How Many 3D Structures Do We Need to Train a Predictor?

Pantelis G. Bagos[1,2]*, Georgios N. Tsaousis[1], and Stavros J. Hamodrakas[1]

[1] *Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens 15701, Greece;* [2] *Department of Computer Science and Biomedical Informatics, University of Central Greece, Lamia 35100, Greece.*

*Corresponding author. E-mail: pbagos@biol.uoa.gr; pbagos@ucg.gr

DOI: 10.1016/S1672-0229(08)60041-8

It has been shown that the progress in the determination of membrane protein structure grows exponentially, with approximately the same growth rate as that of the water-soluble proteins. In order to investigate the effect of this, on the performance of prediction algorithms for both α-helical and β-barrel membrane proteins, we conducted a prospective study based on historical records. We trained separate hidden Markov models with different sized training sets and evaluated their performance on topology prediction for the two classes of transmembrane proteins. We show that the existing top-scoring algorithms for predicting the transmembrane segments of α-helical membrane proteins perform slightly better than that of β-barrel outer membrane proteins in all measures of accuracy. With the same rationale, a meta-analysis of the performance of the secondary structure prediction algorithms indicates that existing algorithmic techniques cannot be further improved by just adding more non-homologous sequences to the training sets. The upper limit for secondary structure prediction is estimated to be no more than 70% and 80% of correctly predicted residues for single sequence based methods and multiple sequence based ones, respectively. Therefore, we should concentrate our efforts on utilizing new techniques for the development of even better scoring predictors.

Key words: membrane protein, secondary structure prediction, alpha-helical, beta-barrel, 3D structure

Introduction

The three-dimensional (3D) structure of a protein is determined by its amino acid sequence in a given environment and, consequently, determines its exact biological function (*1*). However, experimental methods for determining the structure of a given protein such as X-ray crystallography and nuclear magnetic resonance are expensive, time-consuming and in many cases (*i.e.*, concerning membrane proteins) not easy for a number of reasons. Thus, from the early days of computational biology, several attempts were made in order to develop algorithms that can predict the secondary structure of a protein using only information encoded in its primary sequence. Later on, similar algorithms were developed for predicting more specialized secondary structure features such as transmembrane helices and β-strands of transmembrane (TM) proteins. In a typical case, a limited number of non-homologous sequences with known 3D structure are used for training the algorithm and the method is supposedly able to predict the secondary structure of newly discovered and unrelated proteins. Thus, we expect that as newly solved 3D structures are accumulated, the prediction methods would become better.

It has been shown that the progress of protein structure determination is approximately the same for membrane proteins and water-soluble ones, taking into consideration the year of the first published structure (*2*). However, membrane proteins have a delay in the appearance of the first published structure of about 25 years compared to the water-soluble proteins. Moreover, for training a predictor, usually a non-redundant dataset is used. From the early years of the structure prediction algorithms, it was anticipated that an increase in the number of non-homologous sequences with known structure would enhance the prediction accuracy. However, later it became evident that after a particular point, the prediction accuracy could not be further improved just by increasing the size of the training set. In this work, we try to empirically answer the question regarding the relationship between the size of the training set and the prediction accuracy.

We address separately the general problem of predicting the secondary structure of proteins, and that of predicting the TM segments of membrane proteins (α-helical and β-barrels). The methods used for secondary structure prediction of water-soluble proteins appeared much earlier in the progress of biological research and continue to grow, taking advantage of the increasing number of available unique structures determined year by year. However, even using the most advanced computational techniques devoted to this task (neural networks, support vector machines, *etc.*) and including as input evolutionary information in the form of multiple alignments, it is currently acceptable that their prediction performance cannot exceed an upper limit, no matter what the increase of the training set would be. In order to quantify this common belief, we performed a meta-analysis of published results using data from the existing literature.

Concerning membrane proteins, we have conducted a historical prospective study in order to illustrate the potential impact of newly determined 3D structures in the topology prediction by state-of-the-art machine learning computational methods. Along these lines, we have used the hidden Markov model (HMM)-based computational methods recently proposed by our group, namely PRED-TMBB (*3*, *4*) and HMM-TM (*5*), as platforms to get an estimate of the improvement of computational predictive methods, as more (unique) structures become available for both α-helical and β-barrel TM proteins.

