BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i205–i211
DOI: 10.1093/bioinformatics/btg1027

An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins

*Pier Luigi Martelli, Piero Fariselli and Rita Casadio**

Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: All-alpha membrane proteins constitute a functionally relevant subset of the whole proteome. Their content ranges from about 10 to 30% of the cell proteins, based on sequence comparison and specific predictive methods. Due to the paucity of membrane proteins solved with atomic resolution, the training/testing sets of predictive methods for protein topography and topology routinely include very few well-solved structures mixed with a hundred proteins known with low resolution. Moreover, available predictors fail in predicting recently crystallised membrane proteins (Chen *et al.*, 2002). Presently the number of well-solved membrane proteins comprises some 59 chains of low sequence homology. It is therefore possible to train/test predictors only with the set of proteins known with atomic resolution and evaluate more thoroughly the performance of different methods.

Results: We implement a cascade-neural network (NN), two different hidden Markov models (HMM), and their ensemble (ENSEMBLE) as a new method. We train and test in cross validation the three methods and ENSEMBLE on the 59 well resolved membrane proteins. ENSEMBLE scores with a per-protein accuracy of 90% for topography and 71% for topology, outperforming the best single method of 7 and 5 percentage points, respectively. When tested on a low resolution set of 151 proteins, with no homology with the 59 proteins, the per-protein accuracy of ENSEMBLE is 76% for topography and 68% for topology. Our results also indicate that the performance of ENSEMBLE is higher than that of the best predictors presently available on the Web.

Contact: gigi@biocomp.unibo.it; http://www.biocomp.unibo.it

INTRODUCTION

Membrane proteins are involved in almost every cell activity and signal transmission. However their modelling is generally more difficult than that of globular proteins, due to the few examples of membrane proteins known with atomic resolution. For this reason a 2D model of the protein is routinely predicted, highlighting those regions that can interact with the membrane phase. This is done by predicting first the location of transmembrane segments along the protein sequence (topography) and then the location of the N and C terminus with respect to the lipid bilayer (topology). This last step, depending on the predictive method, can be computed using different 'ad hoc' rules derived from experiments and/or statistical analysis (von Heijne, 1999) or using hidden Markov models (Tusnady and Simon, 1998; Krogh *et al.*, 2001).

Two types of membrane proteins have been characterised: the first includes all-alpha proteins that, to a different extent, interact with the lipid bilayer of the cytoplasmic membrane of all cells (White and Wimley, 1999); the second group includes the so called beta-barrel membrane proteins, which interact with the outer membrane with antiparallel beta-strands forming barrels, with an even number of segments (Schulz, 2000). Few methods have been described so far for the prediction of the all-beta membrane proteins (Jacoboni *et al.*, 2001; Martelli *et al.*, 2002; Wimley, 2002, and references therein). On the contrary, several methods have been developed to predict the location of transmembrane segments in the all-helical membrane proteins (for detailed reviews see Möller *et al.*, 2001; Chen *et al.*, 2002).

Routinely, different datasets are used to score the predictor performance. Basically two sets of proteins are considered: the first includes high resolution structures, the second topological models obtained mainly from experimental data (referred to as the low resolution set; Möller *et al.*, 2000). A recent thorough analysis highlights that none of the different advanced methods, based on machine learning and available on the Web (Web predictors), when tested on the high resolution structures of membrane proteins perform consistently best, and that wrong predictions are different for different predictors (Chen *et al.*, 2002).

With the purpose of overcoming the blur introduced by the low resolution training set, we select 59 high-resolution membrane proteins with low sequence identity

*To whom correspondence should be addressed.

to train/test our predictors. We implement a neural network and two HMMs, known to be among the best performing predictors for the task at hand (Chen *et al.*, 2002). We also develop their ensemble (ENSEMBLE) and this is new for the prediction of membrane proteins. Our strategy allows a more thorough comparison between different approaches, based on the high resolution set of membrane proteins, and uses as a blind test the low resolution set. This is different from what was done before, since the predictors previously described were trained on mixed sets of proteins, including also the low resolution models and did not compare predictors on the same training/testing set.

