BIOINFORMATICS

Vol. 15 no.5 1999
Pages 413–421

Improved performance in protein secondary structure prediction by inhomogeneous score combination

Y. Guermeur[1], C. Geourjon[2], P. Gallinari[3] and G. Deléage[2]

[1]LIP, Ecole Normale Supérieure de Lyon,46, Allée d'Italie, 69364 Lyon cedex 07, [2]Institut de Biologie et Chimie des Protéines, 7, passage du Vercors, 69367 Lyon cedex 07 and [3]LIP6, Université Pierre et Marie Curie, Tour 46-00, Boîte 169, 4, Place Jussieu, 75252 Paris cedex 05, France

Received on June 1, 1998; revised on January 4, 1999; accepted on February 17, 1999

Abstract

Motivation: *In many fields of pattern recognition, combination has proved efficient to increase the generalization performance of individual prediction methods. Numerous systems have been developed for protein secondary structure prediction, based on different principles. Finding better ensemble methods for this task may thus become crucial. Furthermore, efforts need to be made to help the biologist in the post-processing of the outputs.*

Results: *An ensemble method has been designed to post-process the outputs of discriminant models, in order to obtain an improvement in prediction accuracy while generating class posterior probability estimates. Experimental results establish that it can increase the recognition rate of protein secondary structure prediction methods that provide inhomogeneous scores, even though their individual prediction successes are largely different. This combination thus constitutes a help for the biologist, who can use it confidently on top of any set of prediction methods. Moreover, the resulting estimates can be used in various ways, for instance to determine which areas in the sequence are predicted with a given level of reliability.*

Availability: *The prediction is freely available over the Internet on the Network Protein Sequence Analysis (NPS@) WWW server at http://pbil.ibcp.fr/NPSA/npsa_server.html. The source code of the combiner can be obtained on request for academic use.*

Contact: *Neural network and ensemble method: Yann.Guermeur@ens-lyon.fr; server and prediction methods: g.deleage@ibcp.fr*

Introduction

The idea of combining models in order to improve performance is well known in statistics and has a long theoretical background (Bates and Granger, 1969; Dickinson, 1973, 1975; LeBlanc and Tibshirani, 1993; Jordan and Jacobs, 1994; Peng *et al*., 1994). The main research issues, for which operational solutions have been proposed, deal with the development of experts with different behaviours, the optimal use of small samples and the design of robust combiners. Theoretical evidence has made rapid strides in forecasting and regression, whereas the specificities of discrimination have seldom been taken into account (Tumer and Ghosh, 1995). Discrimination has many applications in biology, such as the recognition of genes in DNA, or protein structure prediction. For these applications, many classifiers are already available. They have been conceived by different research groups, and they rely on different techniques and on the expertise of these groups. It is thus interesting to develop ensemble methods which allow the combination of existing classifiers so as to improve their recognition rate. This is the subject of the present paper. We present a general system dedicated to the combination of different classifiers operating on sequences. The system is evaluated on an open problem in predictive structural biology: the prediction of protein secondary structure. The prediction of protein structure is perhaps one of the most focused-on questions in molecular biocomputing activities and numerous methods have been developed to predict secondary structure (for reviews, see Eisenhaber *et al*., 1995; Rost and O'Donoghue, 1997). In addition to the amino acid sequences, they ordinarily use data from different knowledge sources (physicochemical properties, homology, etc.). Consequently, whenever secondary structure is to be predicted, several sets of conformational scores are available, each of which can be considered a priori as useful. Most of the current best prediction methods already implement conformational score combination at one stage or another. This combination can take many forms, ranging from the simple linear opinion pool (Rost and Sander, 1993) to the more complex non-linear regression schemes performed by neural networks (Zhang *et al*., 1992; Riis and Krogh, 1996). Symbolic methods based on empirical results have also been implemented, such as the algorithm combine (Biou *et al*., 1988). However, the choice of a particular combiner is hardly ever justified, although it appears to have a crucial effect on performance. Furthermore, the scores combined are systematically homogeneous, i.e. they

represent estimates of the same quantities, whereas the practitioner who needs to make his own prediction based on the results of several methods has most often to deal with inhomogeneous scores.

Our combination method is based on multivariate linear regression (MLR). It allows the combination of inhomogeneous scores and ensures that the resulting prediction will be better than those of the individual methods. Its originality rests on the fact that it estimates the posterior probability of classes, while being simpler than a neural network and far more efficient than a weighted average. This can be done under weak hypotheses regarding the predictors (the experts' outputs). These hypotheses are non-restrictive, as will be shown here. The method thus offers a good compromise between simplicity and efficiency. It is also possible to establish tight distribution-independent bounds on its generalization ability. To our knowledge, this is the first efficient ensemble method for which such bounds have been derived.