Results and Discussion

Table 1 Studies included in the meta-analysis for the accuracy of the secondary structure prediction algorithms

Year	Reference	Training set (No. of proteins)	Q_3	Evolutionary information
1978	*32*	29	53	NO
1978	*33*	25	57	NO
1986	*34*	61	62.2	NO
1987	*35*	59	61.3	NO
1987	*36*	68	63	NO
1987	*37*	25	66	YES
1988	*38*	62	58.7	NO
1988	*39*	106	64.3	NO
1989	*40*	48	63	NO
1990	*41*	62	64	NO
1992	*42*	107	66.4	NO
1993	*43*	91	64.5	NO
1993	*44*	126	72	YES
1993	*45*	110	68	NO
1996	*46*	318	72.9	YES
1996	*46*	318	67	NO
1996	*47*	267	64.4	NO
1996	*48*	126	71.3	YES
1996	*48*	126	66.3	NO
1997	*49*	556	75	YES
1997	*50*	402	67.5	NO
1997	*51*	512	68	NO
1997	*51*	512	72.4	YES
1997	*52*	90	73.5	YES
1997	*53*	304	72	YES
1997	*53*	473	67	NO
1999	*54*	1,180	76.6	YES
1999	*55*	681	76.6	YES
1999	*56*	396	72.9	YES
1999	*57*	187	76.5	YES
2000	*58*	480	76.4	YES
2000	*59*	496	76.7	YES
2000	*60*	1,032	80.6	YES
2000	*61*	452	68.8	NO
2001	*62*	513	73.5	YES
2001	*63*	396	73.7	YES
2001	*63*	396	68.8	NO
2001	*16*	126	75.1	YES
2002	*64*	513	73.5	YES
2002	*64*	513	67.5	NO
2002	*65*	1,180	78.13	YES
2003	*66*	480	78.5	YES
2003	*67*	126	72.8	YES
2003	*68*	1,460	77.07	YES
2004	*69*	513	75.2	YES

The literature search for secondary structure prediction algorithms identified 59 studies that fulfilled our criteria (**Table 1**). The methods are classified into two classes according to the input they use, those using single sequence information (23 methods) and those using evolutionary information in the form of multiple alignments (36 methods). The methods are highly heterogeneous according to the algorithmic technique they utilize; we encountered feed-forward neural networks with various fixed topologies (FFNNs), cascaded correlation neural networks (CC-

Table 1 ***Continued***

Year	Reference	Training set (No. of proteins)	Q_3	Evolutionary information
2004	*70*	1,612	70.2	NO
2004	*71*	513	77	YES
2004	*72*	513	78.44	YES
2004	*73*	513	76.5	YES
2005	*74*	3,553	77.1	YES
2005	*75*	860	78.4	YES
2005	*76*	396	76.3	YES
2005	*77*	2,171	79	YES
2005	*78*	513	79.4	YES
2005	*79*	513	69	NO
2005	*79*	513	76.4	YES
2005	*80*	374	76	YES
2005	*7*	3,925	81.8	YES
2005	*81*	297	70	YES

NNs), recurrent neural networks (RNNs), partially recurrent neural networks (PRNNs), bidirectional recurrent neural networks (BRNNs), hybrid methods such as hidden neural networks (HNNs), linear regression classifiers, support vector machines (SVMs), nearest neighbor methods, Bayesian networks (BNs) and various propensity based statistical methods. The training set used in each method also varied dramatically among methods, from 27 sequences in the earlier works (*6*) to 3,925 sequences in the most recent work (*7*). The datasets used for the historical prospective study (for both α-helical and β-barrel TM proteins) are listed in the supplementary material at http://bioinformatics.biol.uoa.gr/historical/.

In **Table 2**, we list the detailed results of fitting the linear and non-linear curves on the measures of performance (Q, C and SOV) for α-helical and β-barrel TM proteins, as well as on the Q_3 statistic for secondary structure prediction algorithms (see Materials and Methods). From the root mean squared error (RMSE) statistics, it is clear that in the case of β-barrel TM proteins and secondary structure (both with and without the use of multiple alignments), the non-linear model fits better to the data. For α-helical TM proteins, the RMSEs are nearly equivalent in all three cases. However, the growth rate represented by the β_1 coefficients is very small, a fact indicated also in their large standard errors (resulting in marginally statistical significant slopes for Q and C, and in an insignificant one for SOV). Concerning secondary structure predictions, the estimates correspond to an upper limit for the performance of the single sequence methods at around 70%, whereas at the same time for multiple sequence methods this limit is somewhere around 80% (**Figure 1**). The differences between single sequence based methods and those using multiple alignments are reflected in the estimated β_1 coefficient of the model for each class (0.022 vs 0.002). This parameter expresses the shape of the fitted line. For instance, larger values correspond to a rapid initial growth and faster saturation, as opposed to smaller values that correspond to a more smooth increase. Even though one has to have in mind that we are comparing entirely different methods, it seems that there are differences between the two distributions. Thus, the linear phase for the growth of performance for single sequence based methods is estimated to be for datasets <200, whereas the same for multiple alignment based methods (mostly using NNs and SVMs) is for datasets <1000. It seems that methods depending on multiple alignments are more dependent on the size of the training set, perhaps as a consequence of the fact that they utilize many more trainable parameters.

