COMPUTATIONAL METHODS FOR PREDICTING TRANSMEMBRANE ALPHA HELICES

NORMAN HOVIJITRA

BIOC 218: COMPUTATIONAL MOLECULAR BIOLOGY

FINAL PROJECT

DECEMBER 6TH, 2002

Introduction:

Protein crystal structures provide tremendous insight into the complex nature of proteins because they allow for a greater understanding of how a specific protein works and also facilitates the understanding of protein homologues. Their importance is illustrated by the high occurrence of journal articles publishing protein structures in highly respected journals such as *Science* and *Nature*. While the number of known protein crystal structures has grown significantly over recent years, obtaining crystal structures for membrane proteins is fundamentally more difficult than for cytoplasmic proteins and so are there are relatively fewer solved crystal structures for membrane proteins than cytoplasmic proteins. As of December 2002, there are 17395 Protein Data Bank structures for proteins solved by both NMR and X-Ray crystallography and only 57 are integral membrane proteins. Approximately 30% of an organism's genome codes for a membrane protein, while only 0.3% of known structures are structures of membrane proteins. This discrepancy shows that there is a great need for being able to accurately predict the structure of membrane proteins in particular.

The need to predict protein structure in general has been met with large amounts of effort being placed on accurately predicting protein tertiary structure.[1,2] A more simple step towards being able to predict protein structures is to predict the secondary structure based on known primary structure. This venture has had significant gains over previous years and has been utilized to identify secondary structures like alpha helices and beta strands.[3] There have actually been a large number of methods developed to predict protein secondary structure and there is even a program on the web called EVA at http://cubic.bioc.columbia.edu/eva/ that actually automatically evaluates the accuracy of these methods as new protein crystal structures are released! Unfortunately, these methods are not trained to discriminate between membrane proteins and globular (cytoplasmic) proteins and do not evaluate methods used to specifically predict membrane protein structures.

Predicting the structures of membrane proteins is obviously a less general case than predicting the structures of proteins, but there is a significant deficiency of membrane proteins structures and thus, focus should be given to specializing in membrane proteins because the environments in which globular proteins and membrane proteins exist are quite distinct. Fortunately, researchers have focused on predicting membrane protein structures and the amount of work spent in this area is not lacking.[4] The focus of this paper is to give a brief review of the methods that are used to identify membrane protein

---

1 Al-Lazikani, B., J. Jung, et al. (2001). "Protein structure prediction." Current Opinion in Chemical Biology 5(1): 51-56.
2 Moult, J. (1999). "Predicting protein three-dimensional structure." Current Opinion in Biotechnology 10(6): 583-588.
3 Rost, B. (2001). "Review: Protein secondary structure prediction continues to rise." Journal of Structural Biology 134(2-3): 204-218.
4 Rost, B. (2001). "Review: Protein secondary structure prediction continues to rise." Journal of Structural Biology 134(2-3): 204-218.

secondary structure, in particular alpha helices, and then evaluate 4 different proteins with known high-resolution structures that represent different classes of proteins. Alpha helices are chosen as a focus because prediction of alpha helices in membrane proteins is the most advanced and there are several different methods available. There are not nearly as many developed methods that can predict the beta barrels of membrane proteins, but there is at least one Hidden Markov Model based method that was published recently.[5]

It is obvious from work earlier in this class and experimentation in general, that it will be difficult to draw concrete information on the quality of these methods based on only 4 proteins. Other groups that have evaluated the accuracy of transmembrane helix (TMH) prediction methods have gone out of their way to demonstrate the difficulty of evaluating these methods with sets even as large as 36 high-resolution alpha-helical membrane proteins.[6] In spite of this, an evaluation of the accuracy of several different prediction methods is important and will be attempted.

Overview of Methods:

Membrane proteins have several characteristics that allow prediction of transmembrane structures. The first being physical properties such as having regions of hydrophobic residues that are the transmembrane segments, followed by hydrophilic residues that are either periplasmic or cytoplasmic flanks to the transmembrane segments. Effectively, this pattern forms series of alternating hydrophobic and hydrophilic residues. Another physical property that is characteristic of membrane proteins is something called the "positive-inside-rule" developed by von Heijne in 1986.[7] The "positive-inside-rule" is a phenomenon characteristic of transmembrane proteins where positively charged amino acid residues like lysine and arginine have a higher propensity to be in the cytoplasmic flanking segments of a membrane protein's sequence. The plausibility of such a phenomenon is supported by the fact that because a cell is essentially more negatively charged than it's environment due to the proton motive force that exists about a cellular membrane, the movement of a positively charged amino acid residue across the lipid bilayer would be more difficult than for a negatively charged amino acid residue.

