Estimating the length of transmembrane helices using Z-coordinate predictions

COSTAS PAPALOUKAS,[1,2] ERIK GRANSETH,[1] HÅKAN VIKLUND,[1]
AND ARNE ELOFSSON[1]

[1]Stockholm Bioinformatics Center, Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden
[2]Department of Biological Applications and Technology, University of Ioannina, GR 451 10, Greece

(RECEIVED May 29, 2007; FINAL REVISION October 24, 2007; ACCEPTED October 31, 2007)

Abstract

Zpred2 is an improved version of ZPRED, a predictor for the Z-coordinates of α-helical membrane proteins, that is, the distance of the residues from the center of the membrane. Using principal component analysis and a set of neural networks, Zpred2 analyzes data extracted from the amino acid sequence, the predicted topology, and evolutionary profiles. Zpred2 achieves an average accuracy error of 2.18 Å (2.17 Å when an independent test set is used), an improvement by 15% compared to the previous version. We show that this accuracy is sufficient to enable the predictions of helix lengths with a correlation coefficient of 0.41. As a comparison, two state-of-the-art HMM-based topology prediction methods manage to predict the helix lengths with a correlation coefficient of less than 0.1. In addition, we applied Zpred2 to two other problems, the re-entrant region identification and model validation. Re-entrants were able to be detected with a certain consistency, but not better than with previous approaches, while incorrect models as well as mispredicted helices of transmembrane proteins could be distinguished based on the Z-coordinate predictions.

Keywords: membrane proteins; computational analysis of protein structure; protein structure prediction

Supplemental material: see www.proteinscience.org

Membrane proteins are essential for many biological processes such as cell-to-cell signaling and transporting substances into or out of the cell. With the growing number of membrane protein structures, it has been noticed that the diversity with respect both to structure and function of these proteins is of the same magnitude as for globular proteins. Still, they are more difficult to study experimentally than the globular proteins. The records in the Protein Data Bank (PDB) (Berman et al. 2000) can exemplify this, where membrane proteins are clearly underrepresented. In a typical genome, ~20%–25% of the proteins are transmembrane (TM) (Krogh et al. 2001), while in PDB, <1% are of this class. Although this number is still small, there is a steady increase in the number of available TM-structures, for instance, 90 out of 168 structures and 29 out of 56 superfamilies have been deposited into PDB since 2004.

In α-helical membrane proteins, the most notable feature is the existence of long hydrophobic membrane helices. These helices are normally about 20 residues long and are mainly characterized by their hydrophobicity, as well as other features of the primary sequence. However, with the increased number of available 3D structures of TM proteins, it has become clear that the helices actually show a significant variation in their length, slope, and straightness. Bowie and coworkers have shown that many helices contain a kink (Yohannan et al. 2004). In addition, the length of the helices, as defined from their entry points into the membrane, varies from less than 20 residues to up

Reprint requests to: Arne Elofsson, Stockholm Bioinformatics Center, Center for Biomembrane Research, Stockholm University, SE-106 91 Stockholm, Sweden; e-mail: arne@bioinfo.se; fax: 46-8-153679.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.073036108.

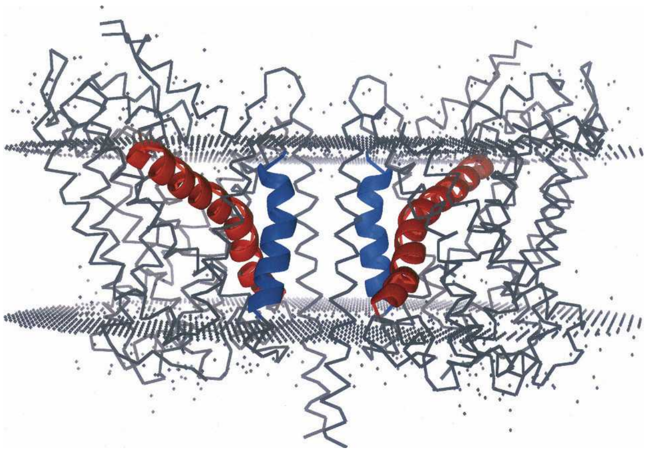

Figure 1. Example of long and short helices in the glutamate receptor leutaa (PDB ID: 2A65), a bacterial homolog of the Na$^+$/Cl-dependent neurotransmitter transporters. (Red) Helix number 2 with a length of 38 residues; (blue) helix number 8 with a length of 20 residues.

to 40 residues. As an example, the bacterial homolog of the Na/Cl neurotransmitter transporters contains a 38-residue-long helix (Fig. 1).

Computer-based methods are capable of overcoming the experimental hurdles that exist for TM proteins and extract new knowledge about their structure and function. Various approaches have been proposed in recent years for predicting the structural features of TM proteins. Topology predictors of TM proteins are based mainly on hidden Markov models (HMMs), artificial neural networks (ANNs), or other machine learning schemes. These systems predict the location of the membrane-spanning regions as well as the orientation of the protein in the membrane.