We first introduce the formal setting of the combination problem and explain how the MLR model can be constrained to solve it provided the predictors are preliminary adequately processed. A neural network is described which performs this pre-processing. We then mention some appealing properties of the model concerning its generalization ability. Finally, we turn to its application for protein secondary structure prediction. We describe experiments which highlight the relevance of our approach for this problem, irrespective of the nature of the scores provided by the experts combined. They establish that combining two of the best methods, namely PHD (Profile network from HeiDelberg) (Rost and Sander, 1993, 1994) and SOPMA (Self-Optimized Prediction from Multiple Alignment) (Geourjon and Deléage, 1995), can generate a statistically significant increase in prediction accuracy. The important question of the benefits that can be expected in structure prediction from class posterior probability estimation is also studied.

Systems and methods

All programs have been written in C ANSI for maximal speed and portability. They are integrated into one single treatment which successively performs the acquisition of the conformational scores provided by different prediction methods and uses them to generate initial class posterior probability estimates from which the final probabilities of class membership are derived, by linear combinations. This process can be invoked from within the NPS@ (Network Protein Sequence Analysis) server. All the secondary structure prediction methods have been implemented as described by the authors (Geourjon and Deléage, 1995; Garnier *et al.*, 1996; Levin, 1997), and are available through the WWW server, except for PHD. For this method, the output files of the leave-one-out cross-validation procedure have been obtained directly from B.Rost.

Algorithm

Problem specification

We consider the case of Q-category discrimination tasks, under the assumption that for each input pattern or instance x, the outputs $f_j(x) = [f_{jk}(x)]$, $(1 \le j \le P)$, $(1 \le k \le Q)$ of P classifiers are available. $D = \{(x_i, y_i)\}$, $(1 \le i \le N)$ denotes the training set. y_i is the canonical coding of the category of x_i, i.e. $y_i = [\delta_{kl}]$, $(1 \le k \le Q)$ when $x_i \in C_l$, where δ is Kronecker's symbol. Let $F(x) = [f_1(x)^{\mathrm{T}}, \ldots, f_j(x)^{\mathrm{T}}, \ldots, f_P(x)^{\mathrm{T}}]^{\mathrm{T}} \in IR^{PQ}$ be the vector of predictors for example x [$f_j(x)^{\mathrm{T}}$ is the transpose of vector $f_j(x)$]. In this context, we address the following problem.

Problem 1. Given a set of models trained independently to perform a multiclass discrimination task, combine the scores they generate in order to produce class posterior probability approximations corresponding, when processed with Bayes' estimated decision rule, to an improved recognition rate.

Multivariate linear regression on class posterior probability estimates

We solve Problem 1 under the additional hypothesis that the outputs of the models are class posterior probability estimates, i.e. $\forall (j,k), f_{jk}(x) = \hat{p}_j(C_k|x)$. They are thus non-negative and sum to one. We will see in the next section that this hypothesis induces no restriction on the nature of the methods that can be combined. To satisfy the requirements of Problem 1, the components of the function g resulting from the combination must themselves be non-negative and sum to one. In what follows, we briefly describe how the MLR model is implemented to compute such a function [details of proofs can be found in Guermeur (1997)].

The MLR model studied here, parameterized by $v = [v_1^T, \ldots, v_k^T, \ldots, v_Q^T]^T \in IR^{PQ^2}$, computes the function g given by:

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_k(x) \\ \vdots \\ g_Q(x) \end{bmatrix} = \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \\ \vdots \\ v_Q^T \end{bmatrix} F(x) \qquad (1)$$

i.e. $$g_k(x) = \sum_{l=1}^{l=P} \sum_{m=1}^{m=Q} v_{klm}\, f_{lm}(x) = \hat{p}(C_k|x), (1 \le k \le Q).$$

Maximizing directly the empirical recognition rate with respect to v is difficult. In this work, we optimize the 'quality' of the class posterior probability estimates. This choice is grounded on the fact that both criteria are asymptotically equivalent. Thus, we consider the set Λ of convex loss functions which ensure that the model outputs can be interpreted as probabilities. It includes as special cases the quadratic and cross-entropic losses (see Bishop, 1995). The training prob-

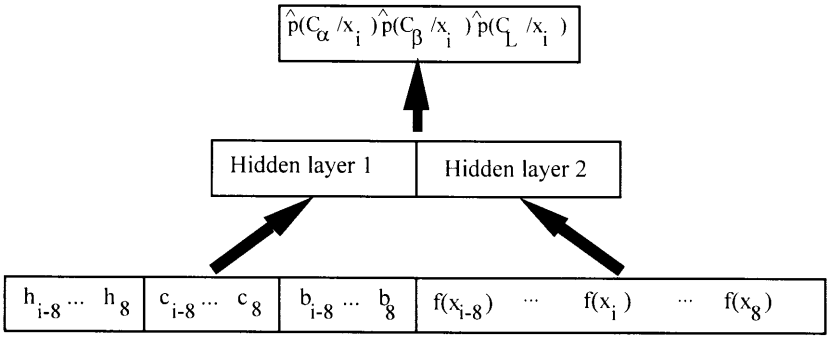

Fig. 1. Architecture of the network used to post-process the outputs of each of the prediction methods which do not estimate the class posterior probabilities. Vector $[f(x_{i-8})^T,\ldots f(x_i)^T,\ldots, f(x_{i+8})^T]^T \in IR^{51}$ contains the original scores for a segment of size 17. Letters *h*, *c* and *b*, respectively, represent the hydrophobicity, charge and bulk of the residues. Two distinct sets of hidden units are devoted to the two types of data, which makes it easier for the network to take them both into account.

lem can then be reformulated simply owing to the following proposition, which holds irrespective of the nature of the objective function *J*.