Alpha helices could simply be predicted based on a measure of hydrophobicity using Kyte-Doolittle hydropathy scales.[8] Identification of an alpha helix was done by simply computing the degree of hydrophobicity for a segment of about 19 amino acid residues, approximately corresponding to the number of residues needed to span a lipid bilayer, and discriminating helices from other structures with a minimum threshold hydrophobicity value.

---

5 Martelli, P. L., P. Fariselli, et al. (2002). "A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins." Bioinformatics 18 Suppl 1: S46-53.
6 Chen, C. P., A. Kernytsky, et al. (2002). "Transmembrane helix predictions revisited." Protein Science 11: 2774-2791.
7 Von Heijne, G. (1986). "The distribution of positively charged residues." Nature 34: 456-458.
8 Kyte J., R. F. Doolittle. (1982). "A simple method for displaying the hydropathic character of a protein." J Mol Biol 157:105-132.

Table 1Kyte-Doolittle Hydrophobicity Values for 20 Amino Acid Residues.

A	1.8	G	-0.4	M	1.9	S	-0.8
C	2.5	H	-3.2	N	-3.5	T	-0.7
D	-3.5	I	4.5	P	-1.6	V	4.2
E	-3.5	K	-3.9	Q	-3.5	W	-0.9
F	2.8	L	3.8	R	-4.5	Y	-1.3

A method was then needed to distinguish between membrane alpha helices and globular alpha helices and this was accomplished with a measure of alpha helix amphiphilicity developed by Eisenberg.[9] Amphiphilicity is a measure of the asymmetry of the hydrophobicity as you go around the alpha helix with an end on view. Alpha helices in globular proteins have one side that is predominantly hydrophobic and faces the hydrophobic core of the protein and one side that is predominantly hydrophilic and faces the aqueous solvent. Membrane alpha helices, on the other hand, do not have this amphiphilicity because the protein is in a hydrophobic environment.

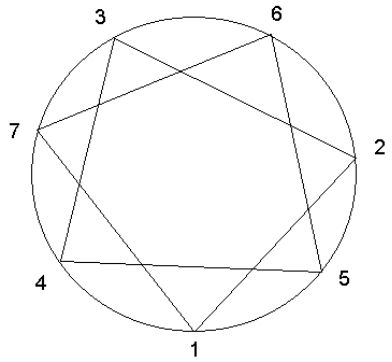

Figure 1 End on view of an alpha helix.

The last major tool that greatly enhanced ability to predict alpha helices in membrane proteins was the use of known membrane alpha helices. Simple statistical preference was used to calculate the propensity of a given amino acid to be in a membrane alpha helix and as expected, methods utilizing statistical preference in addition to physical property scales predicted membrane protein structures more accurately. In current models, known membrane alpha helices are also used to perform homology alignments and then used as part of training sets for more complicated models like neural networks and hidden Markov models.

[9] Eisenberg D., R. M. Weiss, et al. (1982). "The helical hydrophobic moment: a measure of the amphiphilicity of a helix." Nature 299: 371-374.

Review of Methods Used For Case Study:

HMMtop 2.0

This hidden Markov method used to predict transmembrane helices and membrane topology was first published by Gábor Tusnády in 1998.[10] The program was updated in 2001 to allow the user to specify segment localization to increase accuracy. This case study did not take advantage of that user input.

The basis of HMMtop is the hypothesis that different segments of a membrane protein, those being membrane helix, outside transmembrane helix cap, inside non-membrane region, inside transmembrane helix cap, and outside non-membrane region, have differing amino acid compositions. These regions are not found by specifying an ideal amino acid composition for each segment; however, they are found by specifying an ideal difference in amino acid composition between segments.

SOSUI 1.0

SOSUI translates as "being hydrophobic" in Japanese. Accordingly, T. Hirokawa developed the method in Japan in 1996 and published in 1998.[11] The basis of SOSUI is simply a combination of the Kyte-Doolittle hydropathy scale and an amphipathy index proposed by Hirokawa.

SPLIT 4.0

SPLIT is a simple method based on Kyte-Doolittle hydropathy and was developed by Juretic.[12] The method utilizes a non-linear average of hydrophobicity over a segment that was trained on known transmembrane helices.

TMHMM-2.0

TMHMM is a hidden Markov model method and was developed by Sonnhammer, von Heijne, and Krogh.[13] The method is much like HMMtop, however it uses seven states instead of five. Those include the transmembrane helix core, N and C terminal TMH-caps, long and short cytoplasmic non-membrane regions, and a globular domain state for the middle of each non-membrane region on the periplasmic side.