Comparing α-helical TM proteins with β-barrels (**Figures 2 and 3**), we can also observe that the former can achieve a borderline better performance in any measure studied. This is something expected since it is well known that predicting the TM regions in α-helical membrane proteins is a much easier task compared to the β-barrels. Furthermore, β-barrels need less 3D structures in order to train a predictor. This has to be interpreted taking into account the smaller number of parameters used in the models as well as the existence of fewer structural folds of TM β-barrels. Comparing the prediction of TM proteins (α-helical ones and β-barrels) to secondary structure prediction, we have to also note the superior performance of algorithms for TM protein topology prediction. Once again, this is something that we expected since TM protein topology prediction can be seen as a very specialized case of secondary structure prediction. The limitations imposed by the lipid bilayer restrict the possible conformations of a polypeptide chain, making the prediction a relatively easier task. On the contrary, trying to predict the secondary structure is harder since the prediction algorithm has to be able to predict all the available conformations deriving from the large number of structural folds. Recent studies (*8*, *9*) suggest that the possible "folds" of membrane proteins are limited (in a same way that the number of soluble proteins' folds is limited). Thus, we expect that the findings reported in this work

Table 2 Results obtained from the linear and non-linear regression for secondary structure, α-helical membrane and β-barrel membrane proteins

		β_0 (SE)	β_1 (SE)	β_2 (SE)	RMSE
Non-linear			β-barrel TM proteins		
	Q_β	0.869 (0.010)	0.153 (0.034)	-7.876 (2.579)	0.0070
	C_β	0.734 (0.0217)	0.153 (0.036)	-2.183 (1.516)	0.0149
	SOV	0.874 (0.0121)	0.216 (0.041)	-1.398 (1.184)	0.0132
			α-helical TM proteins		
	Q_α	0.884 (0.018)	0.019 (0.020)	-140.0415 (145.267)	0.0093
	C_α	0.776 (0.098)	0.012 (0.018)	-139.895 (183.753)	0.0213
	SOV	0.904 (0.013)	1.984 (–)	14.583 (0.366)	0.0376
			Secondary structure		
	Q_3 (single)	0.679 (0.006)	0.022 (0.004)	-50.405 (17.613)	0.0182
	Q_3 (multiple)	0.790 (0.011)	0.002 (7.4×10^{-4})	-976.918 (351.151)	0.0219
Linear			β-barrel TM proteins		
	Q_β	0.735 (0.014)	0.007 (9.7×10^{-4})	–	0.0157
	C_β	0.462 (0.028)	0.013 (0.002)	–	0.0322
	SOV	0.646 (0.034)	0.012 (0.002)	–	0.0389
			α-helical TM proteins		
	Q_α	0.843 (0.006)	3.3×10^{-4} (8.6×10^{-5})	–	0.0093
	C_α	0.655 (0.013)	7.9×10^{-4} (1.9×10^{-4})	–	0.0206
	SOV	0.879 (0.025)	3.3×10^{-4} (3.7×10^{-4})	–	0.0398
			Secondary structure		
	Q_3 (single)	0.627 (0.009)	7.1×10^{-5} (2.2×10^{-5})	–	0.0349
	Q_3 (multiple)	0.740 (0.007)	2.0×10^{-5} (7.0×10^{-6})	–	0.0269