Proposition 1. Optimal solutions to Problem 1 are obtained by restricting the set of feasible solutions to the convex set *V* with:

$$V = \left\{ v \in IR_+^{PQ^2} / \left\{ \begin{array}{c} \sum_{k=1}^{k=Q} (v_{klm} - v_{klQ}) = 0, (1 \le l \le P), (1 \le m \le Q-1) \\ {}^T 1_{PQ^2} v = Q \end{array} \right\} \right\} \quad (2)$$

V is the restriction to the non-negative orthant of the intersection of a set of hyperplanes in IR^{PQ^2}. Details on the way this set is obtained can be found in Guermeur and Paugam-Moisy (1998).

Since both the objective function *J* and the constraints are convex functions of *v*, the training procedure amounts to solving the following simple convex programming problem.

Problem 2. Given a convex loss function $L \in \Lambda$ and a training set *D*, find in the family of MLR functions *g* a combiner minimizing the empirical cost $\hat{J}(v) = \sum_D L(y_i, g(x_i, v))$, under the constraint $v \in V$.

To sum up, solving Problem 2 with any convex programming algorithm provides us with a statistically rigorous solution to the problem of combining classifiers estimating the class posterior probabilities. In our experiments, we chose the gradient projection method (Rosen, 1960). The active set method (see Fletcher, 1991) incorporated in this algorithm was used to implement a procedure of early stopping (Bishop, 1995). The training criterion was cross-entropy and the regression was performed on a window of $T = 3$ consecutive residues (the model thus had $PQT = 9P$ inputs).

Deriving class posterior probability estimates by non-linear regression on conformational scores

We mentioned in the Introduction that the outputs of protein secondary structure prediction methods were seldom probability estimates. In order to be workable by the MLR combiner, they must be preliminary processed. To perform this task, a structure-to-structure filtering neural network introduced in Guermeur and Gallinari (1996), and described in Figure 1, is used. Its inputs are the conformational scores provided by a prediction method for a segment of size 17 centred on the amino acid to be assigned, plus the coding of physicochemical properties of the corresponding residues. Its outputs are the desired probability estimates. The physicochemical properties are the hydrophobicity as defined in Eisenberg *et al.* (1987), the charge and bulk, quantified in accordance with the scaling proposed in Taylor (1986). As for the MLR combiner, the training criterion is cross-entropy.

Generalization performance

Although the learning process amounts to estimating probability distributions on a finite dataset, the criterion of interest, the generalization ability in terms of recognition rate, can be concomitantly estimated and controlled. The results we established on that subject (Guermeur and Paugam-Moisy, 1998; Guermeur *et al.*, 1998) rest on one fundamental theorem of the statistical learning theory developed by Vapnik (1995): the theorem of uniform convergence of relative frequencies of events to their probabilities. Giving details would go beyond the scope of this article. Suffice it to say that, provided the weak hypotheses of statistical learning apply to biological sequences, propositions of practical interest can be established, such as the following.

Proposition 2. In both experimental set-ups described in the following section, with probability exceeding 0.95, the generalization performance and the training performance of the MLR combiner differ by <1%.

Proposition 2 deals with the 'true' generalization performance and not a sample-based estimate. Furthermore, it only depends on the model complexity and the size of the training sample, not on the distribution of the data, i.e. on the nature of the problem. This makes our combiner the first efficient ensemble method for which tight bounds on the generalization performance have been derived in a real-life experimental set-up. Such bounds are useful since they allow one to keep all the data for training and simply 'extrapolate' the generalization error from the observed one.

Assessment of the method

Experimental protocol

To assess our combiner, we selected two reference secondary structure databases: the Rost and Sander (1993) database, which contains 126 chains of soluble proteins sharing <25% identity, and its extension to 629 chains (Hobohm and Sander, 1994). In both cases, the secondary structure assignment was carried out according to DSSP (Dictionary of Secondary Structures of Proteins) (Kabsch and Sander, 1983). The difference in size dictated the choice of two distinct experimental procedures.