One of the first topology prediction methods, TopPred, combined hydrophobicity plots with the positive-inside rule (i.e., positively charged residues tend to be on the intracellular side) to generate a topology model (von Heijne 1992). A significant improvement was obtained after the introduction of ANNs and HMMs. The first ANNs made use of evolutionary profiles and were also combined with a dynamic programming-like algorithm to optimize the predictions generated by the network (Rost et al. 1996). On the other hand, in the HMM approaches, the various TM protein regions are modeled as separate compartments (e.g., middle of a helix, helix caps, regions close to the membrane, and globular domains) (Krogh et al. 2001; Tusnády and Simon 2001). Certain grammar rules are also used to predict the topology from the amino acid sequence. The addition of evolutionary information extracted from multiple sequence alignments, used in the form of profile-based HMMs, further increased the prediction accuracy (Viklund and Elofsson 2004). Moreover, homology information extracted from global sequence alignments has been used when decoding HMMs, in the form of probabilities of sequence features, in order to provide more specific data during prediction (Käll et al. 2005). Like the HMMs, ANNs have also been trained to encode different protein regions like cytoplasmic, non-cytoplasmic, transmembrane, and signal peptides, improving in that way the topology predictions (Jones 2007). Finally, several other topology predictors for TM proteins have been developed ranging from simple to more sophisticated ones (for review, see Elofsson and von Heijne 2007). Furthermore, some of the latest systems use a set of previously developed efficient predictors in order to produce a consensus prediction (Amico et al. 2006).

Although topology predictions have improved significantly during the last few years and now reach accuracies of ~65%–80%, depending on the data set, they actually do not provide very detailed information about TM proteins at the residue level. The identification of the membrane residues is quite accurate (~90%), but no significant increase has been obtained between the TOPPRED method dated at 1992 (von Heijne 1992) and the more recent state-of-the-art HMM methods (Viklund and Elofsson 2004) in determining the residue disposition, although the topology accuracy has increased from 39% to 66%. Similarly, in TMHMM (Krogh et al. 2001), it has been noticed that the length of the TM helices is not accurately predicted, as most helices are predicted to be 21 residues long (Käll et al. 2005). One reason for the lack of residue-based accuracy is that most information about TM protein structures is of quite low resolution. However, with the recent increase in the number of available three-dimensional (3D) structures of membrane proteins, this has changed.

Recently a different approach from the topology predictors was proposed, namely, determining the distance of each residue from the membrane center, that is, predicting the Z-coordinate (Granseth et al. 2006). Z-coordinate prediction can be considered as an intermediate step toward full 3D structure prediction, while it can also provide detailed residue information. For example, structural details such as re-entrant (Viklund et al. 2006) or interfacial helices (Granseth et al. 2005), the tilt of TM helices, or the protrusion of a loop from the membrane could theoretically be identified based on the Z-coordinates.

In this study, the accuracy of predicting the Z-coordinates of TM proteins is improved by

1. The use of a pre-processing step through signal processing;
2. the utilization of a set of ANNs;
3. the addition of a recurrent input to the ANNs, which corresponds to the previously predicted Z-coordinate;
4. a post-processing step for the ANN output;
5. an overall better exploitation of the information combined with a better tuning of the ANNs.

These adjustments improve the accuracy of the Z-coordinate predictor to an average error rate of 2.18 Å. Most importantly, and based on this accuracy, we show that for the first time, we are able to predict the length of TM helices with a correlation coefficient of 0.41. Moreover, we can recognize with a certain consistency the re-entrant regions in a protein, while we can also identify incorrect models of TM proteins.

Results

Development of Zpred2

The original Z-coordinate predictor (Granseth et al. 2006) is based on a single ANN that uses an evolutionary profile and a predicted topology, from PRODIV-TMHMM (Viklund and Elofsson 2004), to predict the Z-coordinate of a residue. During the development of Zpred2, it was found that there were several parameters that affected its performance and therefore needed tuning. The optimization of these parameters is briefly described below, and the results of these improvements are found in Table 1. The overall architecture of Zpred2 is depicted in Figure 2.

As in ZPRED, topology information is generated from PRODIV-TMHMM, and evolutionary information is obtained from the position-specific scoring matrices (PSSMs) generated from one iteration of PSI-BLAST (Altschul et al. 1997). To improve the predictor, several different encoding schemes were tried (Kawashima and Kanehisa 2000). After testing several transformation tables, an encoding scheme with only the polarity was found to be optimal. For each residue and for the three types of input data, a window of values around it is considered (i.e., the actual values of this window are being used), which is fed to the networks for analysis. Window lengths from 3 up to 31 were tried, and it was found that a slight improvement was obtained if the input window was set to 25 residues.

