Reliability of transmembrane predictions in whole-genome data

Lukas Käll, Erik L.L. Sonnhammer*

Center for Genomics and Bioinformatics, Karolinska Institutet, 17177 Stockholm, Sweden

Received 15 August 2002; revised 15 November 2002; accepted 18 November 2002

First published online 27 November 2002

Edited by Robert B. Russell

Abstract Transmembrane prediction methods are generally benchmarked on a set of proteins with experimentally verified topology. We have investigated if the accuracy measured on such datasets can be expected in an unbiased genomic analysis, or if there is a bias towards 'easily predictable' proteins in the benchmark datasets. As a measurement of accuracy, the concordance of the results from five different prediction methods was used (TMHMM, PHD, HMMTOP, MEMSAT, and TOPPRED). The benchmark dataset showed significantly higher levels (up to five times) of agreement between different methods than in 10 tested genomes. We have also analyzed which programs are most prone to make mispredictions by measuring the frequency of one-out-of-five disagreeing predictions.

Key words: Membrane protein; Topology; Prediction; Reliability

1. Introduction

One of the most common analyses of a newly determined protein sequence is prediction of transmembrane (TM) segments. The presence of such segments suggests a membrane-bound location and is therefore a major determinant of the protein's function. Several programs exist for predicting TM segments and prediction accuracies have been reported as high as 79% correct topology [1–3].

Given the high reported accuracy of these programs, they are potentially of great value for annotating predicted protein sequences in completely sequenced genomes. But how high is the prediction accuracy in a set of genomic sequences? It is likely that the test sets of carefully studied proteins are substantially easier to predict correctly than a complete genomic set. Benchmark datasets used for assessing accuracy of TM prediction programs also make the problem simpler by only including proteins with one or more TM segments. There is furthermore a bias towards proteins with a small number of segments because the experimental data is generally less ambiguous.

We cannot directly assess the accuracy of TM predictions in genomic data, due to the unavailability of classed material. However, we can get a measure of the reliability of the predictions by looking at the correlation between multiple prediction methods. A previous study of TM prediction method consensus carried out on a subset of the *Escherichia coli* genome indicates that the reliability of a TM segment prediction correlates positively with the number of disparate methods producing the same result [4].

2. Materials and methods

We used five methods, TMHMM version 2.0 [5,6], PHDhtm version 2.1 [7,8], HMMTOP version 1.0 [9], MEMSAT version 1.8 [10], and TOPPRED2 version 1.0 [11,12]. We applied some restrictions in how we ran two of the methods. We were running PHDhtm in single-sequence mode. And when running MEMSAT we were counting topologies with negative score as a non-TM protein prediction.

The methods predicted TM segments for all proteins in the genomic datasets of four eukaryotes: *Homo sapiens* (45374 proteins), *Drosophila melanogaster* (14100 proteins), *Caenorhabditis elegans* (19101 proteins) and *Saccharomyces cerevisiae* (6334 proteins), two Archaea: *Sulfolobus tokodaii* (2826 proteins) and *Pyrococcus abyssi* (1765 proteins), three Gram-positive bacteria: *Bacillus subtilis* (4100 proteins), *Streptococcus pneumoniae* (2043 proteins) and *Staphylococcus aureus* (2624 proteins), and one Gram-negative bacteria: *E. coli* (4289 proteins), as well as in a dataset with 160 well-characterized TM proteins (we will refer to this dataset as the benchmark dataset hereon). The genomes were downloaded from ftp://ftp.ncbi.nih.gov/ on the 2nd of May 2002. The benchmark dataset is the one used as training set for TMHMM[6], and is available at http://www.binf.ku.dk/krogh/TMHMM/. Highly overlapping datasets have been used for training and benchmarking other methods. The datasets were run in the condition they where downloaded, and no preprocessing step to scale away signal peptides or other features was applied.