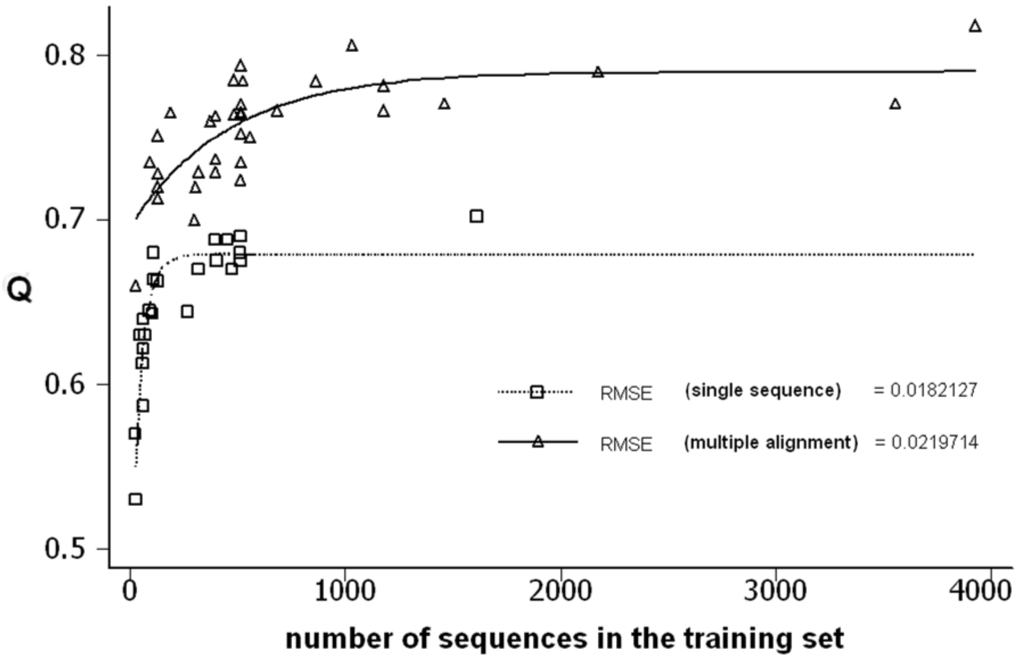

Figure 1 The prediction accuracy (Q_3) of secondary structure prediction algorithms in relation to the size of the training set. Single sequence methods are depicted with squares and multiple alignment-based ones are depicted with triangles. The non-linear regression curves for single sequence and multiple alignment ones are depicted with solid and dotted lines respectively.

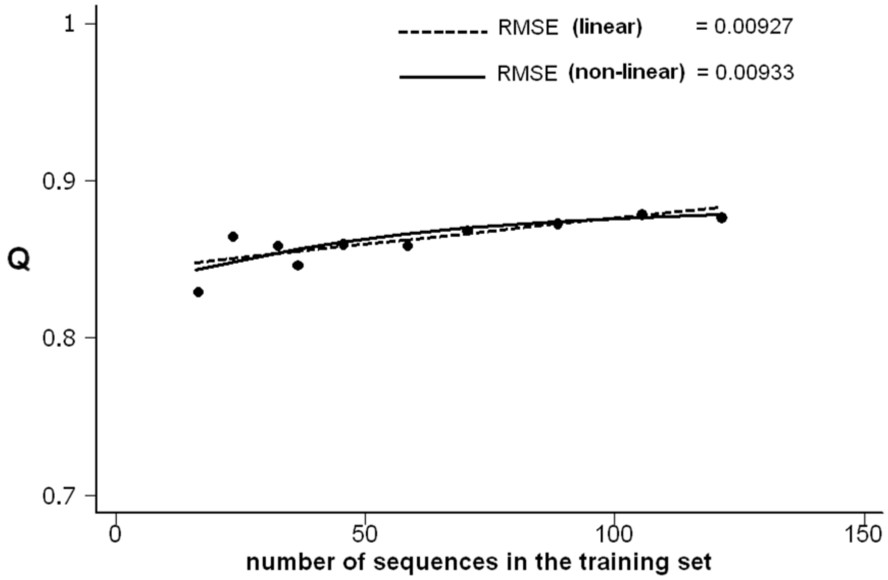

Figure 2 The prediction accuracy (Q_α) of prediction algorithms for α-helical membrane proteins in relation to the size of the training set. The non-linear and linear regression curves are depicted with solid and dotted lines respectively.