Table 1. *Average prediction accuracy errors*

Method	All	TM-dense	TM-sparse
Original ZPRED	2.55	2.80	2.04
Encoding based on polarity	2.54	2.90	1.83
25-residues window	2.52	2.91	1.75
PCA	2.49	2.74	2.01
Previous Z-coordinate	2.46	2.72	1.95
Translate the reference Z-coordinates ($z = z + 0.5$)	2.44	2.69	1.94
Class division	2.38	2.72	1.70
Set of ANNs	2.30	2.68	1.54
Post-process (3–27)	2.18	2.59	1.36
Correct topology	1.92	2.19	1.38
OPM data set	*2.17*	*2.43*	*0.74*

Errors are for the eight steps of improvement during the development of Zpred2. Each row depicts progressively the prediction amelioration when additional improvements are incorporated. Additionally, the last two rows contain the obtained accuracies when the correct topologies are used, instead of the predicted ones, and when the OPM data set is used, respectively. Besides the overall accuracy error, results are also given for the two protein classes in our data set, the TM-dense and the TM-sparse.

For typical values of window lengths, this yields a very high vector dimension, which might be undesirable since this creates generalization problems during the training process (Hagan et al. 1996). Therefore, PCA is used as a dimensionality reduction technique (Fig. 2). The best results were obtained using 13 PCA vectors, describing 99% of the variation in the original data.

The next step in the optimization was to include information about the previously predicted Z-coordinate and in that way avoid any large fluctuations in the Z-coordinate predictions. Another issue that occurred during testing was the "value shifting" in the data obtained from the translation process of the PDB 3D coordinates to the reference (true) Z-coordinates. In some cases, it was observed that if the predicted coordinates from Zpred2 were slightly shifted upward or downward, they could better fit the true ones. This shifting difference could be justified from the application of the translation algorithm, TMDET (Tusnády et al. 2005), since this algorithm exploits the hydrophobicity profiles of the proteins and in some cases likely translates the coordinates slightly off the true ones.

Besides that, we noticed that among the 101 sequences there was a large diversity in the Z-coordinates' "waveforms," which could basically be grouped into two major classes: those that follow a general sinusoidal pattern and those that do not (see Fig. 3). These two classes of proteins are referred to in the following as TM-dense and TM-sparse, respectively. By training separate networks for each class, a significant improvement is achieved, mainly for the residues classified as globular, that is, those that have a Z-coordinate of 25 Å or more. Their mean error drops from 1.47 Å to 1.23 Å, while in the TM-sparse group only, their mean error drops from 1.29 Å to 0.72 Å. Another major improvement of the proposed Z-predictor derives from the use of three different ANNs instead of a single one.

In the final step, post-processing is applied to the ANNs' output following a straightforward procedure that takes the average value of the network outputs and transforms it to the interval of interest, that is, to a value between 5 and 25 Å. More specifically, we trained the ANNs with target values from 3 to 27 Å and then translated their outputs to the desired interval.

From Table 1 we can see that there are three major factors that contribute to the performance improvement. Specifically, the division of the proteins in two classes improves the predictions by 0.06 Å, the use of a set of ANNs by 0.08 Å, while the adjustment of training the ANNs with targets between 3 and 27 Å instead of 5 and 25 Å improves the predictions by 0.12 Å. All the other five steps contribute 0.11 Å, in total. We should mention that

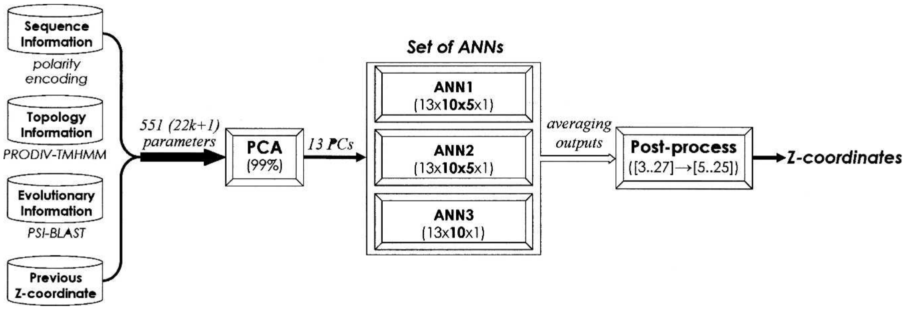

Figure 2. Overall architecture of the improved Z-predictor (Zpred2). The input data are processed through PCA in order to reduce their dimension. The obtained principal components are fed separately into three different ANNs, each one specialized to a certain subgroup of proteins. The output of the three networks is then averaged and subsequently post-processed to produce the final Z-coordinate predictions.

all prediction errors were estimated using fivefold cross-validation, where the original data set is partitioned into five subsets, and one of them is retained as the test set while the rest are used for training. This procedure is repeated five times in total so that all subsets are used once as the test set.