As a measure of correlation between different methods we count how many methods agree on their prediction. We define agreeing predictions such that all TM segments overlap with at least one residue with another TM segment prediction, and that the orientation of the loops (cytoplasmic side or not) of all the predictions coincide. Specifically we observe the largest consensus group (LCG), i.e. the group with the largest number of methods. In cases when this definition is not unique, when all methods disagree or if we have a 2-2-1 grouping, we select the LCG to be the consensus group containing TMHMM. If TMHMM happened to be the single method in a 2-2-1 configuration, the group containing HMMTOP is selected as the LCG.

3. Results and discussion

The output from the five methods was analyzed and the number of methods in the LCG was extracted for the each gene in the different datasets (Fig. 1). There are two ways to use the predictions: either by counting only predictions with one or more TM segments, or including also 'empty' predictions with no TM segments. Since the benchmark dataset only contains proteins with one or more TM segments, we first only considered such predictions (Fig. 1a), while a comparison between the different genomic predictions using also zero-TM segment predictions is shown in Fig. 1b. The percentages of concordance are much higher when including zero-TM seg-

*Corresponding author. Fax: (46)-8-337983.
E-mail address: erik.sonnhammer@cgb.ki.se (E.L.L. Sonnhammer).

PII: S0014-5793(02)03730-4

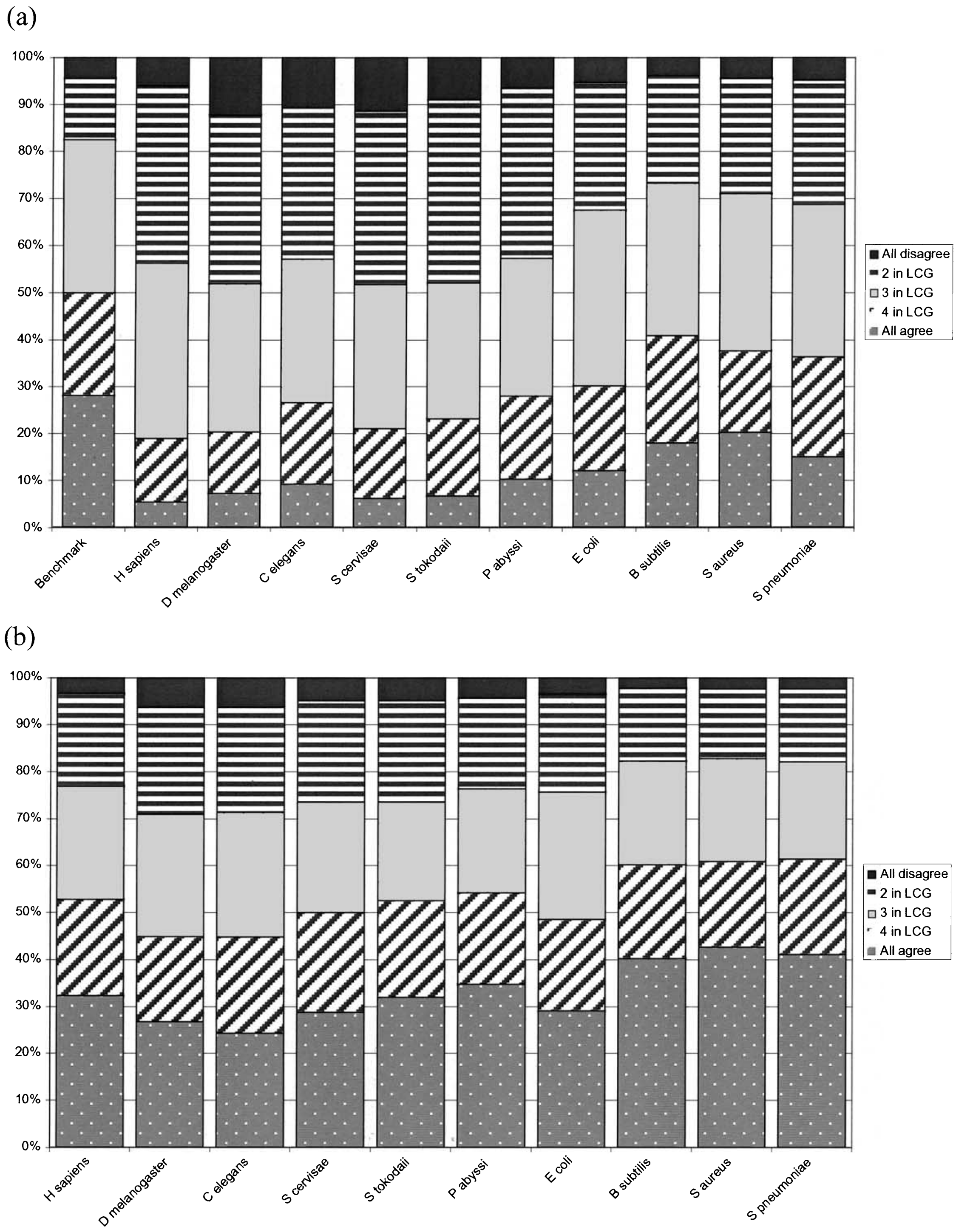

Fig. 1. Method concordance (a) when only counting proteins where the LCG predicts one or more TM segments and (b) when examining all of the proteins of the examined dataset. 'Benchmark' refers to the dataset with well-characterized TM proteins. The Benchmark dataset does not contain any proteins without TM segments, hence it is excluded from panel b.

ment predictions, suggesting that such predictions are relatively reliable.

It is clear from Fig. 1a that the methods produce a more correlated prediction when working on the benchmark dataset than when working on the genomic datasets. As we argued above, the correlation can be seen as an indication of the individual methods' reliability, and we therefore must lower our expectations when working with these methods on genomic data. Especially for eukaryotic data the method concordance is much lower than for the benchmark dataset. It is however hard to quantify a more realistic accuracy figure than the ones reported in previous studies. On one hand one could argue that this kind of correlation study would tell more about the worst performing method than the best. It could on the other hand be argued that the difference between benchmark and genomic dataset performance is even larger

