NAIST-IS-M0251065

Master's Thesis

# Analysis of Transmembrane Helices in Proteins: The Difficult-to-predict Helices

Shin Tanimoto

February 6, 2004

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

Master's Thesis
submitted to Graduate School of Information Science
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of Science

Shin Tanimoto

Thesis Committee: Shin Isii, Professor
Toshio Hakoshima, Professor
Shigehiko Kanaya, Associate Professor
Takeshi Kawabata, Adjunct Associate Professor

# Analysis of Transmembrane Helices in Proteins: The Difficult-to-predict Helices*

Shin Tanimoto

## Abstract

Integral membrane proteins play important functional roles in biological systems. In order to understand the mechanism of their functions, knowledge of their 3D structures is essential. However, due to inherent experimental difficulties, the 3D structures of only a limited number of integral membrane proteins are known. In the absence of 3D structure, prediction of transmembrane (TM) helix topology is extremely important. Although the performances of currently available TM helix prediction methods are high, neither can they predict all TM helices nor are they free of over-prediction. A clear understanding of the failures and successes of the current TM helix prediction methods is necessary before better prediction methods can be proposed. In this work we perform an analysis of a known set of TM helices in proteins in terms of their average hydropathy values, amino acid preferences and lipid accessibility. Specifically we focus on TM helices that are correctly predicted and those that are missed by typical TM helix prediction methods. Compared to correctly-predicetd helices, difficult-to-predict (missed by prediction methods) helices are found to be mostly hydrophilic and they show unique amino acid propensity - over-representation of Trp, Tyr and Phe, and, under-representation of Glu, Asp and Lys. However, the correctly-predicted as well as difficult-to-predict helices show no correlation with lipid accessibility. Our results will help understand the limitations of the current TM helix prediction methods and will be useful in proposing new TM helix prediction strategies.

**Keywords:** transmembrane helices, transmembrane helix prediction, amino acid propensity, hydrophobicity, lipid exposure

*Master's Thesis, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT0251065, February 6, 2004.

# タンパク質における膜貫通ヘリックスの解析*

# :予測困難なヘリックスの特徴解析

谷本 心

## 内容梗概

膜内在性タンパク質は生体システム内において重要な機能的役割を果たしており，この機能メカニズムを理解するためには立体構造は欠かすことができない．しかしながら、膜タンパク質固有の実験上の難しさから，立体構造が既知の膜内在性タンパク質は数が限られている．それゆえに膜貫通ヘリックスのトポロジーを予測することは極めて重要である．現在用いられている膜貫通ヘリックス予測手法の精度は高いものであるが，全ての膜貫通ヘリックスを過不足なく予測することはできない．現在の膜貫通ヘリックス予測手法の成功と失敗について明確に理解することは新しい手法を公開する前に必要となる．本稿では既知の膜貫通ヘリックスについて，特に予測しやすいもの，また予測が外れやすいものについて，疎水性指標，脂質露出度，アミノ酸の出現頻度に着目して解析を行う．予測しやすいヘリックスに比べ，予測が外れやすいヘリックスは，概ね親水性であり，アミノ酸の出現頻度が特徴的であることが分かった．すなわち，Trp，Tyr，Pheが多く出現し，Glu，Asp，Lysは出現が少ない．しかしながら，脂質露出度については顕著な差異を確認できなかった．これらの結果は現在の予測手法の限界を理解するために役立ち，新しい予測手法の提案に繋がるものと考えている．

キーワード

膜貫通ヘリックス予測，タンパク質構造予測，バイオインフォマティクス，情報生命工学

*奈良先端科学技術大学院大学情報科学研究科情報生命学専攻修士論文，
NAIST-ISMT0251065, 2004 年2 月9 日.

# Contents

# Chapter 1

# Introduction

## 1.1 The function and structure of proteins are related

Proteins and nucleic acids are the two major classes of molecules that make up biological systems. Nucleic acids, DNA and RNA, carry genetic information that codes for life. Different proteins are synthesized during the lifecycle of an organism according to its genetic content. It is these proteins that execute all the relevant biological functions of an organism.