The final predictor

Table 1 contains the intermediate prediction accuracies that correspond to each one of the improvements incorporated in the current Z-predictor. When Zpred2 was applied with all the settings mentioned above, the obtained accuracy in predicting the Z-coordinates in the [5, 25] interval was 2.18 Å, compared to the 2.55 of the original ZPRED predictor. For each class of proteins, TM-dense and TM-sparse, the obtained accuracies were 2.59 Å and 1.36 Å, respectively. From Figure 4 it can be seen that the lower error for the second group of proteins is due mainly to the globular domains, which are predicted more accurately.

Moreover, the performance reached an overall accuracy of 1.92 Å when the true topologies (derived directly from the reference Z-coordinates) were used instead of those predicted from PRODIV-TMHMM. This indicates that as the performance of future one-dimensional (1D) topology predictors improves, a marginal improvement of Zpred2 will follow. Besides measuring the prediction accuracy error, we also did a linear regression analysis and measured the correlation between the predicted and the true Z-coordinates. For the 101 sequences in our data set, a correlation coefficient of 0.903 was obtained, while for the two groups, the corresponding values were 0.857 and 0.925, respectively.

Independent test set

Although Zpred2 was carefully cross-validated, it was further tested on an independent test set consisting of 45 proteins. In this data set, the average error was 2.17 Å despite the fact that the Z-coordinates were not calculated using the TMDET algorithm (Tusnády et al. 2005), but the one used in the OPM (Orientations of Proteins in

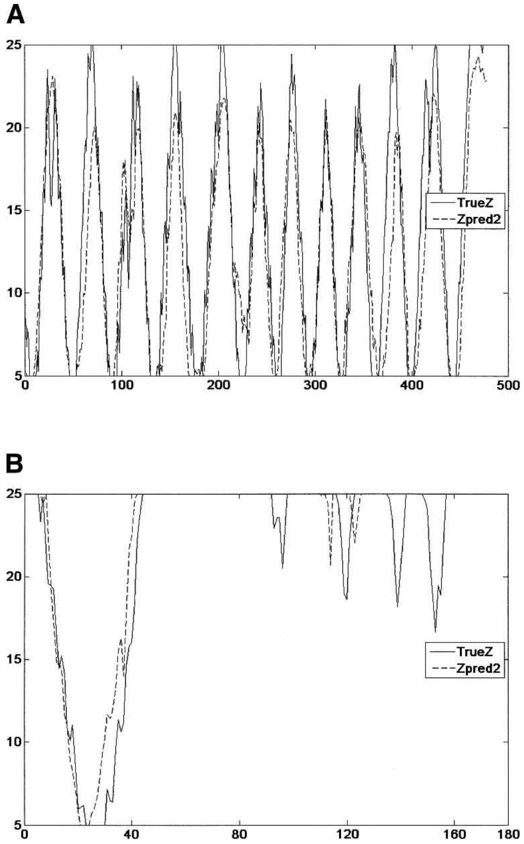

Figure 3. Example of the two main waveform patterns of the Z-coordinates in the 101 sequences of our data set: (*A*) TM-dense with the sinusoidal-like pattern and (*B*) TM-sparse where a large part is ≥25 Å. (Solid line) The true Z-coordinates are plotted; (dotted line) the predicted ones. For comparison, the two depicted sequences have 2.44 Å (*left*) and 1.19 Å (*right*) prediction errors, respectively.

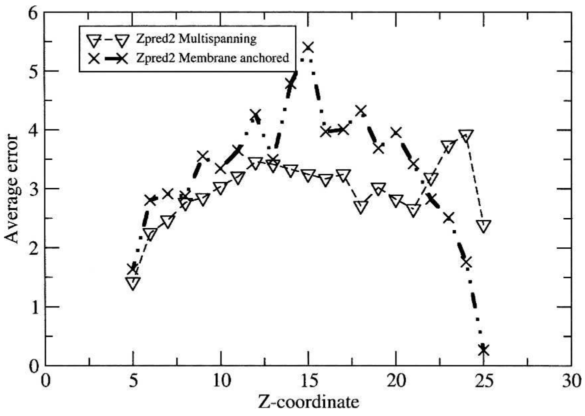

Figure 4. Average error of Zpred2 predictions plotted against the Z-coordinates. The prediction errors of the two classes of membrane proteins are shown. It can be seen that the prediction of globular domains (Z-coordinates ≥25 Å) is significantly better for the TM-sparse class, and this the main reason for the better performance on this class. The predictions of the most central residues in the membrane (~5 Å) are also better than the rest in both cases. This is more or less expected since it is possible to predict only higher and not much lower values for these.