The first database has been divided by its authors into seven disjoint parts of roughly equal size [see Riis and Krogh (1996) for details]. This splitting is retained here to implement a two-stage cross-validation procedure. Each subset is iteratively used as test set. With the six remaining training sets, a variant of stacked generalization (Wolpert, 1992) is applied, in which the initial leave-one-out cross-validation procedure is replaced with a more computationally efficient 6-fold cross-validation. This implementation, also suboptimal, has been observed not to deteriorate the generalization performance, which is consistent with other results, for instance those reported in Breiman (1996). The PDBSELECT extension is more than six times larger than the former release in terms of residues. We took advantage of this abundance of data to implement a simpler experimental procedure. The database is also split into seven subsets of approximately the same size. As before, each subset is iteratively used for testing. The training set is divided into two halves. The first one is used to train the filtering networks associated with each of the classifiers. The combiners are trained on the sequences of the second one.

Choice of classifiers and combiners

We conceived the experiments in order to characterize the behaviour of the combiner with respect to two factors: its capacity to improve the recognition rate of the current best prediction methods and its robustness, i.e. its faculty to maintain its accuracy even in the case of a pool of experts with high disparity in performance. To that end, we chose two different sets of classifiers. In a first series of experiments, the combination of PHD and SOPMA was assessed on the database of Rost and Sander. Precisely, two sets of posterior probability estimates were derived from the scores provided by SOPMA by means of two different structure-to-structure networks. Although the initial outputs of the PHD method are actually class posterior probability estimates, we only had access in this study to a degenerate coding of its predictions: the nature of the conformational states predicted and the values of the reliability index as defined in Rost and Sander (1993). These data were used to compute rough initial approximations of the posterior probabilities. Since they already appeared to be of reasonable quality, we included them in the combination, with two other sets provided by two new filtering networks. On the whole, five classifiers were thus combined. This first series of experiments was devoted to the assessment of the highest accuracy reachable. In the second one, performed with the extended database, SOPMA was combined with the GOR IV method [Garnier–Osguthorpe–Robson (prediction method); Garnier *et al*., 1996], which uses the formalism of the information theory, and SIMPA96 (where SIMPA is SIMilarity Peptide Analysis) (Levin, 1997), a nearest-neighbour method. As can be seen in Table 2, the difference in accuracy after filtering between SIMPA96 and GOR IV amounts to 2.9% in recognition rate, so that the diversity of the methods was not a priori sufficient to warrant the success of the combination.

Several other ensemble methods were implemented for comparison. We only report here the results obtained with two methods lying at opposite ends of the spectrum in terms of capacity: majority voting and a single hidden layer perceptron. In case equality occurs in majority voting, the conformational state selected is the one with the highest average estimated posterior probability.

Prediction accuracy

The prediction accuracy has been assessed by means of four different measures: (i) the percentage of correctly predicted residues Q_3 for a three-state description of secondary structure (helix, extended and aperiodic); (ii) Pearson's/Matthews' correlation coefficient C (Matthews, 1975); (iii) the segment overlap measure Sov (Rost *et al*., 1994); (iv) the standard deviation in the secondary structure content σ.

Table 1. Compared accuracy of the ensemble methods on the Rost and Sander database

	SOPMA	SOPMA+	PHD	PHD+	NN	Vote	MLR
Q_3	68.9	69.5	71.5	71.4	71.5	72.0	72.4
C_α	0.55	0.58	0.62	0.62	0.62	0.63	0.63
C_β	0.49	0.48	0.52	0.52	0.52	0.52	0.52
C_c	0.48	0.48	0.51	0.51	0.51	0.52	0.52
Sov	0.72	0.70	0.73	0.72	0.72	0.73	0.73
Sov_α	0.74	0.74	0.75	0.73	0.73	0.75	0.76
Sov_β	0.72	0.67	0.72	0.70	0.72	0.72	0.70
Sov_c	0.70	0.70	0.72	0.70	0.71	0.71	0.72
σ_α	11.5	11.0	11.7	11.4	11.1	11.0	10.6
σ_β	11.6	12.2	11.2	11.4	11.1	11.2	11.5
σ_c	11.7	12.6	12.0	12.3	12.6	12.0	12.9

Columns 2–5 contain the statistics for individual classifiers. Methods post-processed with the network of Figure 1 are designated by their name followed by a plus. The last three columns give the results of the combination with a one hidden layer perceptron (neural network, NN), majority voting and our MLR combiner.

Table 2. Relative accuracies of the ensemble methods on the PDBSELECT database

	GOR IV	GOR IV+	SOPMA	SOPMA+	SIMPA	SIMPA+	Vote	NN	MLR
Q_3	64.1	66.5	68.4	69.7	69.2	69.4	70.2	71.2	71.3
C_α	0.47	0.51	0.55	0.58	0.56	0.57	0.59	0.60	0.60
C_β	0.39	0.43	0.48	0.49	0.49	0.49	0.49	0.52	0.52
C_c	0.44	0.46	0.49	0.50	0.49	0.49	0.51	0.52	0.52
Sov	0.66	0.68	0.72	0.71	0.71	0.70	0.72	0.72	0.72
Sov_α	0.63	0.67	0.72	0.73	0.74	0.72	0.73	0.73	0.74
Sov_β	0.67	0.64	0.73	0.68	0.67	0.66	0.69	0.70	0.68
Sov_c	0.68	0.70	0.72	0.72	0.72	0.71	0.73	0.73	0.73
σ_α	13.9	12.5	10.8	10.7	10.8	10.6	10.5	10.1	10.3
σ_β	11.5	11.6	10.3	11.1	11.2	10.7	10.3	10.1	10.9
σ_c	9.4	10.1	9.9	10.6	11.6	11.1	10.1	10.5	11.4

Notations are those of Table 1. The classifiers combined are GOR IV, SOPMA and SIMPA96.