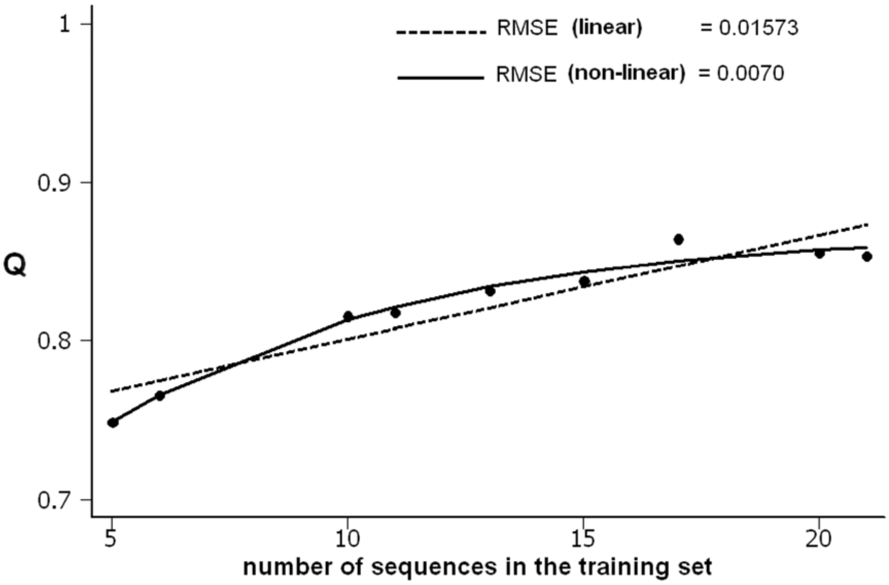

Figure 3 The prediction accuracy (Q_β) of prediction algorithms for β-barrel membrane proteins in relation to the size of the training set. The non-linear and linear regression curves are depicted with solid and dotted lines respectively.

could be extrapolated to the future provided that a fold, which is completely different compared to what we already have seen, will not appear. Given that the basic principles governing membrane protein folding (such as hydrophobicity) have already been taken into account when designing these algorithms, we have no reason to expect a dramatical change in the future.

In our historical prospective study, we used solely methods with single sequence information. In the case of α-helical TM proteins, HMM-TM has been shown to outperform the high-scoring methods currently available such as TMHMM and HMMTOP that use single sequences, and compares favorably against the newly developed methods that use multiple alignments. In case we used multiple alignments in our method, perhaps a higher plateau could be reached and maybe we had a slightly different shape of the growth curve; however, the general conclusions would remain unaffected. In the case of β-barrels, PRED-TMBB has been shown to be one of the most successful prediction algorithms outperforming even methods that use evolutionary information. Furthermore, it has been shown that HMM methods outperform other methods based on NNs and SVMs in both prediction of α-helical (*10*) and β-barrel TM proteins (*11*). Thus, the results of this study are not likely to be inflated by the type of the prediction method used.

The type of the algorithmic technique used for secondary structure prediction has a direct impact on the performance, and the accumulated experience over the years has provided researchers with useful heuristic rules that increase the performance. Furthermore, for algorithms using evolutionary information derived from multiple alignments, the choice of a particular algorithm such as BLAST or PSI-BLAST (*12*), HMMER (*13*), or CLUSTAL (*14*) in order to perform the database search and the alignment may influence the results. In addition, the size of the database on which the search is performed has been shown to influence the results greatly, thus favoring the more recently published methods (*15*, *16*). However, the results reported here clearly indicate that, using existing algorithmic techniques, the performance of secondary structure prediction algorithms cannot be further improved by increasing the size of the training set.

In the case of membrane proteins, the study that we conducted eliminates all the possible sources of variation (different methods for training, different selection criteria for the dataset, *etc.*), thus it is expected to produce unbiased estimates for the dependence on the size of the training set. The total number of freely estimated parameters in the model used for β-barrel membrane proteins is 175, whereas the respective number for the model used for α-helical membrane proteins is 304. These numbers are adequate for training a prediction method using some dozens of proteins (*i.e.*, thousands of amino acids as the observations) and in any case are significantly smaller compared to the number of freely estimated parameters (weights) needed by an NN method. Perhaps if we used an NN method, different estimates would have been produced. However, HMMs have been proved to be not only the most parsimonious among the machine learning algorithms, but also the most efficient for predicting the topology of TM proteins. Furthermore, the particular HMM methods used here have been found to be among the top-scoring ones in the literature (*5*, *11*).

The major finding of this work is the identification of an upper limit for the performance of the prediction algorithms. We have shown that using the existing algorithmic techniques, the prediction performance cannot be further improved by simply adding sequences to the training set. Thus, we need to develop new algorithmic techniques entirely different from the ones used up to now. Such methods definitively need to be able to exploit long-range interactions (correlations) along the sequence (*17*). All currently available techniques are based (one way or the other) on the use of the statistical properties of neighboring amino acid residues along the sequence. Thus, they all use local information and ignore long-range dependencies, which are highly important for the stability of the secondary structure elements and in some cases such as the β-sheet, are responsible for their formation. A few methods have been used already for incorporating long-range interactions along a protein sequence in the secondary structure prediction problem using neural networks (*18*) or variations of the stochastic context free grammars (*19*, *20*), whereas other methods mainly based on neural networks are devoted solely to predict the long-range interactions (*21–24*). Such techniques are computationally more demanding, but given that computational power continues to grow, their use should be exploited further in the context of structure prediction algorithms in the near future.

