	924	$4, \hat{I}=\{1,2,\cancel{3},4,5,\cancel{6}\}$
		0.9999	74.63 ± 1.76%	267	$3, \hat{I}=\{1,2,\cancel{3},\cancel{4},5,\cancel{6}\}$
		0.99999	71.04 ± 1.69%	200	$2, \hat{I}=\{1,2,\cancel{3},\cancel{4},\cancel{5},\cancel{6}\}$
Experiment 2 ■	Selectivity μ	0	75.85 ± 1.52%	1600	$6, \hat{I}=\{1,2,3,4,5,6\}$
		0.3	75.23 ± 1.72%	1501	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.5	75.01 ± 1.63%	1247	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.6	75.01 ± 1.65%	1127	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.8	75.01 ± 1.65%	924	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.9999	75.10 ± 1.73%	278	$3, \hat{I}=\{1,2,\cancel{3},\cancel{4},5,\cancel{6}\}$
		0.99999	67.60 ± 0.80%	49	$3, \hat{I}=\{1,2,\cancel{3},\cancel{4},5,\cancel{6}\}$
Experiment 3 ◀	Selectivity μ	0	75.70 ± 1.22%	1600	$6, \hat{I}=\{1,2,3,4,5,6\}$
		0.3	75.30 ± 0.79%	1531	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.5	75.10 ± 0.94%	1317	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.6	75.08 ± 0.99%	1183	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.8	75.08 ± 0.99%	971	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.9999	74.74 ± 0.79%	280	$3, \hat{I}=\{1,2,\cancel{3},\cancel{4},5,\cancel{6}\}$
		0.99999	41.84 ± 2.33%	51	$3, \hat{I}=\{1,2,\cancel{3},\cancel{4},5,\cancel{6}\}$
Experiment 4 ●	Selectivity μ	0	75.33 ± 0.99%	1600	$6, \hat{I}=\{1,2,3,4,5,6\}$
		0.3	75.30 ± 0.95%	1514	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.5	75.07 ± 0.97%	1275	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.6	75.03 ± 0.99%	1150	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.8	75.03 ± 0.99%	933	$5, \hat{I}=\{1,2,\cancel{3},4,5,6\}$
		0.9999	74.27 ± 1.53%	318	$3, \hat{I}=\{1,2,\cancel{3},\cancel{4},5,\cancel{6}\}$
		0.99999	64.16 ± 1.81%	12	$3, \hat{I}=\{1,2,\cancel{3},\cancel{4},5,\cancel{6}\}$

Beyond this limit, a further increase of selectivity results in a drastic loss of both recognition accuracy and stability with respect to different training sets.

6 Conclusions

Application of the machine learning techniques to the problems of bioinformatics, in particular feature generation and selection in the space of amino acid sequences, represents a fruitful direction of research both in computer science and in computational biology. In this proof-of-principle study, we applied a method based on the Relevance Vector Machines (RVM) methodology to the problem of the protein secondary structure prediction. A unique characteristic of this method is that it permits automatic selection of the most appropriate features (modalities) from the total number of possible modalities.

In our study, the average accuracy of the strand prediction was approximately 75%, a comparable accuracy to the current state-of-the-art. However, the use of relevance vector principles means that this accuracy figure is achievable with only a small fraction (less than a quarter) of the totality of features, representing a potentially significant advantage in terms of parsimony, robustness and interpretability of the resulting classifications.

Acknowledgments. We would like to acknowledge support from grants of the Russian Foundation for Basic Research 11-07-00409 and 11-07-00728, grant of the Ministry of Education and Science of the Russian Federation, contract 07.514.11.4001, and from UK EPSRC grant EP/F069626/1 (ACASVA).

References

1. Branden, C., Tooze, J.: Introduction to Protein Structure, 2nd edn., p. 410. Garland Publishing, Inc., New York (1999)
2. Rost, B.: Protein secondary structure prediction continues to rise. Journal of Structural Biology 134(2-3), 204–218 (2001)
3. Yoo, P., Zhou, B., Zomaya, A.: Machine learning techniques for protein secondary structure prediction: An overview and evaluation. Current Bioinformatics 3(2), 74–86 (2008)
4. Critical Assessment of the Protein Structure Prediction. Protein Structure Prediction Center. Sponsored by the US National Library of Medicine (NIH/NLM), `http://predictioncenter.org/http://predictioncenter.org/index.cgi?page=proceedings`
5. Aloy, P., Stark, A., Hadley, C., Russell, R.: Predictions without templates: new folds, secon-dary structure, and contacts in CASP5. Proteins 53(suppl. 6), 436–456 (2003)
6. Torshin, I.Y.: Bioinformatics in the Post-Genomic Era: The Role of Biophysics. Nova Biomedical Books, NY (2007) ISBN: 1-60021-048
7. Vapnik, V.: Statistical Learning Theory, p. 736. John-Wiley & Sons, Inc. (1998)
8. Ward, J., McGuffin, L., Buxton, B., Jones, D.: Secondary structure prediction with support vector machines. Bioinformatics 19(13), 1650–1655 (2003)

9. Bishop, C., Tipping, M.: Variational Relevance Vector Machines. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, pp. 46–53. Morgan Kaufmann (2000)
10. Seredin, O., Mottl, V., Tatarchuk, A., Razin, N., Windridge, D.: Convex Support and Relevance Vector Machines for selective multimodal pattern recognition. In: The 21th International Conference on Pattern Recognition, Tsukuba Science City, Japan, November 11-15 (2012)
11. Engel, D., DeGrado, W.: Amino acid propensities are position-dependent throughout the length of α-helices. J. Mol. Biol. 337, 1195–1205 (2004)
12. Ni, Y., Niranjan, M.: Exploiting long-range dependencies in protein β-sheet secondary structure prediction. In: Proceedings of the 5th IAPR International Conference on Pattern Recognition in Bioinformatics, Nijmegen, The Netherlands, September 22-24, pp. 349–357 (2010)
13. Cole, C., Barber, J., Barton, G.: The Jpred 3 secondary structure prediction server. Nucl. Acids Res. 36 (suppl. 2), W197–W201 (2008)
14. Duin, R., Pekalska, E., de Ridder, D.: Relational discriminant analysis. Pattern Recognition Letters 20, 1175–1181 (1999)
15. Wang, L., Zhu, J., Zou, H.: The doubly regularized support vector machine. Statistica Sinica 16, 589–615 (2006)
16. Dayhoff, M., Schwarts, R., Orcutt, B.: A model of evolutionary change in proteins. Atlas of Protein Sequences and Structures 5(suppl. 3), 345–352 (1978)
17. Sulimova, V., Mottl, V., Kulikowski, C., Muchnik, I.: Probabilistic evolutionary model for substitution matrices of PAM and BLOSUM families. DIMACS Technical Report 2008-16, Rutgers University, p. 17 (2008)