Results

The comparison of the predictive success of native methods and post-processed ones on the Rost and Sander database (1993) (columns 2–5 of Table 1) shows that the PHD accuracy is already maximal in the original method, whereas the Q_3 of the SOPMA method is improved by 0.6%. This improvement is more important (+1.3%) in the extended set (Table 2). Most of the gain affects the helix state.

The phenomenon obviously springs from the fact that SOPMA and the structure-to-structure network use the data in different ways. On the contrary, little can be expected from filtering the rough estimates derived from the prediction and reliability index of the connectionist method, apart from a restoration of the original probability approximations. When considering the combination, the MLR model consistently yields the best results even if the σ parameter is somewhat more unstable than for individual methods. It is noteworthy that α helix is predicted with the best correlation coefficients and Sov parameters. On the database of Rost and Sander, the accuracy of the prediction resulting from the combination is 72.4%, which is 0.9% higher than the accuracy of PHD. Under the hypotheses of normality classically used in the case of large samples, the difference is statistically significant with a confidence exceeding 0.96. Figures in Table 2 demonstrate the ability of SOPMA to generalize to large databases without noticeable loss in accuracy. They also illustrate the robustness of our combiner, which succeeds in increasing significantly the recognition rate even in the case where the spectrum of quality among the classifiers is wide. Precisely, the improvement in recognition rate over the best

	SOPMA	*SOPMA*	*PHD*	*PHD*	*PHD*
SOPMA	1.00	0.99	0.87	0.88	0.88
SOPMA	0.99	1.00	0.87	0.89	0.88
PHD	0.87	0.87	1.00	0.97	0.96
PHD	0.88	0.89	0.97	1.00	0.99
PHD	0.88	0.88	0.96	0.99	1.00

	GOR4	*SOPMA*	*SIMPA*
GOR4	1.00	0.89	0.91
SOPMA	0.89	1.00	0.89
SIMPA	0.91	0.89	1.00

Fig. 2. Empirical error correlation matrices of the classifiers used in the first series of experiments ($_1$ on the left) and the second one ($_2$ on the right). In order of increasing row/column indexes, the classifiers corresponding to $_1$ are two post-processed versions of SOPMA, PHD in its initial configuration and combined with two filtering networks. Similarly, the rows/columns of matrix $_2$ correspond to post-processed versions of GOR IV, SOPMA and SIMPA96.

individual method exceeds 1.9%, in a context where majority voting, for instance, appears clearly inappropriate. This phenomenon is even more prominent if additional 'weak' classifiers are used. To highlight it, we studied the incorporation in the combination of a hierarchical neural network classifier and a statistical module (Guermeur and Gallinari, 1996). Their respective recognition rates on the PDBSELECT database are 65.4 and 62.8%. The results of the connectionist combiner and the majority voting were both negatively affected by this inclusion (loss in accuracy of 0.2 and 1.5%, respectively), which had no effect at all on the MLR combination (the new classifiers were simply ignored). The gain in prediction is spread out on most proteins since there are as many as 397 over 629 chains that benefit from the combination and 54 for which the quality is even. The difference in raw performance compared with the first set of experiments obviously springs from the fact that the quality of the individual classifiers is lower. The leave-one-out predictions of PHD are not available for the largest database.

The main lesson that can be drawn from the analysis of the error correlation matrices of Figure 2 is that as different as the principles of the methods may be, errors remain highly correlated. This obviously induces a limit on the number of methods which can be combined profitably. A theoretical study on this matter can be found in Clemen and Winkler (1985). Indeed, in both series of experiments, we found it useless to increase the number of experts.

The results of the study of the recognition rate as a function of the highest value of the outputs are summarized in Figure 3. They corroborate the observations already made in Rost and Sander (1993, 1994) and Riis and Krogh (1996): this value is a good indicator of the confidence that one can have in the prediction. More precisely, there is a strong linear dependency between both criteria. One can make use of this property to determine anchor points useful in simulations of protein folding. Unfortunately, the confidence exceeds 90% for only 10.4% of the residues.