---

[10] Tusnády, G. E., I. Simon. (1998). "Principles governing amino acid composition of integral membrane proteins: applications to topology prediction." J Mol Biol 283: 489-506.
[11] Hirokawa T., BC. Seah, et al. (1998) "SOSUI: Classification and secondary structure prediction system for membrane proteins." Bioinformatics 14: 378-379.
12 Juretic, D., D. Zucic, et al. (1998). "Preference functions for prediction of membrane-buried helices in integral membrane proteins." Comput Chem 22(4): 279-294.
13 Sonnhammer, E. L., G. von Heijne, et al. (1998). "A hidden Markov model for predicting transmembrane helices in protein sequences." Proc Int Conf Intell Syst Mol Biol 6: 175-182.

MPEx 2.04

Wimley and White developed the Membrane Protein Explorer to predict transmembrane helices and topology via hydropathy plots.[14] This method is based on a hydropathy scale developed by Wimley and White.

TMpred

TMpred predicts transmembrane regions and orientation and was developed by Hofmann and Stoffel in 1993.[15] While there is not very much documentation on this method, it finds TMH segments by combining several weight scoring matrixes that were trained on TMbase, expert compiled database of TMH segments.

TopPred 2

TopPred utilizes a sliding trapezoidal window, emphasizes the positive-inside-rule, and evaluates segments based on a hydropathy scale developed by Engelman.[16] Gunnar von Heijne developed the method in 1992.[17]

MEMSAT

MEMSAT utilizes dynamic programming to maximize expectation and was developed by Jones.[18] The method assigns one of five categories to each residue, those being inside loop, outside loop, inside helix end, helix middle, and outside helix end. The dynamic programming then searches through many possible predictions and maximizes the expectation score, based on experimentally derived data.

TM-Finder

TM-Finder was developed by Liu and Deber in 1999.[19] The method is based on experimental data that measured the propensity of individual amino acids to be in an alpha helical state with circular dichroism and their hydrophobic character with a HPLC.

---

[14] White, S. H., Wimley W. C. (1999). "Membrane protein folding and stability: Physical principles." Annu Rev Biophys Biomol Struct 28:319-365.
[15] Hofmann K., W. Stoffel. (1993). "TMBASE – a database of membrane spanning protein segments." Biol Chem Hoppe-Seyler 374: 166.
[16] Engelman D. M., T. A. Steitz, et al. (1986). "Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins." Annu Rev Biophys Biophys Chem 15: 321-353.
[17] von Heijne, G. (1992). "Membrane protein structure prediction. J Mol Biol 225: 487-494.
[18] Dones D. T., W. R. Taylor, et al. (1994). "A model recognition approach to the prediction of all-helical membrane protein structure ad topology." Biochem 33: 3038-3049.
[19] Liu, L. –P. and C. M. Deber. (1999). "Combining hydrophobicity and helicity: A novel approach to membrane protein structure prediction. Bioorg & Med Chem 7: 1-7.

PHDhtm and PHDRhtm

PHDhtm utilizes multiple alignments from protein families and a neural network to identify transmembrane helices and was developed by Rost.[20] PHDRhtm is an updated PHDhtm that utilizes dynamic programming-like algorithm to refine the PHDhtm results that normally result in transmembrane helices that are too long.

PHDhtm uses a three layer computational method. The first inputs a window of 13 amino acids and a global alignment to a second layer that is comprised of several different neural networks that were trained on different sets of data. Some sets are balanced, meaning that the overall set has a representative number of transmembrane helices, loops, and other structures. Some sets are unbalanced, meaning they have underrepresented types of secondary structures. The third layer then reports the final prediction that has the highest score.

Case Studies:

Experimental Method:

Eleven different methods were used to evaluate the accuracy of four different proteins.

Potassium Ion Channel: KcsA_Strco from *Streptomyces coelicolor*, a 160 AA membrane protein with two alpha helices.

Iron(III) dicitrate transport protein: FecA_Ecoli from *Escherichia coli*, a 774 AA beta barrel membrane protein with three short alpha helices.

Sensory rhodopsin II: Bact_Natph from *Natronomonas pharaonis*, a 239 AA membrane protein with seven alpha helices.

Aspartate transcarbamylase: PyrB_Ecoli from *Escherichia coli,* a 310 AA globular protein.

The last three proteins have crystal structures that could not have been used for training the prediction methods because they were only released in the past two months. PyrB and Natph have structural homologues that have solved crystal structures so that may affect the accuracy of their prediction, favoring methods that train on known transmembrane helices. KcsA is quite an old protein and has had its structure solved since 1998. Luckily, the top homologues matches of these proteins were released in 1999 onward and the majority of the methods were published around 1999. This search was done at http://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver. In addition, most methods were developed around 1999.