Membranes) database (Lomize et al. 2006). It is worth mentioning that from the 101 sequences, 53 are contained also in the OPM database. For these, the average difference between the Z-coordinates determined by TMDET and those deposited in OPM is 1.74 Å.

Discussion

The average prediction accuracy error obtained for Zpred2 was quite low, 2.18 Å, but how can this be carried over to practical applications? The prediction of the Z-coordinates can provide additional information not available from standard topology predictors. Primarily it can be used to identify TM regions mis-predicted by the topology predictors. A common limitation of topology predictors is that they do not always predict the correct number of TM regions. The Z-coordinate predictions can be used as an indicator for the reliability of topology predictions. More specifically, from the predictions of Zpred2, it was found that actual TM residues not identified correctly by PRODIV-TMHMM (i.e., false negatives) had, on average, a lower predicted Z-coordinate (14.7 Å) than non-TM residues correctly labeled by PRODIV-TMHMM (21.1 Å). Similarly, residues that were mis-predicted to be in a TM region (i.e., false positives) had a higher predicted Z-coordinate (11.8 Å) than the correctly predicted TM residues (7.8 Å). Even though our initial attempts to make use of the Z-coordinate predictions to improve the predicted topology did not actually result in better topology predictions, we believe that this information is worthwhile to be exploited for future topology predictors.

Estimating the length of TM helices

Next we used Z-coordinates to predict the length of the helices. The exact location of a TM helix can either be defined based on the secondary structure or alternatively from the sites where the protein chain enters and exits the membrane environment. The second definition is in better agreement with topology predictions (data not shown) and is therefore used here. It is also in better agreement with the predicted helices from the topology predictors, as they are trained to recognize the hydrophobic segment and not the secondary structure state. Moreover, definition of a helix length purely based on Z-coordinates also has the advantage of simplicity, since determining the length from the predicted Z-coordinates is a rather straightforward task. Using this definition, the TM helices vary in length between 16 and 42 residues, but it should be noted that several of the longest helices were severely kinked.

In Table 2 we can see the length predictions from PRODIV-TMHMM, Phobius (Käll et al. 2004), and Zpred2. Zpred2 tends to predict slightly longer helices, while PRODIV-TMHMM to a certain degree predicts helices that are short. Conversely, from the small standard deviation of PRODIV-TMHMM, caused by its predictions of helices with a specific length, 21 residues (see also Fig. 5), was confirmed. In contrast, Phobius, which uses a different model to determine the length of the helices, has a similar standard deviation as the observed helix lengths. The average error in the length predictions for all methods is similar (2–3 residues), but a minor improvement on the actual prediction of the start and end positions is obtained with Zpred2. Strikingly, it can be observed that the correlation coefficient obtained for PRODIV-TMHMM and Phobius is low (<0.1), while it is higher for Zpred2 (0.41). Moreover, in the case where the approximate ends are known, the correlation coefficient for Zpred2 is further improved, reaching 0.62. Also, as can be seen in Figure 5, all helices that are predicted to be longer than 30 residues actually are. This indicates that using Zpred2 it is possible

Table 2. *Prediction results for the helix lengths using PRODIV-TMHMM, Phobius, and Zpred2*

Method	Correlation	Length error	Start error	Average length
True Z-coordinate	—	—	—	22.3 ± 3.3
PRODIV	0.08	2.1	2.7	21.1 ± 1.1
Phobius	0.07	2.6	2.8	22.1 ± 2.5
Zpred2	0.41	2.5	2.5	23.2 ± 3.2
Zpred2 on the OPM data set	*0.31*	*3.9*	*2.7*	*21.9 ± 6.0*

The correct helices were determined from the reference Z-coordinates. Besides the mean length, the mean prediction errors and the average shift of the predictions are also presented. The final row shows the performance on the independent OPM test set.

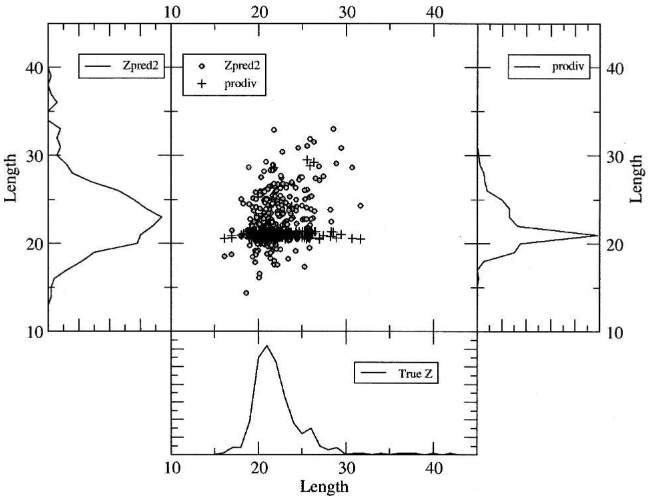

Figure 5. In the three side plots, the distribution of the helix lengths for Zpred2, the PDB files, and PRODIV-TMHMM are plotted. In the central plot, the actual lengths for each corresponding helix between the true (*X*-axis), Zpred2 (circles), and PRODIV-TMHMM (pluses) are shown. To avoid two dots ending up on top of one another, each dot is moved randomly ± half a unit in the *X* and *Y* directions. The correlation between the true Z-coordinates and Zpred2 is 0.41, while between the true and PRODIV-TMHMM, it is 0.08.