The MLR combination of different prediction schemes is available on the Internet (http://pbil.ibcp.fr/NPSA/npsa_mlr.html) by using the NPS@ server. A typical output

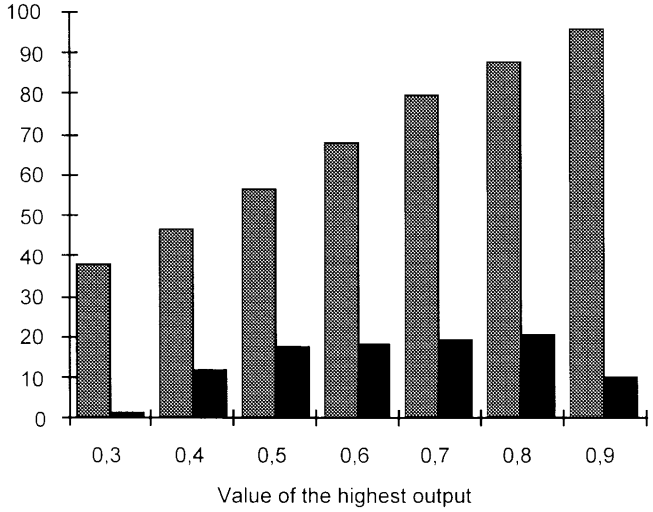

Fig. 3. Recognition rate (striped bars) and percentage of residues (black bars) as a function of the value of the highest output of the combiner. Predictions are those of the MLR combiner on the Rost and Sander database (see Table 1).

from the server is given in Figure 4 for the RBTR_KLEAE. The output comprises a colour-formatted text incorporated below the sequence. The conformational scores can be viewed directly by spawning the ANTHEPROT [ANalyse THE PROTeins (package)] (ftp.ibcp.fr) software with the chemical/x-antheprot MIME type. In the graphical viewer, a movable cursor displays the amino acid sequence (from –5 to +5 of the current cursor position), the predicted secondary structure state and the conformational scores for each state.

Discussion

Although the idea of combining protein secondary structure prediction methods is not new, its implementation remains difficult for the biologist. For instance, a fully satisfactory answer is still to be given to the fundamental question of the selection of the committee of experts. Making joint predic-