[20] Rost, B. (1996). "PHD: predicting 1D protein structure by profile based neural networks." Meth in Enzym 266: 525-539.

The goal of choosing the previous four proteins was to investigate evaluate the methods' abilities to predict alpha helices in a proteins that are predominantly transmembrane alpha helices (KcsA and Natph), proteins that are membrane beta barrels and a few alpha helices (FecA), and a globular protein (PyrB).

Results:

All methods were used with default settings. In cases where two results were given, representing a conservative prediction and a non-conservative prediction, the conservative prediction was used.

Graphical views of all predictions have been generated to give an overall impression of the prediction methods. It is difficult to evaluate which method is the best, however for Natph, it appears to be SOSUI and for KcsA, it appears to be MPEx. All of the predictions for FecA miss all the transmembrane helices. Some predictions for PyrB are able to get 100% correct, however MPEx, TMpred, and TM-Finder incorrectly identify transmembrane helices in the globular protein. You can also see that some of the misses correspond to alpha helices in the globular protein.

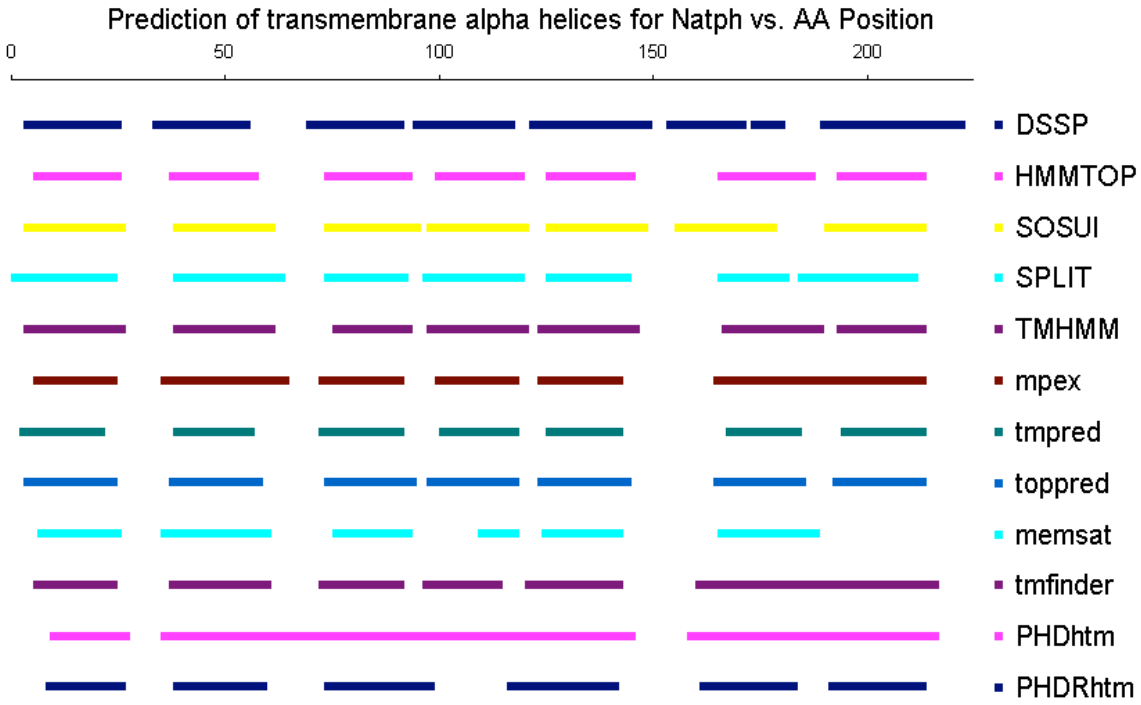

Figure 2 Prediction of transmembrane alpha helices for Natph.