With our approach, we find that all methods perform similarly; however the performance is maximal only when the ensemble of predictors is used, including the neural network and the two HMMs, all trained on evolutionary information. Furthermore, when predicting both the high resolution and low resolution sets of membrane proteins, ENSEMBLE outperforms the best performing Web predictors.

ABSTRACT SYSTEM AND METHODS

Datasets

We use three datasets for different purposes. The first one (S59) is derived from the database of membrane proteins available at http://blanco.biomol.uci.edu (Jayasinghe *et al.*, 2001). S59 comprises 59 high resolution membrane proteins, which are used for training and scoring the predictive methods (available at http://www.biocomp.unibo.it/gigi/ENSEMBLE). The second (S151) is a Möller's database subset (Möller *et al.*, 2000) containing only low resolution proteins, whose sequences do not have similarity with those in S59. The third dataset (S1396) is a non redundant set of 1396 globular proteins, whose structures are known and whose sequences are less than 25% similar (http://www.cbrc.jp/papia/papia.html).

Each predictor is trained using evolutionary information in the form of sequence profiles after multiple sequence alignments. Sequence alignments were obtained using PSI-BLAST (Altschul *et al.*, 1997); three rounds with threshold equal to 0.001) to search against the non-redundant database (available at http://www.ncbi.nlm.nih.gov/BLAST). To train and test the methods a 41-fold cross validation procedure was adopted, in order to ensure that no detectable sequence similarity among training and testing sets were present.

The neural network-based predictor

A feed-forward neural network (NN) is implemented and trained with the back-propagation algorithm to discriminate transmembrane (TM) alpha helices from extra membrane regions, similarly to what described elsewhere (Rost *et al.*, 1995). The network architecture basically consists of a perceptron with one hidden layer containing 15 hidden nodes and an input window spanning 17 residues (for a total of 340 input nodes; each residue is coded with 20 neurons). Two output nodes are considered (TM helices and loops). The architecture of the predictor is extended to include a second cascade network to filter out spurious assignments. This second network consists of 34 inputs (2*17), 5 hidden and 2 output nodes.

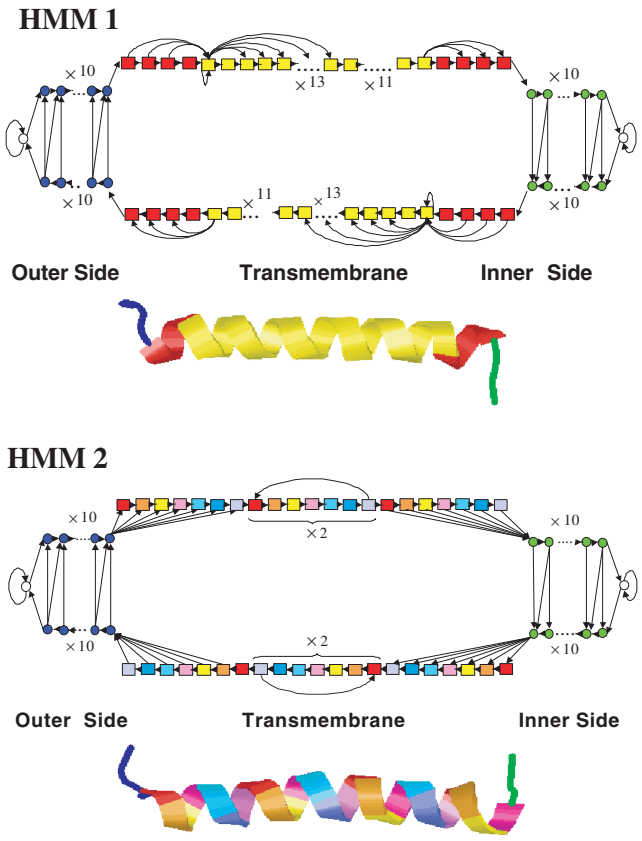

Fig. 1. Graphic models of the two HMMs implemented in this paper. HMM1 models hydrophobic transmembrane helices and HMM2 captures the helix amphipathy. The number of trainable parameters is 173 and 258 for HMM1 and HMM2, respectively. The states filled with the same color share the same emission parameters.