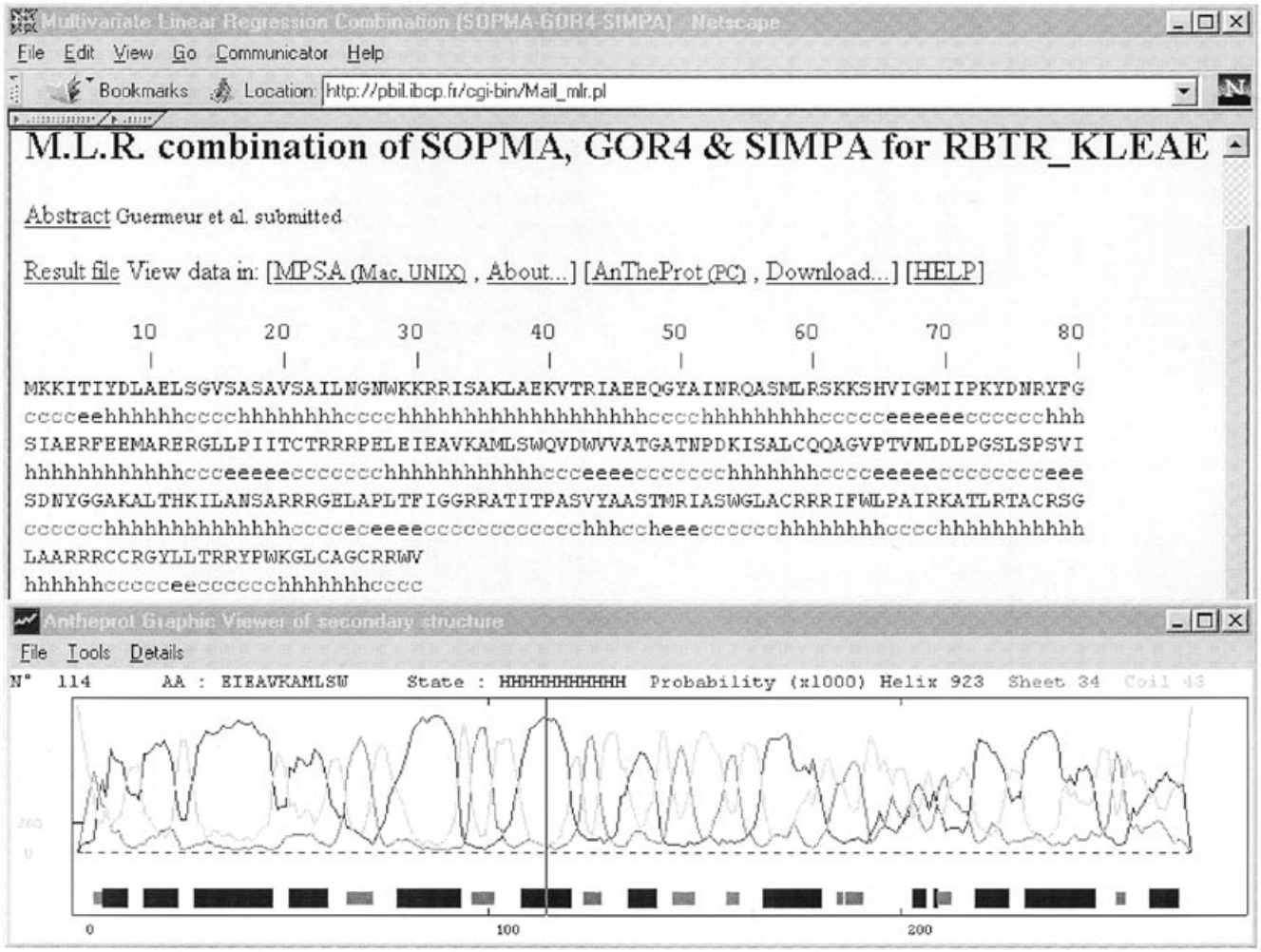

Fig. 4. The combination of three methods, SOPMA, GOR IV and SIMPA, has been performed by means of the multivariate linear regression on the ribitol operon repressor protein (RTBR_KLEAE code from SWISS-PROT 36.0) through our Web server at http://pbil.ibcp.fr/NPSA/npsa_mlr.html. Colour-coded outputs are provided with the prediction, as well as a synthetic graphic image of the result and a graphical display of the conformational scores (not shown). The result can be directly loaded within the AnTheProt 4.0 (graphic view of the figure) software by using the chemical/x-antheprot MIME type and by clicking on the appropriate link (View data in AnTheProt PC). The AnTheProt software (ftp.ibcp.fr) allows the user to move the cursor along the sequence and the conformational scores (probabilities multiplied by 1000) are given at the location of the cursor.

tions (considering residues predicted in the same conformational state by different methods) is helpful to stress regions that are predicted with high confidence (Zhang *et al*., 1992; Geourjon and Deléage, 1995), which makes an experimental determination of the structure considerably easier. However, this only concerns limited parts of the structure. In this paper, we have highlighted the fact that the user can always take advantage of the combination of methods, even when they provide scores of different natures or exhibit large differences in prediction accuracy. The main requirement to ensure an improvement lies in the choice of a combiner of appropriate complexity. It springs from the experimental results that hybrid secondary structure prediction methods can benefit from the use of ensemble methods more sophisticated than the common weighted averages. However, due to the limited size of the databases currently available, their bias, as well as the bias of the classifiers, combining subsystems with complex non-linear models can lead to poor generalization performance. Our combiner thus appears as a good compromise. It has proved capable of increasing significantly the recognition rates of the two current best protein secondary structure prediction methods. Multiple extensions of this work can be thought of to make a better use of the approximations of the class posterior probabilities than simply

computing a reliability index. We have highlighted in Guermeur (1997) the fact that their quality was compatible with the implementation of higher level modules such as dynamic programming algorithms or hidden Markov models. This property is very attractive, which should make it possible to implement for structure prediction powerful techniques developed in other fields of sequence processing.

As a conclusion, our ensemble method may provide the user with a prediction that benefits from the advantages of each individual classifier. Moreover, it can exploit any type of prediction. This makes the system evolutive towards the most reliable prediction since it has been observed consistently to yield better results than any individual method. At the biological level, it can be used as the first step for protein threading, topology and *de novo* three-dimensional predictions. The MLR combination of different prediction schemes is available on the Web, thus allowing the biologist to use it without any implementation effort.

Acknowledgements

The authors would like to thank Dr B.Rost for providing them with the PHD predictions for the 126 proteins. Thanks are also due to Dr J.Garnier, Dr J.-F.Gibrat and Dr J.-M.Levin for the availability of the software for the GOR4 and SIMPA96 prediction methods.

References

Bates,J.M. and Granger,C.W.J. (1969) The combination of forecasts. *Opl Res. Q.*, **20**, 451–468.

Biou,V., Gibrat,J.F., Levin,J.M., Robson,B. and Garnier,J. (1988) Secondary structure prediction: combination of three different methods. *Prot. Eng.*, **2**, 185–191.

Bishop,C.M. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Breiman,L. (1996) Stacked regressions. *Machine Learn.*, **24**, 49–64.

Clemen,R.T. and Winkler,R.L. (1985) Limits for the precision and value of information from dependent sources. *Oper. Res.*, **33**, 427–442.

Dickinson,J.P. (1973) Some statistical results in the combination of forecasts. *Opl Res. Q.*, **24**, 253–260.

Dickinson,J.P. (1975) Some comments on the combination of forecasts. *Opl Res. Q.*, **26**, 205–210.

Eisenberg,D., Wilcox,W. and Eshita,S. (1987) Hydrophobic moments as tools for analysis of protein sequences and structures. *Proteins Struct. Funct.*, 425–436.