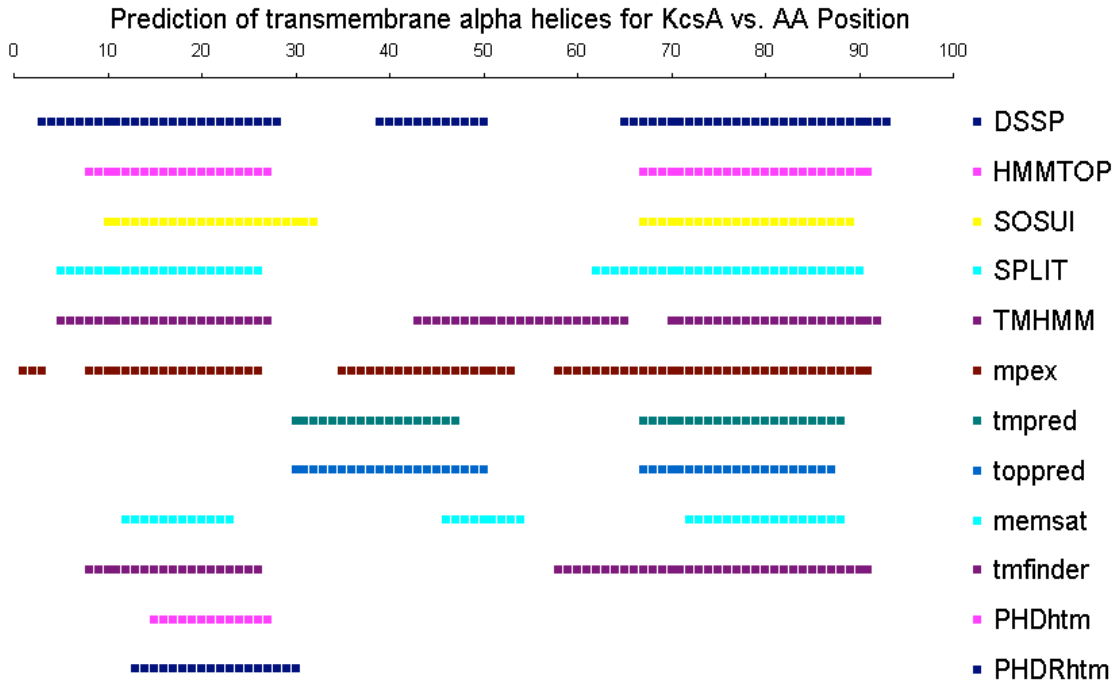

Figure 3 Prediction of transmembrane helices for KcsA.

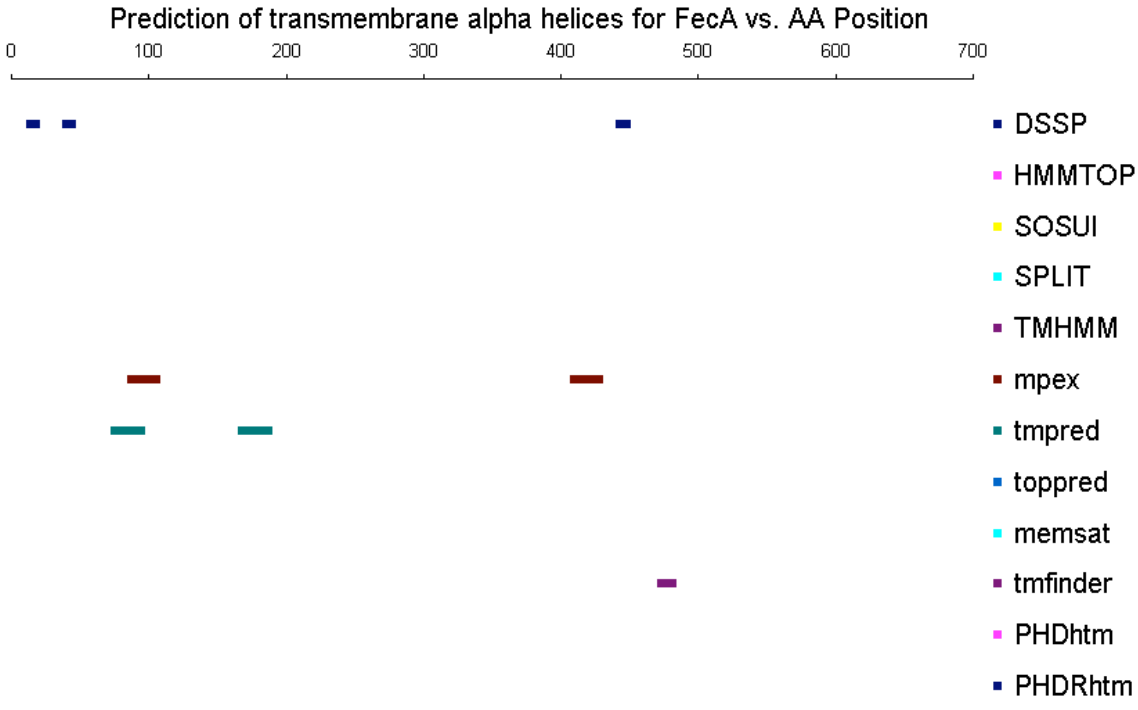

Figure 4 Prediction of transmembrane helices for FecA.