It is now well established that the three-dimensional (3D) structure of proteins is primarily responsible for its function (Branden and Tooze, 1999). In other words, in order to understand how a protein works, along with its amino acid sequence, knowledge of the 3D structure is required.

Recent initiative in structural genomics projects around the world has made it possible to build a knowledge-base for a large number of 3D structures, available at the protein data bank (PDB; http://www.rcsb.org/pdb/). Analysis of these structures provides important information about the mechanism of protein function.

## 1.2 There are two classes of proteins: integral-membrane and soluble

The smallest unit of a biological system is the cell. A typical picture of a eukaryotic cell is shown in Fig. 1.1. As shown in Fig. 1.1, the cell itself is defined by a boundary (plasma membrane). In addition, the cell contains many compartments, like lysosome, mitochondria, smooth endoplasmic reticulum etc.
The cell is filled with mostly aqueous cytosol. On the other hand, the boundary of the cell and the boundaries of many internal compartments are defined by a more non-polar environment.

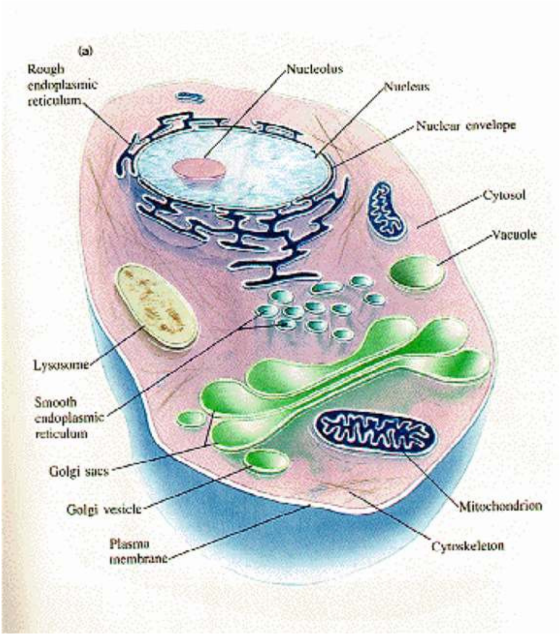

Fig. 1.1 A cartoon of an eukaryotic cell

This non-polar boundary is typically made up of lipid bilayers. A lipid molecule is a long hydrocarbon (non-polar) with a polar head. The non-polar tails can associate in aqueous environments forming bilayers, as shown in Fig. 1.2.

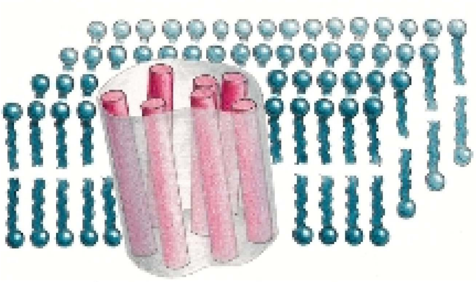

Fig. 1.2 Cartoon representation of a membrane formed by lipid bilayer. The polar head of each lipid molecule is shown by a sphere and the tail is hydrophobic. A protein (red cylinders) is shown embedded in the membrane.

Proteins in a cell are distributed in the cytosol or are embedded in the membrane. Because the membrane environment is very hydrophobic, proteins that are found embedded in the membrane are not water-soluble. From this simple viewpoint, proteins can be divided as integral-membrane (IM) or soluble.

## 1.3 IM proteins play an important role in biological systems

Integral membrane proteins play a crucial role in several key functions in biological systems, like transport of ions and other small molecules across the membrane, cell-cell signaling, receptors for external signals to the cell and as components for energy transduction. The membrane bound receptors are of particular interest since they are important targets for drugs and are therefore important to the pharmaceutical industry (Marchese, 1999). It is therefore very important that the structures of membrane proteins are known.