Materials and Methods

Performing a literature search in PubMed (www.pubmed.gov), we identified studies describing an algorithm for secondary structure prediction that reported: (1) explicitly the use of a non-redundant training set, and (2) the prediction performance using the percentage of correctly predicted residues (Q_3) in a three-state mode (H-helix, E-extended, C-coil) in a test set having no significant similarity with the set used for training. For the latter, we accepted either the test on an independent set or the results of a cross validation or a jackknife test. We further classified the algorithms into two classes, those that depend on single sequence information, and those that use evolutionary information derived from multiple alignments. If a certain prediction method reports both the results using single sequences and multiple alignments, these results are counted separately. Finally, if a method reports the performance on two or more large independent sets, we kept only the one with the highest accuracy.

For the analysis regarding TM proteins, in order to eliminate the inherent variability of the different methods applied on different datasets, we decided to conduct a prospective study based on historical records (a so-called "historical prospective study"). We used PDB_TM (*25*, *26*) in order to collect all the available high-resolution structures of α-helical and β-barrel TM proteins deposited in the Protein Data Bank (*27*). Consequently, we ranked these structures according to the year of publication. Thus, we were able to create datasets corresponding to the structures available for each year in the range 1995–2005. Since there was a slight delay between the elucidation of the first structure of an α-helical membrane protein (1986) and that of the first structure of a porin (1992), we decided to subtract the offset of 6 years, and thus obtain datasets for each year following the first published structure. For each dataset, we performed a redundancy check, using algorithm 2 from Hobohm *et al* (*28*). Non-redundant datasets were created by removing all chains for which a putative homologous entry was already in the set, with the threshold defined as <30% pairwise sequence similarity (in a length of more than 80 residues) in a BLAST alignment (*12*). For sequences shorter than 80 residues, which are frequent among single-spanning membrane proteins, we used the similarity of less than 50% as threshold in a length of more than 30 residues.

For each such set, we trained separately a different HMM in order to predict the TM segments. The model used for the β-barrels was identical to the one introduced with the PRED-TMBB method (*3*), whereas the model for α-helical membrane proteins was the same as that used in HMM-TM (*5*). Concerning β-barrels, we evaluated the performance on the jackknife test (*i.e.*, removing a protein from the training set, training the model with the remaining proteins and performing the test on the protein removed). In the case of α-helical TM proteins, where the training sets were larger, we used a seven-fold cross-validation procedure. Since the sequences do not show any significant similarity (no more than 30% identities in a BLAST comparison), the results of the study were approximately to what would have been observed if such an algorithm was applied at that particular time. We used the Matthews correlation coefficient (C_α and C_β for α-helical and β-barrel TM proteins respectively) and the percentage of correctly predicted residues (Q_α and Q_β for α-helical and β-barrel TM proteins respectively) (*29*), as well as the segment overlap measure (SOV) (*30*) against the structures used for training each HMM. In both cases (α-helical and β-barrel TM proteins), the observed structures, against which the comparisons were performed, were obtained by visual inspection of the 3D structures. Especially for α-helical TM proteins, as explained in detail in the respective paper (*5*), a procedure for the refinement of the boundaries of the TM segments was performed prior to train the final model.

The relationship between the sizes of the training set with the performance of the prediction algorithms was assessed using linear and non-linear models. We fitted a simple linear regression line on each of the parameters (C, Q and SOV denoted here as y) against the number of proteins in the training set (x):

$$y = \beta_0 + \beta_1 x$$

Here of interest is the coefficient β_1, which denotes the amount of increase in the predictive performance that we can achieve by adding one more protein to the training set. In order to check for non-linearity with respect to the training set, we used the non-linear model of von Bertalanffy (*31*):

$$y = \beta_0 \left(1 - e^{-\beta_1(x-\beta_2)}\right)$$

This model requires the estimation of three parameters β_0, β_1 and β_2. β_0 corresponds to the maximal prediction performance, β_1 corresponds to the growth rate and β_2 is an offset corresponding to the hypothetical size of a training set required in order to have a y equal to zero. The parameters were estimated iteratively by non-linear least squares. In order to decide which model fits better to the data (linear vs non-linear), we used the RMSE statistic given by the formula:

$$RMSE = \sqrt{\frac{\sum\limits_i (y_i - \widehat{y_i})^2}{n}}$$

where $\widehat{y_i}$ is the predicted model value for the i^{th} observation. Smaller values of RMSE denote a better fit.

Acknowledgements

PGB was supported by a scholarship from the State Scholarships Foundation of Greece (SSF) for postdoctoral research in the Department of Cell Biology and Biophysics of the University of Athens (Machine Learning Algorithms for Bioinformatics).