to reliably identify long TM helices. It is worth mentioning that when Zpred2 was applied to the OPM data, the corresponding helix lengths were estimated with a correlation coefficient of 0.31.

Other applications

It has been noted that many TM proteins contain re-entrant regions, that is, regions that start on one side of the membrane, enter the membrane, and then return to the same side of it. In the Z-coordinate representation, such regions have the waveform pattern of shorter peaks around the 15 Å level. Because of their smaller size, Zpred2 encounters some problems in predicting these regions accurately. In addition, even when they are predicted accurately, it is difficult to differentiate them from the helices, in particular when they go deeper than 5 Å. Nevertheless, by defining a set of structural rules based on the Z-coordinates, the length of the TM regions, and their intermediate distance, we identified 23 candidate regions out of which 12 were correctly predicted (positive predictive accuracy, PPA: 43.5%). By relaxing the threshold values used by the rules, more re-entrant regions were able to be detected but with a serious decrease in the PPA. In comparison, TOP-MOD, a system previously developed for re-entrant detection, correctly identified 17 out of the 36 re-entrant regions with a total of eight false-positive predictions (Viklund et al. 2006).

A third application of Z-coordinate prediction is to use it to identify incorrect models of TM proteins. Specifically, when there is a significantly larger deviation between the predicted and the observed Z-coordinates than observed in the training set, this might be an indication that the model is incorrect. Table 3 contains the average difference between the predicted and the observed Z-coordinates for erroneous and correct structures of three proteins, EmrE, MsbA, and KvAP. The two erroneous EmrE models have larger differences than expected, while in the MsbA case, the predicted Z-coordinates cannot be used to distinguish the correct from the incorrect models. Most likely the main errors in the MsbA models are from the helices placed at incorrect *XY* positions but not in the Z-coordinate plane, while the errors in the EmrE models include shifts in the Z-coordinates (Fleishman et al. 2006). As for the KvAP models, the 2A79 is in better agreement with the Z-coordinate predictions than the 1ORQ and 1ORS. This is also in agreement with previous results from Lee et al. (2005). However, we should note that 2A79 belongs to the TM-sparse class, as defined in the Zpred2 development section, and when taking this into account, the errors are comparable, that is, we cannot clearly distinguish between the KvAP models.

Future outlook

To further improve the Z-coordinate predictions, we believe that it is necessary first to obtain better topology predictions. Then again, even with 100% accurate topology predictions, Zpred2 would reach a prediction accuracy of 1.92 Å. But it should be noted that the average RMSD between the TMDET and OPM Z-coordinates is only slightly lower than that (1.74 Å). This

Table 3. *Model validation results*

	Average prediction accuracy error		
	Mean	Z-score	Class
MsbA			
1JSQ	1.33	0.2	TM-sparse
1PF4	1.32	0.2	TM-sparse
1Z2R	1.44	0.1	TM-sparse
1L7V	*1.57*	*0.0*	*TM-sparse*
2HYD	*1.62*	*0.0*	*TM-sparse*
EmrE			
1S7B	4.33	−1.9	TM-dense
2F2M	2.79	−0.4	TM-dense
KvAP			
1ORQ	3.19	−0.8	TM-dense
1ORS	2.96	−0.6	TM-dense
2A79	*1.94*	*−0.3*	*TM-sparse*

Results in terms of average and Z-score for the three sets of erroneous (and correct in italics) proteins using the predictions from Zpred2 and the data obtained from the TMDET algorithm. The corresponding class is also given for each model.

could indicate that the performance of Zpred2 is close to optimal. Therefore, we believe that it is necessary to use a more consistent transformation of the 3D structures in order to substantially improve the Z-coordinate predictions. Anyhow, even small improvements might provide significant improvements on the helix-length predictions.

Conclusions

In this study, Zpred2, an improved version of the original Z-predictor, was developed and applied to three different problems, in order to examine its practical value. Zpred2 uses the same types of information as ZPRED, but it incorporates them differently and more efficiently, thus reducing the average error from 2.55 to 2.18 Å. The dimensionality reduction through PCA, the class partitioning of the proteins, the use of a set of ANNs, and the post-processing are the steps that contribute most to the increase in performance. Moreover, the recurrent input of the previously predicted Z-coordinate as well as the primary sequence encoding further assist in improving the predictions.