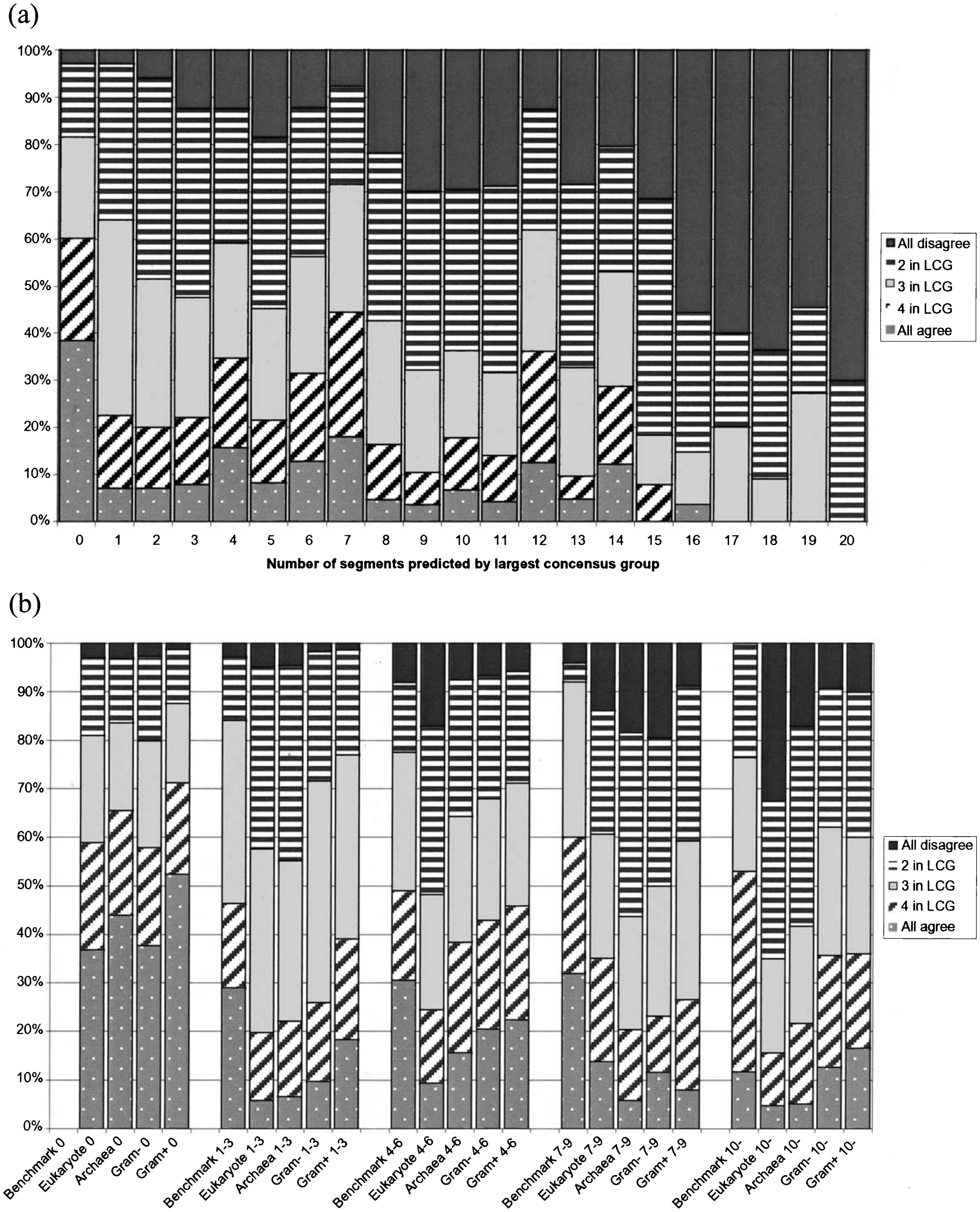

Fig. 2. Method concordance as a function of the number of TM segments for all the examined genomes together (a) and for different taxonomic groups (b). In panel b, ranges of the number of TM segments predicted by the LCG are used. For instance, 'Gram+4-6' means the genes predicted to contain 4–6 TM segments in the examined Gram-positive genomes. Note that the benchmark set does not contain any proteins without TM regions.

in reality, due to a built-in correlation of the methods stemming from their training on similar datasets. In any case, we have shown that the accuracy reported on the benchmark set is highly inflated. The level of 'all methods agree' was 28% for the benchmark set, and between 5% and 20% for the genomic datasets. In general, the concordance was lowest for eukaryotes and archaea, and highest for bacteria.

We looked for explanations for this observation and have examined the length distribution of the hydrophobic regions in the predicted TM segments and the distribution of the predicted number of TM segments for different species but could not find any trends that would explain the difference (data not shown). The discrepancy could simply be an effect of higher diversity of TM protein structure in eukaryotes and archaea than in bacteria.

We also investigated the methods concordance for different numbers of predicted TM segments. The results are shown both for all genomes together (Fig. 2a) or separated by taxonomic groups (Fig. 2b). It is worth noting that the downwards trend is sometimes broken at certain numbers of predicted