The hidden Markov model-based predictors

We implement two types of hidden Markov models (HMM) in order to capture different features of TM helices present in the data base. The first HMM, (HMM1 in Fig. 1), is conceptually similar to that introduced by Krogh *et al.* (2001). In HMM1 the TM segments are modelled by means of two types of states, one for the

helix core and one for the caps. This model captures the hydrophobic nature of most TM helices. The second HMM is used to model also amphipathic TM helices. HMM2 (Fig. 1) is endowed with a larger number of free parameters, in order to mimic the periodic pattern of hydrophobic and hydrophilic residues that characterise some TM helical segments in S59. This is obtained using a state-tying repetition each 7 residues. In either model, the inner and outer loops are described with different sets of emission parameters, capturing the topological information. The allowed transitions between the states describe the grammar of TM proteins and constrain the minimum length of TM segments to 15 and 16 for HMM1 and HMM2, respectively. The maximal length is unbound in both models, in order to increase their flexibility. Differently from previous implementations (Tusnady and Simon, 1998; Krogh *et al.*, 2001), our HMMs take advantage of evolutionary information derived from sequence profile (Martelli *et al.*, 2002). Training and testing algorithms are described elsewhere (Martelli *et al.*, 2002).

THE ENSEMBLE PREDICTOR

It is possible to take advantage of the disagreement among different predictors by using an ensemble method that averages over the different answers (e.g. Sollich and Krogh, 1996, and references therein). More formally (following Sollich and Krogh, 1996), for a given input x, if $\langle \varepsilon(x) \rangle$ is the error obtained by averaging the errors of the single methods separately, the ensemble error $e(x)$ can be evaluated as

$$e(x) = \langle \varepsilon(x) \rangle - \langle a(x) \rangle \tag{1}$$

where $\langle a(x) \rangle$ is the average disagreement of the single methods with respect to the mean ensemble value. Since both quantities are positive, no improvement is obtained when using a joint method if there is no disagreement ($\langle a(x) \rangle \cong 0$). On the contrary, when there is disagreement among different methods, we can expect an improvement from Equation (1) if an ensemble method is used. This is so, provided that single methods perform similarly. Using this notion, we define a meta-predictor (ENSEMBLE) that averages the predictive answer over the three methods (NN, HMM1 and HMM2). Differently from a consensus method, ENSEMBLE computes the local average of the three methods for each residue in the sequence. This is possible, since both NN and HMMs compute the residue probability of being or not in a TM helix.

More formally, for each sequence position i of a protein p we can define the difference between the TM helical (H) and loop (L) probabilities of the neural network outputs as:

$$\Delta NN(p,i) = NN(H,p,i) - NN(L,p,i) \tag{2}$$

Then we can define the difference between the a posteriori probability for each of the two HMMs of being in a TM helical state (H) and the a posteriori probability of being in a loop state (inner I or outer O) as:

$$\Delta HMM(p,i) = AP(H,p,i) - (AP(I,p,i) + AP(O,p,i)) \tag{3}$$

The ENSEMBLE predictor computes the average propensity value as:

$$E(p,i) = (\Delta NN(p,i) + \Delta HMM1(p,i) + \Delta HMM2(p,i))/3 \tag{4}$$

In this way, for each sequence position i in a protein p, ENSEMBLE computes a value in the range of [-1,1], where positive values indicate that the residue is likely to be in a TM helix.

Selecting the topographical model

The optimal topographical model is computed by using the MaxSubSeq algorithm (Fariselli *et al.*, 2003) based on dynamic programming. MaxSubSeq uses the outputs of a given predictive method and by model optimisation locates the TM segments along the protein sequence. Briefly, a recursive algorithm generates a scoring matrix for each predicted sequence, by evaluating the total sum of the output differences along a segment of fixed length. Minimal and maximal lengths are derived from the database of selected proteins. A model is selected by evaluating the optimal score among those satisfying the observed constraints.

For a given sequence position j and for a given model i (i is the number of TM helical segments) the scoring matrix **S** is computed as:

$$S^i(j) = \max_{m=\lambda_{\min} \to \lambda_{\max}} \{S^i(j-1), S^{i-1}(j-m-1) + s^j_{j-m}\} \tag{5}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum length of a helical TM segment, respectively; s^j_{j-m} is the score of the segment that spans from the sequence positions $j-m$ to j.