Eisenhaber,F., Persson,B. and Argos,P. (1995) Protein structure prediction: recognition of primary, secondary and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 1–94.

Fletcher,R. (1991) *Practical Methods of Optimization*, 2nd edn. John Wiley & Sons.

Garnier,J., Gibrat,J.-F. and Robson,B. (1996) GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.*, **266**, 540–553.

Geourjon,C. and Deléage,G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Applic. Biosci.*, **11**, 681–684.

Guermeur,Y. (1997) Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines. Thèse de Doctorat de l'Université Paris 6, Paris.

Guermeur,Y. and Gallinari,P. (1996) Combining statistical models for protein secondary structure prediction. *Proceedings of ICANN'96*, pp. 599–604.

Guermeur,Y. and Paugam-Moisy,H. (1998) Linear ensemble methods for multiclass discrimination. Research Report 1998–52, LIP, ENS Lyon, http://www.ens-lyon.fr/LIP/research_reports.us.html.

Guermeur,Y., Gallinari,P. and Paugam-Moisy,H. (1998) Multivariate linear regression on classifier outputs: a capacity study. *Proceedings of ICANN'98*, pp. 693–698.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Jordan,M.I. and Jacobs,R.A. (1994) Hierarchical mixture of experts and the EM algorithm. *Neural Comput.*, **6**, 181–214.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

LeBlanc,M. and Tibshirani,R. (1993) Combining estimates in regression and classification. Technical Report 9318, Department of Preventive Medicine and Biostatistics and Department of Statistics, University of Toronto, Toronto.

Levin,J.M. (1997) Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng.*, **10**, 771–776.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Peng,F., Jacobs,R.A. and Tanner,M.A. (1994) Bayesian inference in mixture-of-experts and hierarchical mixture-of-experts architectures. Technical Report, Department of Biostatistics, University of Rochester, June 1994.

Riis,S. and Krogh,A. (1996) Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol.*, **3**, 163–183.

Rosen,J.B. (1960) The gradient projection method for nonlinear programming. Part I. Linear constraints. *J. Soc. Indust. Appl. Math.*, **8**, 181–217.

Rost,B. and O'Donoghue,S. (1997) Sisyphus and prediction of protein structure. *Comput. Applic. Biosci.*, **13**, 345–356.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rost,B. and Sander,C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.

Rost,B., Sander,C. and Schneider R. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.

Taylor,W.R. (1986) The classification of amino acid conservation. *J. Theor. Biol.*, **119**, 205–218.

Tumer,K. and Ghosh,J. (1995) Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. TR-95-02-98, The Computer and Vision Research Center, The University of Texas at Austin, Austin, TX.

Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory.* Springer, New York.

Wolpert,D.H. (1992) Stacked generalization. *Neural Networks*, **5**, 241–259.

Zhang,X., Mesirov,J.P. and Waltz,D.L. (1992) Hybrid system for protein secondary structure prediction. *J. Mol. Biol.*, **225**, 1049–1063.

Appendix

We give below the definitions of several of the mathematical concepts used in this article. These definitions are standard in the various fields to which the concepts belong.

For all n in IN^*, IR^n denotes the n-dimensional Euclidian space.

The canonical coding of category C_k is the vector of IR^Q, the components of which are all equal to 0, except the kth one, which is equal to 1.

The class posterior probability $p(C_k|x)$ is the probability that example x belongs to class C_k.

The loss function L is the function which measures, for each example x, the discrepancy between the desired output y and the observed one, $g(x)$. Let p be the probability distribution function on the Cartesian product $X \times Y$ of the input and output spaces and $G = \{g(.,v), v \in V\}$ the set of MLR functions expressed in parametric form. The generalization error, or risk, or cost function associated with function $g(.,v)$ is defined as the mathematical expectation:

$$J(v) = \int_{X \times Y} L(g(x, v), y) \mathrm{d}p(x, y)$$

The estimate of the risk obtained on the training sample D is called the empirical risk.

L is convex if, and only if, for all $(x,y) \in X \times Y$, for all $\theta \in [0,1]$ and for all $(v^{(1)},v^{(2)}) \in V^2$,

$$L(g(x, \theta v^{(1)} + (1 - \theta)v^{(2)}), y) \leq \theta L(g(x, v^{(1)}), y) + (1 - \theta)L(g(x, v^{(2)}), y)$$

Note that the convexity of L implies the convexity of J.

The sample-based estimate of the quadratic cost function is given by:

$$\hat{J}(v) = \frac{1}{N} \sum_{i=1}^{i=N} \| g(x_i, v) - y_i \|^2$$

The sample-based estimate of the cross-entropic error function is:

$$\hat{J}(v) = -\frac{1}{N} \sum_{i=1}^{i=N} \sum_{k=1}^{k=Q} y_{ik} \ln\left(\frac{g_k(x_i, v_k)}{y_{ik}}\right)$$

where y_{ik} is the kth component of vector y_i.