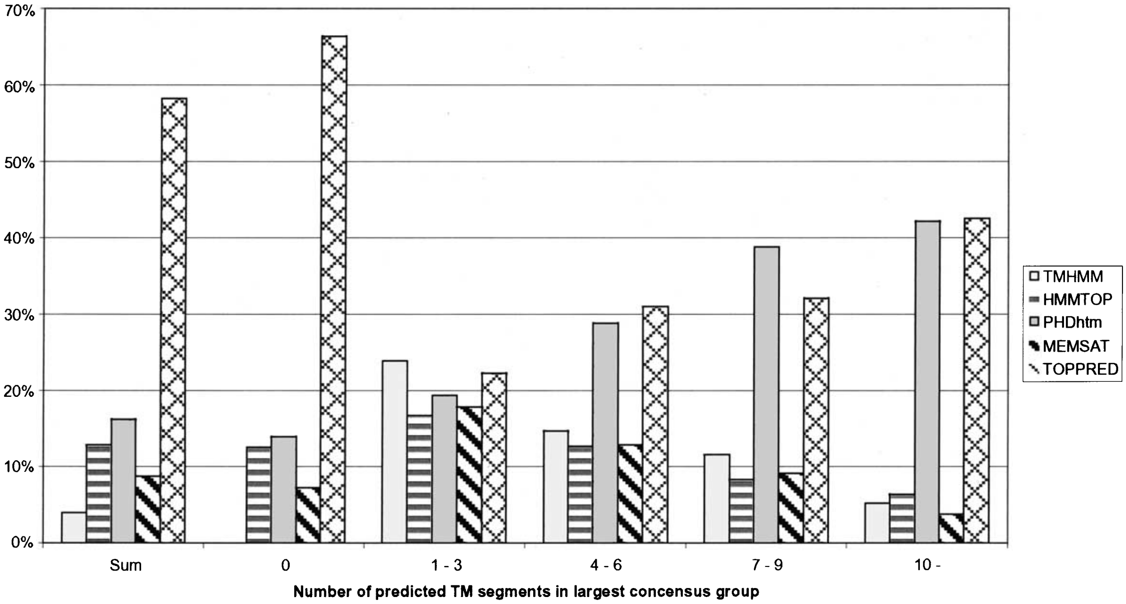

Fig. 3. Singled out (single disagreeing) method when the four other methods agree both in total (Sum) and as a function of the number of TM segments in the LCG. Note the absence of TMHMM as singled out when the LCG has no TM segments. This can be seen as an indication that TMHMM has a higher selectivity than the other methods, i.e. it is more restrictive in classifying a protein as a TM protein compared to the other methods.

TM segments, e.g. at four, seven or 12 segments. This could be an effect of having had more such examples when training the method, or simply that such topologies are easier to predict.

Are there large differences in the reliability of the individual methods? A comparison between the methods was made by analyzing proteins from all of the examined genomes that had an LCG size of for four out of five methods, and by determining which method was the 'odd one out', or singled out (Fig. 3). If the errors of the methods were to be treated as independent of each other, the number of times a method appears as singled out would be strongly correlated to the number of errors the method makes. However, as mentioned above there is a common bias for the examined methods and we have to see this measurement as a rough indication of the number of errors. One clear conclusion that can be drawn from the results is that TMHMM is the most selective method for avoiding false positive predictions. In only one out of the 15615 genes having an LCG with no segments was TMHMM singled out. Overall, TOPPRED is by far most frequently singled out while TMHMM is the least singled out method. PHDhtm is frequently singled out for large numbers of TM segments.

Acknowledgements: This work was supported by grants from Pharmacia Corporation and from the Swedish Knowledge Foundation.

References

[1] Moller, S., Croning, M.D. and Apweiler, R. (2002) Bioinformatics 18, 218.
[2] Moller, S., Croning, M.D. and Apweiler, R. (2001) Bioinformatics 17, 646–653.
[3] Ikeda, M., Arai, M., Lao, D.M. and Shimizu, T. (2002) In Silico Biol. 2, 19–33.
[4] Nilsson, J., Persson, B. and von Heijne, G. (2000) FEBS Lett. 486, 267–269.
[5] Sonnhammer, E.L., von Heijne, G. and Krogh, A. (1998) Proc. Int. Conf. Intell. Syst. Mol. Biol. 6, 175–182.
[6] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) J. Mol. Biol. 305, 567–580.
[7] Rost, B., Fariselli, P. and Casadio, R. (1996) Protein Sci. 5, 1704–1718.
[8] Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995) Protein Sci. 4, 521–533.
[9] Tusnady, G.E. and Simon, I. (1998) J. Mol. Biol. 283, 489–506.
[10] Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994) Biochemistry 33, 3038–3049.
[11] von Heijne, G. (1992) J. Mol. Biol. 225, 487–494.
[12] Claros, M.G. and von Heijne, G. (1994) Comput. Appl. Biosci. 10, 685–686.