All the predictions reported for the methods described in this paper (NN, HMM1, HMM2 and ENSEMBLE) are filtered using MaxSubSeq.

Assigning the topology

NN predictors code only local information in the input window. Therefore topology can only be assigned to a given sequence by means of statistical rules derived from a data base of known topologies, such as the positive inside rule (von Heijne, 1999).

In the case of HMMs, when exploiting the Viterbi's decoding, or the k-best variant (Tusnady and Simon,

1998; Krogh *et al.*, 2001), it is possible to automatically assign the protein topology. As previously demonstrated when predicting the topology of outer membrane proteins (Martelli *et al.*, 2002), the a posteriori decoding (Durbin *et al.*, 1998) performs better. However a drawback of this approach is that sometimes predictions can be incoherent. In our application this is particularly relevant with ENSEMBLE: we can have loop clashes (for instance, HMM1 may assign inside, while HMM2 may assign outside), or a helix can be deleted due to the synergic predictions of the three methods. To overcome this problem we devise a specific set of topological rules. Given a protein sequence, with a list of predicted TM segments, we consider the odd and even loops flanking each TM region. The maximum number of residues included in a loop is 60 for intra-segment loops and 30 if the loop is located at the N or C terminus. Finally we compute the topology of a protein p as

$$Top(p) = \sum_{k=1}^{L} (-1)^k (P(p, I, k) - P(p, O, k)) \quad (6)$$

where L is the loop number, $P(p, I, k)$ and $P(p, O, k)$ are the loop propensities to be inside or outside the membrane, respectively.

The sign of *Top(p)* selects the predicted topology:

- if *Top(p)* > 0 the predicted protein topology is OUT,
- if *Top(p)* < 0 the predicted protein topology is IN,
- if *Top(p)* $= 0$ the predicted protein topology is AMBIGOUS.

Depending on the method, $P(p, I, k)$ and $P(p, O, k)$ are computed from:

***Rule1*:** the von Heijne's rule, where $P(p, I, k)$ is the number of positive charges in the k-th loop and $P(p, O, k)$ is set equal to 0.

***Rule 2*:** the sum of the HMM1 a posteriori propensity computed over the residues in the k-th loop.

***Rule 3*:** the sum of the HMM2 a posteriori propensity computed over the residues in the $k-$th loop.

***Rule 4*:** the sum of the average of the two HMM a posteriori propensities computed over the residues in the k-th loop.

***Rule 5*:** (combining Rule 1 and 4) the sum of the average of the two HMM a posteriori propensities and of the number of positive charges in the k-th loop.

Scoring the prediction

The most relevant accuracy index is Q_{ok}, which computes the topography accuracy of a set comprising Np proteins, and is computed as

$$Q_{ok} = 100 P_{ok}/Np \quad (7)$$

following a recent definition (Chen *et al.*, 2002), where P_{ok} is the number of proteins whose topography is correctly assigned. For each protein the topography is a binary measure, since we consider 1 (correct) or 0 (wrong) depending on the fact that a prediction meets both of the following conditions

(i) the number of predicted segments equals the observed one;

(ii) the overlap between the predicted and expected segments equals at least 9 residues.

This is in agreement with a previous stringent definition (Chen *et al.*, 2002).

The second most relevant index is Q_T, which accounts for the topology predictions and is obtained scoring a given set of Np proteins

$$Q_T = 100 P_T/Np \quad (8)$$

where P_T is the number of proteins whose topology is correctly assigned. Since Q_{ok} and Q_T are the two most critical accuracy measures, we also compute the error associated with them, assuming that the underlying distributions are binomial. Both Q_{ok} and Q_T are evaluated after filtering with MaxSubSeq.

The Sov index computes the overlapping between the predicted and the expected TM segment (Zemla *et al.*, 1999). Finally the per-residue performance is also evaluated using Q_2(accuracy), C (correlation coefficient), Q (coverage) and P (precision) as previously described (Martelli *et al.*, 2002).

RESULTS AND DISCUSSION

Topography prediction

We want first to compare machine learning approaches based on evolutionary information on the topography prediction of membrane proteins. With the S59 high resolution set, we score the NN, the HMMs and the ENSEMBLE methods. The per-protein and per-residue performances, evaluated using a cross validation procedure, are listed in Table 1. Both NN and HMMs, when implemented with evolutionary information have a comparable performance, with NN scoring slightly better than HMMs. However the ensemble of methods (ENSEMBLE) shows a large improvement of the Q_{ok} accuracy, from 7 to 9 percentage points.