## 1.4 Experimental determination of 3D structures of IM proteins is difficult

The difference between IM and soluble proteins become most prominent if one surveys the known 3D structures of proteins in PDB. Only about 1% of all structures in the PDB correspond to IM proteins, although the fraction of IM proteins among all proteins is estimated to be between 20-30% (Wallin and von Heijne, 1998). The main reason why so few structures of IM proteins are known is because IM are difficult to manipulate in solution and crystallize, a prerequisite for the determination of 3D structure by X-ray diffraction.

## 1.5 Topology of IM proteins: definition

The inherent difficulty of experimentally determining the structure of IM proteins makes prediction of 3D structure an important field for IM proteins. Fortunately, due to the intrinsic constraint imposed by the membrane, the problem of 3D structure prediction of IM proteins can be reformulated into a 2D problem. As shown in Fig. 1.3, the topology of IM proteins of type a) and b), which consists of one or more transmembrane helices can simply be defined in terms of two criteria: 1) identification of TM helix boundaries in the sequence, 2) identifying whether or not the N-terminal is inside or outside (the cell or the cell compartment that is enclosed by the membrane).

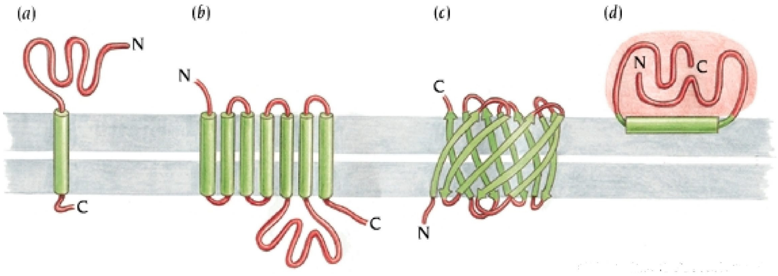

**Fig. 1.3. Four possible orientations of IM proteins in a membrane: a) only a single helix crosses the membrane, b) several helices crosses the membrane, c) a cage formed by beta-sheets is embedded in the membrane, d) the protein is attached to the surface of the membrane.**

## 1.6 Prediction of topology of TM helices in IM proteins

Several prediction methods are available for predicting the topology of TM helices (Chen and Rost, 2002). These methods range from simple amino acid hydropathy-based methods to methods that take advantage of a large dataset of IM proteins with known topologies. In addition the in/out prediction is typically based on the "positive-inside" rule that states that positively charged side chains are biased more towards the inside of the cell (von Heijne, 1986).

Because of the limited dataset with known topologies, it is difficult to accurately estimate the efficiency of the prediction methods. However, the estimates for the prediction accuracy, in terms of fraction of proteins for which all TM helices are predicted correctly, range between 50-70% of all proteins (Chen and Rost, 2002). When tested on soluble proteins, about 10% proteins are falsely predicted as IM proteins. The correct in/out prediction ranges around 60%.

## 1.7 The limitations of TM helix prediction

As mentioned in the review by Chen and Rost (1998), the fraction of all observed TM helices predicted correctly, ~90%, is higher than the correct prediction of all TM helices in IM protein. The per-residue accuracy is lower, at about 80%. As mentioned earlier, these estimates vary as one changes the test set, but overall, it can be said that prediction of TM helices by currently available methods is generally quite successful. The methods can be improved, but probably the prediction accuracy has reached its limit using current strategies. In order to capture the TM helices missed by current prediction methods, it is important that one investigates the nature of TM helices that are difficult to predict. Also it is important that one re-examines the "correct" answers that typical TM prediction methods are supposed to predict.

## 1.8 Hydropathy-based *ab initio* prediction of TM helices: what can be learned

Hydropathy-based *ab initio* prediction of TM helices was the first attempt to predict TM helices (Kyte and Doolittle, 1982). Although it is no more the best prediction method, still the method has some distinct characteristics. Most importantly, since amino acid hydropathy value is a measure of the residue's preference of membrane environment over an aqueous environment, the prediction has some clear physical meaning. Therefore, even if simple hydropathy-based predictions don't match with the observed TM helices, it is important that the actual predictions are carefully analyzed.