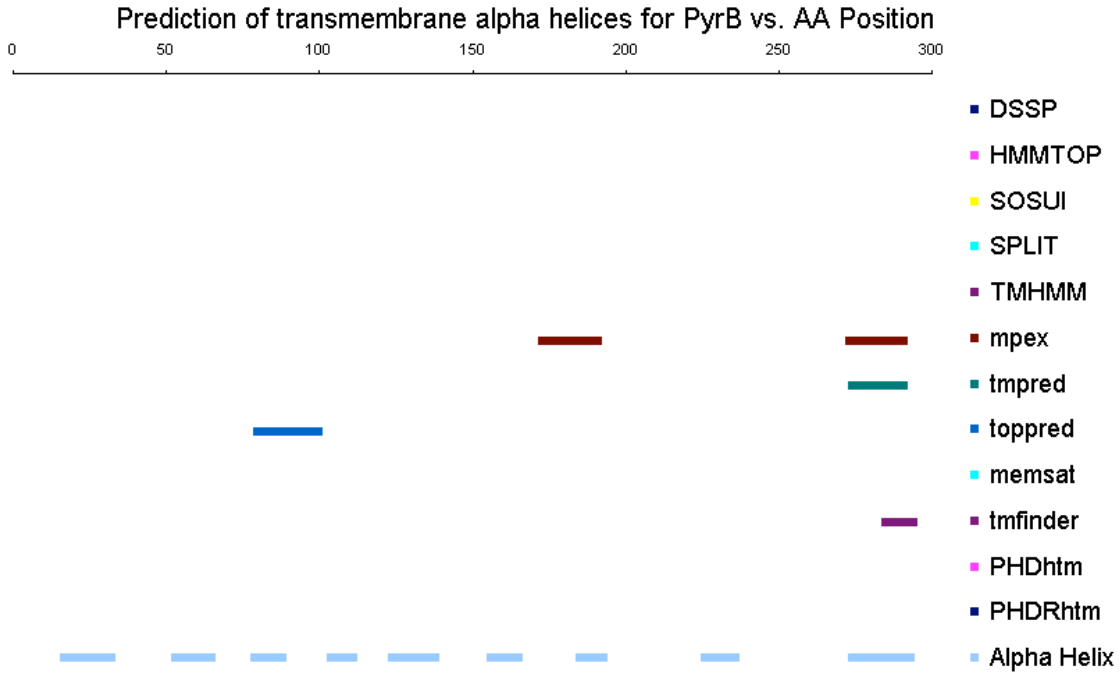

Figure 5 Prediction of transmembrane helices for PyrB and one row for globular helices.

A more quantitative way of evaluating the methods is with measures of Matthew's correlation coefficient, sensitivity, and precision.

Matthew's correlation coefficient (MCC) is a measure that is not swayed by the actual percentage of true positives in a sample and is the most accurate way of evaluating different methods. Sensitivity is a measure of how many observed positive occurrences are actually predicted. Specificity is a measure of how many observed negative occurrences are actually predicted. Precision is a measure of how many predicted positives are observed positives.

(1.1)

(1.2)

(1.3)

(1.4)

Where TP is true positive, FP is false positive, FN is false negative, and FP is false positive. All observations were made on a per residue basis in comparison to the Database of Secondary Structure in Proteins (DSSP) server, which was the gold standard for each protein.

Table 2 Quantitative results for eleven different prediction methods.