Table 1. Performance of different methods in cross-validation on the S59 dataset

	NN	**HMM1**	**HMM2**	**ENS**
Q_{ok}%	83	81	81	90
(correct/total)	(49/59)	(48/59)	(48/59)	(53/59)
Q2%	86	84	82	85
Corr	0.714	0.692	0.658	0.708
Q_{TM}%	84	89	88	87
Q_{Loop}%	87	81	78	84
P_{TM}%	85	80	77	82
P_{Loop}%	87	89	88	88
SOV	0.908	0.896	0.872	0.926

Indexes when indicated are computed as percent value. The correctly predicted proteins over the total are indicated among brackets. According to the binomial distribution the associated maximal standard deviation of Q_{ok} is 5%. For the definition of the different indexes see System and Methods.

Table 2. Blind test of the S151 dataset

	NN	**HMM1**	**HMM2**	**ENS**
Q_{ok}%	68	65	62.	75
(correct/total)	(103/151)	(98/151)	(94/151)	(114/151)
Q2%	84	85	83	88
Corr	0.626	0.695	0.663	0.740
Q_{TM}%	84	96	97	96
Q_{Loop}%	84	82	78	86
P_{TM}%	64	64	60	69
P_{Loop}%	94	98	99	98
SOV	0.870	0.864	0.839	0.894

According to the binomial distribution the associated maximal standard deviation of Q_{ok} is 4%. For the meaning of the indices see System and Methods.

In Table 2 our methods are tested on the S151 low resolution set, which comprises 151 protein chains with sequence identity <25% to those of S59 and is used as a blind test. In this case, the TM annotation is derived from low resolution experiments (based on molecular biology or biochemical methods). Basically a decrease of the general performance of the predictors is noticed. Again ENSEMBLE outperforms the single methods. As previously discussed (Chen *et al.*, 2002), the observed decrease may reflect that the low resolution set contains new motifs but also that the low resolution assignment over- or under-annotates TM helices.

Our predictors, including ENSEMBLE, and others in the literature, wrongly predict signal peptides as TM helices. This is the case for a subset of 34 proteins in S151, containing the signal peptide. We however can take advantage of well performing predictors of signal peptides (Nielsen *et al.*, 1999). When a signal peptide is predicted, this can be excluded and the sequence is then predicted.

Table 3. The prediction of S59 topology using different rules

Method/ Rule	Q_T	*Number of ambiguous*
NN		
Rule 1	56% (33/59)	7
HMM1		
Rule 1	56% (33/59)	9
Rule 2	68% (40/59)	0
HMM2		
Rule 1	54% (32/59)	10
Rule 3	68% (40/59)	0
ENSEMBLE		
Rule 1	61% (36/59)	11
Rule 2	76% (45/59)	0
Rule 3	75% (44/59)	0
Rule 4	76% (45/59)	0
Rule 5	76% (45/59)	0

According to the binomial distribution the associated maximal standard deviation is 6%. Rules are defined in System and Methods

The data shown in Table 2 are done after deletion of the signal peptides; 30 out of the 34 proteins are then correctly predicted.

Topology prediction

With the predictors at hand we can also compare how the different methods assign the protein topology on the high resolution set. The results are reported in Table 3. NN and the positive inside rule (Rule 1, as implemented by our method) are clearly overcome by both HMM assignments. Particularly, no ambiguity is detected with HMMs, whereas the positive inside rule implementation predicts a significant percentage of ambiguous cases. ENSEMBLE, that is superior when predicting the protein topography, reaches a noteworthy 76% accuracy also when predicting protein topology. This is so, provided that the HMM-derived information is considered (Rule 2, 3 and 4). No further improvement is detected when the positive inside rule is used in combination with HMM information (Rule 5).