## 1.9 Propensity based scores of TM helices

One of the more successful TM prediction methods is based on amino acid propensity values of different regions of TM helices and the loops that join them (MEMSTAT; Jones et al., 1994). Although the success of the method lies more on the way the program reaches its solution dynamically, the amino acid propensities do play an important role. Analysis of amino acid propensities can also give us information about the nature of the helices.

## 1.10 Consensus method of prediction of TM helices

Recently, one strategy of improving the prediction of current prediction methods has been to use a consensus method of prediction (Nilsson et al., 2002). In this method, several prediction methods are used to predict TM segments in a protein. Finally, based on some arbitrary consensus rule, the predicted segments from all the methods are combined to produce the final answer.

## 1.11 Outline of this thesis

The main objective of this thesis is to understand the current methods and strategies of TM helix prediction. Specifically, we want to study the failures and understand any common reason behind such failures. In order to achieve this goal we have performed a detailed analysis of TM helices in proteins.

The thesis is divided into three chapters and the contents of the sections are outlined here. In chapter 2 we sequence hydropathy of TM sequences are analyzed. In chapter 3 amino acid propensities of TM helices are analyzed. Finally in chapter 4 we analyze TM helices from a viewpoint of its 3D structure.

# Chapter 2

# Analysis of Transmembrane Helices: Sequence Hydropathy

## 2.1 Introduction

Over the last two decades, a number of prediction methods have been developed that are capable of predicting TM helices given an input sequence (Chen and Rost, 2002). These methods can broadly be classified into two groups. The first is based on some amino acid hydropathy scale (Kyte and Doolittle, 1982). The second is knowledge based, where some *a priori* knowledge about known TM helixes is used to predict TM helices in new sequences (for example, TMHMM; Sonnhammer et al., 1998).

While the overall accuracy of the best available TM prediction methods is rather high, still there is room for improvement (Chen and Rost, 2002). In order to begin to understand how such improvements can be made, we have to first understand some typical cases where TM prediction fails. In this chapter we attempt to do that by first analyzing known TM helices and then identifying those that are difficult to predict. We want to know if the TM helices that are difficult to predict share some common property. If common properties are found, then it would help in designing prediction methods that perform better than the currently available ones.

As far as the dataset is concerned, in this chapter we only analyze TM helices in proteins whose 3D structures are not known. The TM segments in these proteins were assigned using a variety of experimental methods. This dataset will be called the 2D dataset and is described in the Methods section. In the next chapter we will analyze TM helices in proteins for which 3D structures are known.

We have chosen three top predicting methods, HMMTOP (Tusnady and Simon, 1998), TMHMM (Sonnhammer et al., 1998) and MEMSAT (Jones et al., 1994), to predict TM helices in the 2D dataset. Although the prediction accuracy of these knowledge-based methods is typically higher than simple hydropathy based methods, we also use hydropathy-based prediction for a number of reasons.

Because hydropathy is a measure of the partition of an amino acid between water and a non-polar medium, there is a clear physico-chemical basis for the correctly predicted or incorrectly predicted TM helix. Secondly, although the set of twenty amino acid hydropathy values, originating from some specific scale, is invariant in hydrophobicity-based methods, there are more than one scales available. We use two scales, the KD (Kyte and Doolittle, 1982) and the WW (Wimley and White, 1999) scales, to understand the applicability of the hydropathy values in predicting TM helices.

## 2.2 Materials & Methods

### *2.2.1 Dataset*

I used the TransMembrane Protein DataBase, ( TMPDB, http://bioinfo.si.hirosaki-u.ac.jp/~TMPDB/ ) which is a collection of TM proteins with topologies based on definite experimental evidences such as X-ray crystallography, NMR, gene fusion technique, substituted cysteine accessibility method, Asp( N)-linked glycosylation experiment and other biochemical methods. There are 276 alpha-helical sequences, 17 beta-stranded sequences, and 9 alpha-helical sequences with short pore-forming alpha helices buried in the membrane like potassium channel. In this chapter, I used non-redundant alpha-helical sequences; the non-redundant means that sequences were subjected to a sequence similarity check using CLUSTAL W and the similarity was less than 30%. Also members with X-ray structures were removed to construct the TMPDB 2D database. The final database contained 997 alpha helices.