NATPH	HMMtop 2.0	SOSUI 1.0	SPLIT 4.0	TMHMM-2.0	MPEX 2.04	TMPRED	TopPred 2	MEMSAT	TM-Finder	PHDhtm	PHDRhtm
MCC	0.356	**0.475**	0.266	0.243	0.318	0.424	0.409	0.157	**0.430**	0.201	0.284
Sensitivity	0.724	**0.828**	0.747	0.741	0.764	0.678	0.759	0.540	0.805	**0.868**	0.667
Specificity	0.686	0.686	0.549	0.529	0.588	**0.824**	**0.706**	0.647	0.667	0.314	0.667
Precision	0.887	**0.900**	0.850	0.843	0.864	**0.929**	0.898	0.839	0.892	0.812	0.872
KCSA	HMMtop 2.0	SOSUI 1.0	SPLIT 4.0	TMHMM-2.0	MPEX 2.04	TMPRED	TopPred 2	MEMSAT	TM-Finder	PHDhtm	PHDRhtm
MCC	**0.627**	0.464	**0.577**	0.376	0.400	0.163	0.190	0.361	0.430	0.266	0.209
Sensitivity	0.672	0.627	0.716	**0.821**	**0.881**	0.463	0.493	0.507	0.687	0.194	0.239
Specificity	**1.000**	0.871	0.903	0.548	0.484	0.710	0.710	0.871	0.774	**1.000**	0.935
Precision	**1.000**	0.913	0.941	0.797	0.787	0.775	0.786	0.895	0.868	**1.000**	0.889
FECA	HMMtop 2.0	SOSUI 1.0	SPLIT 4.0	TMHMM-2.0	MPEX 2.04	TMPRED	TopPred 2	MEMSAT	TM-Finder	PHDhtm	PHDRhtm
MCC	n/a	n/a	n/a	n/a	**-0.039**	**-0.040**	n/a	n/a	-0.019	n/a	n/a
Sensitivity	0.000	n/a	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Specificity	1.000	n/a	1.000	1.000	**0.941**	**0.938**	1.000	1.000	0.986	1.000	1.000
Precision	n/a	n/a	n/a	n/a	0.000	0.000	n/a	n/a	n/a	n/a	n/a
PYRB	HMMtop 2.0	SOSUI 1.0	SPLIT 4.0	TMHMM-2.0	MPEX 2.04	TMPRED	TopPred 2	MEMSAT	TM-Finder	PHDhtm	PHDRhtm
MCC	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Sensitivity	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Specificity	1.000	1.000	1.000	1.000	**0.872**	**0.939**	0.929	1.000	0.966	1.000	1.000
Precision	n/a	n/a	n/a	n/a	0.000	0.000	0.000	n/a	0.000	n/a	n/a

This table high lights the top two values in green for KcsA and Natph and the bottom two values in orange for FecA and PyrB.

Table 3 Quantitative results for eleven predictions for all proteins.

ALL	HMMtop 2.0	SOSUI 1.0	SPLIT 4.0	TMHMM-2.0	MPEX 2.04	TMPRED	TopPred 2	MEMSAT	TM-Finder	PHDhtm	PHDRhtm
MCC	0.737	0.768	0.730	0.718	0.599	0.532	0.647	0.594	0.712	0.668	0.615
Sensitivity	0.665	0.724	0.693	0.716	0.747	0.580	0.642	0.498	0.724	0.638	0.514
Specificity	0.984	0.980	0.975	0.963	0.890	0.926	0.956	0.978	0.958	0.966	0.981
Precision	0.914	0.903	0.873	0.829	0.630	0.662	0.786	0.853	0.812	0.824	0.874

Predictions for all residues were lumped together to determine an overall best prediction method.

Table 4 Mean and standard deviation of the evaluation parameters for the prediction methods for the KcsA and Natph proteins.

Mean	HMMtop 2.0	SOSUI 1.0	SPLIT 4.0	TMHMM-2.0	MPEX 2.04	TMPRED	TopPred 2	MEMSAT	TM-Finder	PHDhtm	PHDRhtm
MCC	**0.492**	0.470	0.422	0.310	0.359	0.294	0.300	0.259	0.430	0.234	0.247
Sensitivity	0.698	0.728	0.732	0.781	0.823	0.571	0.626	0.524	0.746	0.531	0.453
Specificity	0.843	0.779	0.726	0.539	0.536	0.767	0.708	0.759	0.721	0.657	0.801
Precision	0.944	0.907	0.896	0.820	0.826	0.852	0.842	0.867	0.880	0.906	0.881
Stand. Dev.	HMMtop 2.0	SOSUI 1.0	SPLIT 4.0	TMHMM-2.0	MPEX 2.04	TMPRED	TopPred 2	MEMSAT	TM-Finder	PHDhtm	PHDRhtm
MCC	0.192	0.008	0.220	0.094	0.058	0.185	0.155	0.144	0.000	0.046	0.053
Sensitivity	0.037	0.142	0.022	0.057	0.083	0.152	0.188	0.023	0.083	0.477	0.303
Specificity	0.222	0.131	0.250	0.013	0.074	0.081	0.003	0.158	0.076	0.485	0.190
Precision	0.080	0.009	0.064	0.033	0.054	0.109	0.079	0.040	0.017	0.133	0.012

In this table, the KcsA and Natph cases were taken as random independent samples and a mean value is reported in the table without any weighting.

Discussion:

The results of my own visual guesses of the best prediction method matched the quantitative results for the Natph protein, but did not match for the KcsA protein. Although it is nice to look at these pictures to get an understanding of what the methods are predicting, nothing quantitative can be drawn from visual representation and so it is a poor method for evaluating the prediction methods.