Comparison with Web predictors

The performance of ENSEMBLE is compared to that of other predictors recently scored as the best ones available (Chen *et al.*, 2002). The results are shown in Table 4. It should however be noticed that only ENSEMBLE is scored by adopting a cross validation procedure since some of the predicted proteins are present in the training sets of the other methods. When topography and topology are predicted, it is evident that ENSEMBLE scores higher than the other Web predictors both on S59 and S151 (for all predictions signal peptides were excluded)

Table 4. Performance of ENSEMBLE and other Web methods on the S59 and S151 datasets

METHOD	S59		S151	
	Q_{ok}	Q_T	Q_{ok}	Q_T
ENSEMBLE*	90%	76%	75%	68%
(Rule 4)	(53/59)	(45/59)	(114/151)	(103/151)
TMHMM 2.0[+]	71%	54%	72%	63%
	(42/59)	(32/59)	(109/151)	(95/151)
MEMSAT[°]	71%	55%	73%	58%
	(35/49)	(27/49)	(107/146)	(85/146)
PHD[§]	73%	49%	68%	62%
	(43/59)	(29/59)	(103/151)	(94/151)
HMMTOP[#]	76%	66%	72%	64%
	(45/59)	(39/59)	(108/151)	(97/151)

*ENSEMBLE is used adopting a cross validation procedure. Web predictors contain some of the tested proteins in the training set. [+]Krogh *et al.*, 2001; [°]MEMSAT does not predict chains without PSI-BLAST alignment (Jones *et al.*, 1994); [§](Rost *et al.*, 1996); [#](Tusnady and Simon, 1998). For the definition of the different indexes see System and Methods.

Predicting globular proteins

When assigning membrane proteins in a large-scale genome analysis, it is important to know the rate of missing membrane proteins (false negatives) and the rate of false positive globular proteins. To evaluate this, we compare the average propensity values predicted with ENSEMBLE both for S59 and S1396, a set containing 1396 non redundant globular proteins. The classification error rate of the two sets is plotted as a function of the maximal peak value found among all the putative TM helices in each sequence (Fig. 2). From this plot it is clear that by rejecting propensity values $\leqslant$0.92, about 3% of membrane proteins are missed (rate of false negatives) and about 3% of globular proteins are wrongly classified (rate of false positives). These error rates are in the range of those reported for the best methods available (Chen *et al.*, 2002).

The length distribution of TM segments

A crucial question in predicting TM helices is how the predicted length compares to that expected in the high resolution set of membrane proteins. Routinely, predictive methods assign the majority of segment length to one extreme of their minimal or maximal allowed value (Chen and Rost, 2002). Minimal and maximal TM segment lengths are implemented as direct constraints in the dynamic programming filter (Jones *et al.*, 1994; Rost *et al.*, 1996), or in the HMM grammars (Tusnady and Simon, 1998; Krogh *et al.*, 2001).

We overcome this problem using MaxSubSeq and filtering the ENSEMBLE outputs. We allow minimal and maximal lengths of 15 and 40 residues, respectively. These limits are derived from the dataset (S59). Interestingly, and

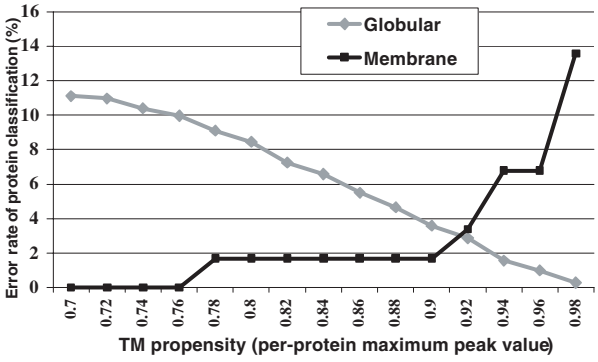

Fig. 2. Error rate of membrane and globular protein classification.

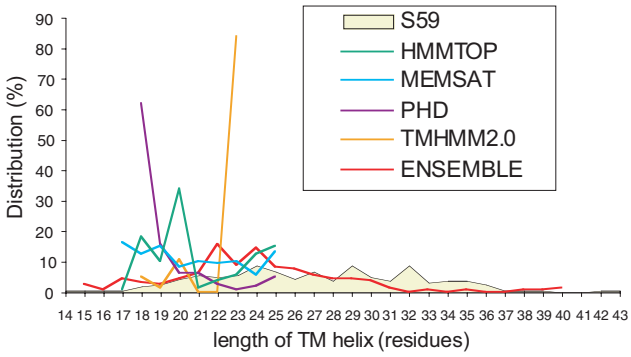

Fig. 3. The TM length distribution of S59 as compared to that predicted with different methods.

differently from other predictors, the length distribution of the TM helices predicted with ENSEMBLE is comparable to that derived from S59 (Fig. 3). This indicates that our constraints are more suited than others to partially overlap the expected length distribution.