Possibly the most valuable application of Zpred2 is the prediction of the helix length, as this is a problem where current topology prediction methods provide little or no information. Here, Zpred2 managed to predict the helices length with a correlation coefficient of 0.41. Although this accuracy is far from perfect, it can be seen as a first step toward detailed predictions of individual helices. In addition, it was found that Zpred2 can provide information about possible re-entrant regions and missed or overpredicted helices, as well as to identify dubious models of TM proteins.

Materials and Methods

Data sets

For training and testing Zpred2 in predicting Z-coordinates from the amino acid sequence, we used 46 PDB structures obtained by X-ray diffraction (see the Supplemental material). These constitute 101 nonhomologous protein chains with 21,589 residues in total. The same data set was also used to test the performance of Zpred2 in estimating the length of the helices. Additionally, 45 protein sequences were extracted from the OPM database in order to construct an independent test set. These sequences are characterized by <50% identity between them and the 101 sequences mentioned above. We should note that all tests concerning Zpred2 were made using MATLAB version 7.1.

Z-coordinate definitions

Since PDB does not contain any information concerning the membrane location, we applied the TMDET algorithm (Tusnády et al. 2005) to these files to obtain the reference (true) Z-coordinates. The TMDET algorithm finds the most probable localization of the membrane layer and, based on that, the PDB coordinates can be rotated and translated so as to determine the target Z-coordinates.

Network training

Predicting the Z-coordinates to their full extent, that is, from $-M$ to M, with M being a sufficiently large number, is a too complex problem. Therefore, further simplifications needed to be considered. One such simplification was to work with absolute values and predict values from 0 to M. However, we chose to consider a more limited interval of interest, specifically the 5–25 Å region (as in Granseth et al. 2006) and hereby focus on the region where the structural properties of α-helical membrane proteins change the most (White and Wimley 1999).

Furthermore, when including the previously predicted Z-coordinate, a starting value is required for the first training epoch. We examined the whole range of starting values (i.e., from 5 to 25), and the choice of 22 gave the best predictions, which actually coincides with the average of the starting values in our data set.

In what concerns ‘‘value shifting,’’ some of the structures could be translated a few angstroms along the Z-coordinate in order to better fit their hydrophobicity profiles. To study that effect, we did some tests by shifting with a constant value the coordinates obtained from TMDET for all sequences. Particularly, we used values from −1 Å to 1 Å, and, indeed, 0.5 gave slightly better results.

Profile generation

PSI-BLAST (Altschul et al. 1997) profiles were obtained after one round of searches against UniRef90 (Wu et al. 2006) and with the E-value cutoff set to 10^{-5}. The log-odds profiles obtained were subsequently normalized to values between zero and one using the logistic function: $1/(1 + e^{-x})$, in order to be in accordance with the other two types of input data (Jones 1999).

Sequence encoding

Several different encoding schemes were used, based on predetermined transformation tables with normalized values that correspond to different physicochemical properties of the 20 amino acids, including the polarity, the hydrophobicity, volume, molecular weight, buriability, electron–ion interaction potential, solvation free energy, HPLC parameter, and the steric parameter (Kawashima and Kanehisa 2000).

PCA

Before applying PCA we scaled the mean and standard deviation of the training set to zero and unity, respectively. This is a typical procedure with PCA that normalizes all data values. During our analysis, we examined different variation thresholds for the PCA ranging from 80% to 99%, while we also tested other signal processing methodologies for dimensionality reduction like the Fourier transform, parametric modeling, and wavelets, but PCA produced slightly better results.

Set of ANNs

Using a set of ANNs instead of a single one, the Z-coordinate prediction is obtained by averaging the outputs of the different networks. From the Z-coordinate definition, it is apparent that the prediction process can be seen as a function approximation problem for which ANNs are widely considered as an excellent tool among other machine learning approaches (Hagan et al. 1996). During the initial experiments, we realized that if we used only one ANN, the predicted Z-coordinates varied significantly between different tests with similar networks. For this reason, we used several ANNs with different architectures. Specifically, we used feed-forward neural networks, radial basis networks, recurrent networks, and adaptive ones, but the feed-forward ANNs consistently gave the best results; thus we tried using from one up to a combination of five of them. For their architecture, we used one or two hidden layers with five to 25 nodes in the first hidden layer and one to 10 in the second. As a training algorithm, we tried the resilient backpropagation, the scaled conjugate gradient, one-step secant, Levenberg-Marquardt, and Bayesian regularization. We also examined the case of early stopping, so we set the iteration epochs from 50 up to 1000.