Authors' contributions

PGB conceived the study, designed the algorithms, performed the statistical analysis and wrote the manuscript. GNT collected the datasets, performed the training procedure and participated in writing the manuscript. SJH supervised the project and co-wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Anfinsen, C.B. 1972. The formation and stabilization of protein structure. *Biochem. J.* 128: 737-749.
2. White, S.H. 2004. The progress of membrane protein structure determination. *Protein Sci.* 13: 1948-1949.
3. Bagos, P.G., *et al.* 2004. A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics* 5: 29.
4. Bagos, P.G., *et al.* 2004. PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res.* 32: W400-404.
5. Bagos, P.G., *et al.* 2006. Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinformatics* 7: 189.
6. Chou, P.Y. and Fasman, G.D. 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* 13: 211-222.
7. Lin, H.N., *et al.* 2005. HYPROSP II—a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* 21: 3227-3233.
8. Martin-Galiano, A.J. and Frishman, D. 2006. Defining the fold space of membrane proteins: the CAMPS database. *Proteins* 64: 906-922.
9. Oberai, A., *et al.* 2006. A limited universe of membrane protein families and folds. *Protein Sci.* 15: 1723-1734.
10. Viklund, H. and Elofsson, A. 2004. Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.* 13: 1908-1917.
11. Bagos, P.G., *et al.* 2005. Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics* 6: 7.
12. Altschul, S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
13. Eddy, S.R. 1995. Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3: 114-120.
14. Thompson, J.D., *et al.* 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.
15. Przybylski, D. and Rost, B. 2002. Alignments grow, secondary structure prediction improves. *Proteins* 46: 197-205.
16. Rost, B. 2001. Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.* 134: 204-218.
17. Kihara, D. 2005. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci.* 14: 1955-1963.
18. Krogh, A. and Riis, S.K. 1996. Prediction of beta sheets in proteins. In *Advances in Neural Information Processing Systems 8* (eds. Touretzky, D.S., *et al.*), pp. 917-923. MIT Press, Cambridge, USA.
19. Mamitsuka, H. and Abe, N. 1994. Predicting location and structure of beta-sheet regions using stochastic tree grammars. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2: 276-284.
20. Waldispuhl, J., *et al.* 2006. Predicting transmembrane beta-barrels and interstrand residue interactions from sequence. *Proteins* 65: 61-74.
21. Punta, M. and Rost, B. 2005. PROFcon: novel prediction of long-range contacts. *Bioinformatics* 21: 2960-2968.
22. Vullo, A. and Frasconi, P. 2003. Prediction of protein coarse contact maps. *J. Bioinform. Comput. Biol.* 1: 411-431.
23. Vullo, A. and Frasconi, P. 2002. A bi-recursive neural network architecture for the prediction of protein coarse contact maps. *Proc. IEEE Comput. Soc. Bioinform. Conf.* 1: 187-196.
24. Pollastri, G. and Baldi, P. 2002. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 18: S62-70.
25. Tusnady, G.E., *et al.* 2004. Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* 20: 2964-2972.
26. Tusnady, G.E., *et al.* 2005. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.* 33: D275-278.