CONCLUSIONS

In this paper we implement three machine learning systems, and their ENSEMBLE, as a new method. We show that this new approach highly performs on a cross validated data set of high resolution proteins (S59), and scores higher than the best performing methods both on the set of high resolution and low resolution proteins (S151). This is noteworthy, if we consider that our results are obtained using a cross validation procedure and are compared to performances of other Web predictors containing some of the tested proteins in the training set (Chen *et al.*, 2002). We have also introduced different

types of rules for protein topology prediction, verifying that the best performing one must contain the information extracted by the HMM systems. Moreover the ensemble predictor is quite efficient in discriminating membrane from globular proteins. Overall these results suggest that ENSEMBLE, when coupled with a signal peptide predictor, can be used for large-scale annotation of all-alpha membrane proteins.

AKNOWLEDGEMENTS

This work was partially supported by a grant of the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) for the project 'Hydrolases from Thermophiles: Structure, Function and Homologous and Heterologous Expression', a grant for a target project in Biotechnology, a project on Molecular Genetics, both of the Italian Centro Nazionale delle Ricerche (CNR), a PRIN 2002 for Development and implementation of algorithms for predicting protein structure and a PNR 2001-2003 (FIRB art.8) project on Postgenomics, delivered to RC. PLM is the recipient of a fellowship from the Italian Center for National Researches (CNR) devoted to a target project of Molecular Genetics (Law No 449-1997).

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.*, **25**, 3389–3402.

Chen,C.P., Kernytsky,A. and Rost,B. (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774–2791.

Chen,C.P. and Rost,B. (2002) Long membrane helices and short loops predicted less accurately. *Protein Sci.*, **11**, 2766–2773.

Durbin,R., Eddy,S., Krogh,A. and Mitchinson,G. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

Fariselli,P., Finelli,M., Marchignoli,D., Martelli,P.L., Rossi,I. and Casadio,R. (2003) MaxSubSeq: an algorithm for segment-length optimization. The case study of the transmembrane spanning segments. *Bioinformatics*, **19**, 500–505.

Jacoboni,I., Martelli,P.L., Fariselli,P., De Pinto,V. and Casadio,R. (2001) Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci.*, **10**, 779–787.

Jayasinghe,S., Hristova,K. and White,H.H. (2001) MPtopo: a database of membrane protein topology. *Protein Sci.*, **10**, 455–458.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.

Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Möller,S., Kriventseva,E.V. and Apweiler,R. (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.

Möller,S., Croning,M.D. and Apweiler,R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Martelli,P.L., Fariselli,P., Krogh,A. and Casadio,R. (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18**, S46–S53.

Nielsen,H., Brunak,S. and von Heijne,G. (1999) Machine learning approaches to the prediction of signal peptides and other protein sorting signals. *Protein Engng*, **12**, 3–9.

Rost,B., Casadio,R., Fariselli,P. and Sander,C. (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci.*, **4**, 521–533.

Rost,B., Fariselli,P. and Casadio,R. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704–1718.

Schulz,G.E. (2000) β-barrel membrane proteins. *Curr. Opin. Struct. Biol.*, **10**, 443–447.

Sollich,P. and Krogh,A. (1996) Learning with ensembles: how over-fitting can be useful. *Advanced in Neural Information Processing Systems,* 8. pp. 190–196.

Tusnady,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**, 489–506.

von Heijne,G. (1999) Recent advances in the understanding of membrane protein assembly and structure. *Q. Rev. Biophys.*, **32**, 285–307.

White,S.H. and Wimley,W.C. (1999) Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.*, **28**, 319–365.

Wimley,W.C. (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci.*, **11**, 301–312.

Zemla,A., Venclovas,C., Fidelis,K. and Rost,B. (1999) A modified definition of Sov, a segment-based measure of protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.