Prediction of the length of the helices

To predict the length of a helix, the beginning and ending points have to be determined. From the topology predictions, this is implicit in the model, but from the Z-coordinate predictions, it is not. Therefore, we defined a simple set of rules using only the absolute (true or predicted) Z-coordinates to exactly define a helix. First, an approximate location of the helix endpoints was determined either from the topology predictions or from the 3D structure. Thereafter, a helix endpoint was defined to be the residue that passed through the membrane border, set at 15 Å, and within 10 residues from the true, or predicted, helix endpoint.

Re-entrants' identification

For the problem of detecting re-entrant regions, a slightly different data set was adopted. Seventy-nine chains derived from PDB were used with a homology between them of $\leq 30\%$ (Viklund et al. 2006). This set contains 36 cases of re-entrant regions and 302 TM regions. The determination of the re-entrant regions was realized based on rules that take into account structural features like the depth penetration (Viklund et al. 2006). As for Zpred2, from all the TM regions that it detects with the membrane border set at 15 Å, we consider only those with a length varying from five up to 40 residues. Additionally, the TMs that have more than nine consecutive residues with coordinates at 5 Å are also excluded (as being helices). The obtained regions are finally merged if they are closer than 15 residues to each other. A membrane limit of 14 Å is also used to define the end of each region; that is, a re-entrant region must end between 14 and 15 Å. We should also note that one re-entrant region was excluded from the tests since it was located at the beginning of the sequence, before the twelfth residue, where Zpred2 cannot be applied consistently because of the window implementation.

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 Acids Res.* **25:** 3389–3402.

Amico, M., Finelli, M., Rossi, I., Zauli, A., Elofsson, A., Viklund, H., von Heijne, G., Jones, D., Krogh, A., Fariselli, P., et al. 2006. PONGO: A web server for multiple predictions of all-α transmembrane proteins. *Nucleic Acids Res.* **34:** W169–W172. doi: 10.1093/nar/gkl208.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Elofsson, A. and von Heijne, G. 2007. Membrane protein structure: Prediction versus reality. *Annu. Rev. Biochem.* **76:** 125–140.

Fleishman, S.J., Harrington, S.E., Enosh, A., Halperin, D., Tate, C.G., and Ben-Tal, N. 2006. Quasi-symmetry in the cryo-EM structure of EmrE provides the key to modeling its transmembrane domain. *J. Mol. Biol.* **364:** 54–67.

Granseth, E., von Heijne, G., and Elofsson, A. 2005. A study of the membrane–water interface region of membrane proteins. *J. Mol. Biol.* **346:** 377–385.

Granseth, E., Viklund, H., and Elofsson, A. 2006. ZPRED: Predicting the distance to the membrane center for residues in α-helical membrane proteins. *Bioinformatics* **22:** e191–e196. doi: 10.1093/bioinformatics/btl206.

Hagan, M.T., Demuth, H.B., and Beale, M.H. 1996. *Neural network design*. PWS Publishing, Boston, MA.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292:** 195–202.

Jones, D.T. 2007. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **23:** 538–544. doi: 10.1093/bioinformatics/btl677.

Käll, L., Krogh, A., and Sonnhammer, E.L. 2004. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338:** 1027–1036.

Käll, L., Krogh, A., and Sonnhammer, E.L. 2005. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* **21:** i251–i257. doi: 10.1093/bioinformatics/bti1014.

Kawashima, S. and Kanehisa, M. 2000. AAindex: Amino acid index database. *Nucleic Acids Res.* **28:** 374.

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.

Lee, S.Y., Lee, A., Chen, J., and MacKinnon, R. 2005. Structure of the KvAP voltage-dependent K^+ channel and its dependence on the lipid membrane. *Proc. Natl. Acad. Sci.* **102:** 15441–15446.

Lomize, M.A., Lomize, A.L., Pogozheva, I.D., and Mosberg, H.I. 2006. OPM: Orientations of proteins in membranes database. *Bioinformatics* **22:** 623–625. doi: 10.1093/bioinformatics/btk023.

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

Tusnády, G.E. and Simon, I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17:** 849–850.

Tusnády, G.E., Dosztanyi, Z., and Simon, I. 2005. TMDET: Web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics* **21:** 1276–1277.

Viklund, H. and Elofsson, A. 2004. Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.* **13:** 1908–1917.

Viklund, H., Granseth, E., and Elofsson, A. 2006. Structural classification and prediction of reentrant regions in α-helical transmembrane proteins: Application to complete genomes. *J. Mol. Biol.* **361:** 591–603.

von Heijne, G. 1992. Membrane protein structure prediction. *J. Mol. Biol.* **255:** 487–494.

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

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., et al. 2006. The Universal Protein Resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res.* **34:** D187–D191. doi: 10.1093/nar/gkj161.

Yohannan, S., Faham, S., Yang, D., Whitelegge, J.P., and Bowie, J.U. 2004. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl. Acad. Sci.* **101:** 959–963.