27. Berman, H.M., *et al.* 2002. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* 58: 899-907.
28. Hobohm, U., *et al.* 1992. Selection of representative protein data sets. *Protein Sci.* 1: 409-417.
29. Baldi, P., *et al.* 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16: 412-424.
30. Zemla, A., *et al.* 1999. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins* 34: 220-223.
31. von Bertalanffy, L. 1938. A quantitative theory of organic growth (inquiries on growth laws. II). *Human Biol.* 10: 181-213.
32. Chou, P.Y. and Fasman, G.D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* 47: 45-148.
33. Garnier, J., *et al.* 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120: 97-120.
34. Levin, J.M., *et al.* 1986. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.* 205: 303-308.
35. Deleage, G. and Roux, B. 1987. An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* 1: 289-294.
36. Gibrat, J.F., *et al.* 1987. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* 198: 425-443.
37. Zvelebil, M.J., *et al.* 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* 195: 957-961.
38. Gascuel, O. and Golmard, J.L. 1988. A simple method for predicting the secondary structure of globular proteins: implications and accuracy. *Comput. Appl. Biosci.* 4: 357-365.
39. Qian, N. and Sejnowski, T.J. 1988. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* 202: 865-884.
40. Holley, L.H. and Karplus, M. 1989. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. USA* 86: 152-156.
41. Shestopalov, B.V. 1990. Prediction of protein conformation using a doublet code method. *Mol. Biol. (Mosk.)* 24: 1117-1725.
42. Zhang, X., *et al.* 1992. Hybrid system for protein secondary structure prediction. *J. Mol. Biol.* 225: 1049-1063.
43. Reczko, M. 1993. Protein secondary structure prediction with partially recurrent neural networks. *SAR QSAR Environ. Res.* 1: 153-159.
44. Rost, B. and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232: 584-599.
45. Yi, T.M. and Lander, E.S. 1993. Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.* 232: 1117-1129.
46. Chandonia, J.M. and Karplus, M. 1996. The importance of larger data sets for protein secondary structure prediction with neural networks. *Protein Sci.* 5: 768-774.
47. Garnier, J., *et al.* 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* 266: 540-553.
48. Riis, S.K. and Krogh, A. 1996. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol.* 3: 163-183.
49. Frishman, D. and Argos, P. 1997. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 27: 329-335.
50. Ito, M., *et al.* 1997. Prediction of protein secondary structure using the 3D-1D compatibility algorithm. *Comput. Appl. Biosci.* 13: 415-424.
51. Rychlewski, L. and Godzik, A. 1997. Secondary structure prediction using segment similarity. *Protein Eng.* 10: 1143-1153.
52. Salamov, A.A. and Solovyev, V.V. 1997. Protein secondary structure prediction using local alignments. *J. Mol. Biol.* 268: 31-36.
53. Thompson, M.J. and Goldstein, R.A. 1997. Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. *Protein Sci.* 6: 1963-1975.
54. Baldi, P., *et al.* 1999. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 15: 937-946.
55. Chandonia, J.M. and Karplus, M. 1999. New methods for accurate prediction of protein secondary structure. *Proteins* 35: 293-306.
56. Cuff, J.A. and Barton, G.J. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 34: 508-519.
57. Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292: 195-202.
58. Cuff, J.A. and Barton, G.J. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40: 502-511.
59. Ouali, M. and King, R.D. 2000. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.* 9: 1162-1176.
60. Petersen, T.N., *et al.* 2000. Prediction of protein secondary structure at 80% accuracy. *Proteins* 41: 17-20.
61. Schmidler, S.C., *et al.* 2000. Bayesian segmentation of protein secondary structure. *J. Comput. Biol.* 7: 233-248.

62. Hua, S. and Sun, Z. 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.* 308: 397-407.
63. Pan, X.M. 2001. Multiple linear regression for protein secondary structure prediction. *Proteins* 43: 256-259.
64. Kloczkowski, A., *et al.* 2002. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins* 49: 154-166.
65. Pollastri, G., *et al.* 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* 47: 228-235.
66. Kim, H. and Park, H. 2003. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng.* 16: 553-560.
67. Nguyen, M.N. and Rajapakse, J.C. 2003. Multi-class support vector machines for protein secondary structure prediction. *Genome Inform.* 14: 218-227.
68. Ward, J.J., *et al.* 2003. Secondary structure prediction with support vector machines. *Bioinformatics* 19: 1650-1655.
69. Guo, J., *et al.* 2004. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins* 54: 738-743.
70. Liu, X., *et al.* 2004. Prediction of protein secondary structure based on residue pairs. *J. Bioinform. Comput. Biol.* 2: 343-352.
71. Liu, Y., *et al.* 2004. Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics* 20: 3099-3107.
72. Wang, L.H., *et al.* 2004. Predicting protein secondary structure by a support vector machine based on a new coding scheme. *Genome Inform.* 15: 181-190.
73. Wood, M.J. and Hirst, J.D. 2004. Predicting protein secondary structure by cascade-correlation neural networks. *Bioinformatics* 20: 419-420.
74. Lin, K., *et al.* 2005. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* 21: 152-159.
75. Adamczak, R., *et al.* 2005. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* 59: 467-475.
76. Nguyen, M.N. and Rajapakse, J.C. 2005. Two-stage multi-class support vector machines to protein secondary structure prediction. *Pac. Symp. Biocomput.* pp. 346-357.
77. Pollastri, G. and McLysaght, A. 2005. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 21: 1719-1720.
78. Wood, M.J. and Hirst, J.D. 2005. Protein secondary structure prediction with dihedral angles. *Proteins* 59: 476-481.
79. Qin, S., *et al.* 2005. Predicting protein secondary structure and solvent accessibility with an improved multiple linear regression method. *Proteins* 61: 473-480.
80. Ceroni, A., *et al.* 2005. Learning protein secondary structure from sequential and relational data. *Neural Netw.* 18: 1029-1039.
81. Sadeghi, M., *et al.* 2005. Prediction of protein secondary structure based on residue pair types and conformational states using dynamic programming algorithm. *FEBS Lett.* 579: 3397-3400.