The quantitative measurements of MCC, sensitivity, specificity, and precision allow for a more thorough evaluation of the methods. Clearly, the best prediction method for Natph is SOSUI and the best prediction method for KcsA is HMMtop. These two methods have the highest evaluation values in the most number of categories for the two proteins.

Surprisingly, these two methods are at the opposite ends of the spectrum in terms of the manner in which they predict transmembrane helices. HMMtop is a hidden Markov model and SOSUI is a hydropathy – amphiphilicity model.

Notably, MPEx, TMpred, and TM-finder did the worst with FecA because they predicted false positives. Although, none of methods were able to positively identify the few transmembrane helices present in FecA. These same methods also did the worst for PyrB. MPEx, TopPred, TM-Finder, and TMpred incorrectly predicted transmembrane helices in the globular protein PyrB. Some of these methods confused globular helices with transmembrane helices, which can be observed at the C-terminus of the PyrB protein sequence. Fortunately, these errors were not extreme and all programs did fairly well.

Overall, the two most accurate methods are HMMtop and SOSUI based on the compilation of residues results. These two methods have the highest evaluation values for accurately predicting the combination of all residues from all four proteins. This evaluation method may be faulted by weighting the results of some proteins with more

residues, like FecA, however it retains KcsA's best predictor SOSUI, and Natph's best predictor HMMtop.

A reasoning for why SOSUI and HMMtop are the best and the others do not perform as well is not as clear as one would hope. SOSUI is a physical property method like TM-Finder, TopPred, MPEx, and SPLIT. HMMtop is a trained method like PHDhtm, PHDRhtm, MEMSAT, TMbase, and TMHMM. One could argue that less accurate physical property methods did not account for the physical properties in an accurate way, and that the trained methods were either trained incorrectly, or even over trained and so are only good at predicting proteins that they were trained on. On the other hand, it can easily be seen that the two best methods have standard deviations that imply that this analysis cannot significantly distinguish SOSUI and HMMtop from the other methods.

When looking at the averages for the evaluation values, it appears that HMMtop is slightly better. This would be my expected result because from previous examples in class where we used hidden Markov model programs to align sequences, hidden Markov models are able to train for things that we don't understand and so are sometimes better than simple techniques. However, an argument can be made that predicting transmembrane alpha helices is not actually that difficult because of most of the physical constraints, those being a hydrophobic environment, a limited segment length, and amphiphilicity are understood and can easily be applied. As mentioned previously, it is difficult to draw a conclusion that one method is better than another from this analysis and so it seems that the complicated models, in fact, are just as good as the simple models.

Recent work by an independent group has claimed that TMHMM is the best prediction method.[21] Interestingly, my results rarely have TMHMM as the best prediction method in any category. Perhaps an argument can be made again that I did not use unique membrane proteins, most of the programs were developed even before the structures of the homologues were released. For example, SOSUI was developed in 1996 and HMMtop in 1998, so no homologues existed to be trained on at that time.

In addition, the refinement of PHDhtm to PHDRhtm does not make any significant difference in terms of quantitative accuracy; however, the visual representation of the two methods does make it look like the refined method would be more accurate.

A final comment that can be made concerns the usability of the different methods. All of these methods were quite easy to use and provided prompt results, except for PHDhtm and PHDRhtm. Those two methods took upwards of 12 hours to get results!

[21] Möller, S., M. D. R. Croning, et al. (2001). "Evaluation of methods for the prediction of membrane spanning regions." Bioinformatics 17(7): 646-653.

Conclusions:

The reported accuracies of the different methods used to prediction of membrane alpha helices are quite high and range from 90 to 95% and although it is believed that these values are much higher than what exists in reality, I was able to observe precisions around 98% and specificity around 90%.[22] These high accuracies are possible because the lipid bilayer environment actually simplifies the prediction problem by restricting the allowable lengths of segments and providing an additional parameter governed by von Heijne's "positive-inside-rule." This is a good sign for people who are interested in membrane proteins.

The results of my case studies show that HMMtop and SOSUI are the most accurate methods for the prediction of transmembrane helices, however the standard deviations show that this result may be within the error of the techniques. To improve upon this result, it may be better to have sampled more proteins; however, I believe that even after many samples, these prediction methods would actually give results that are quite similar. Significant gains are not made with more complicated models that incorporate neural networks, dynamic programming, or hidden Markov models because of fairly good understanding of what comprises a transmembrane alpha helix.

[22] Chen, C. P., B. Rost. (2002). "State-of-the-art in membrane protein prediction." Applied Bioinformatics 1: 21-35.