### *2.2.2 Method*

The hydropathy-scales used in this chapter are that proposed by Kyte and Doolittle (KD scale; Kyte and Doolittle, 1982) and Wimley and White (WW scale; Wimley and White, 1999). The sign of all hydropathy scales used in this chapter is such that negative numbers indicates hydrophobic. In addition, we have 1.0 from the KD values to compensate for the extra energy of burying the peptide backbone in the membrane, as has been suggested (Jayasinghe et al., 2001). A summary outlining the origin of the KD scale is given in the Appendix to this chapter. The two hydropathy values are shown in Fig. 2.1 for all the amino acids. The two scales correlate well, the main difference is for residues Trp and Tyr. There are other differences as shown in Fig. 2.1.

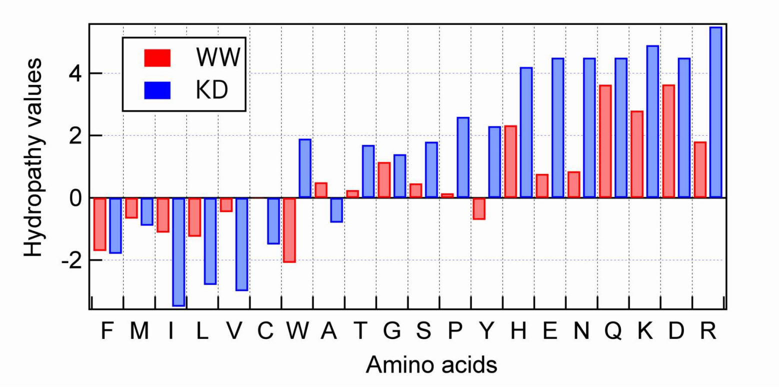

**Fig. 2.1. KD and WW hydropathy values for the 20 amino acid residues. The KD hydropathy values are changed as KD = 1.0 – KD following Jayasinghe et al. (2001)**

The hydropathy values, averaged over some fixed window size correlates quite well with known TM helical segments. This is shown in Fig. 2.2 for the L subunit of photosynthetic reaction center of *R. spheroids.*

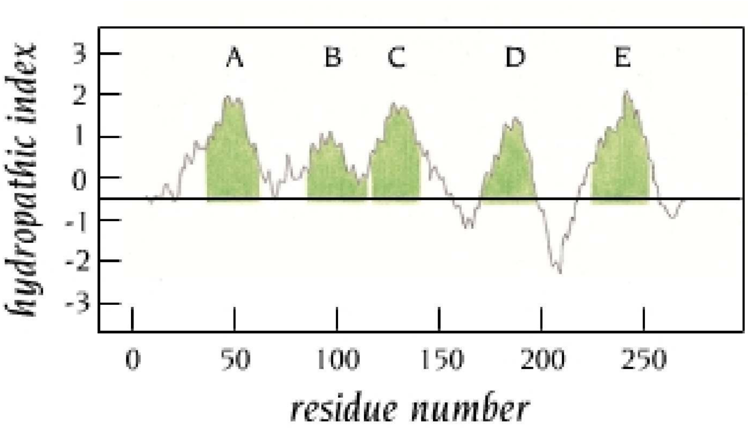

**Fig. 2.2. Hydropathy plot for the L subunit of photosynthetic reaction center of R. sphaeroides. A window size of 19 was used with the KD hydropathy scale.**

This is the basis of predicting TM helices using hydropathy scales. However, as is clear from Fig. 2.2, given the window-averaged hydropathy values, the exact location of the TM segments can only be guessed. Typically some threshold value is used to demarcate the TM regions. In addition some algorithm is needed to precisely define the TM helices.

All hydropathy-based TM prediction method used in this chapter consists of the three following steps:

1. Assign the appropriate hydropathicity values to each residue in a given amino acid sequence.
2. Calculate the average of hydropathicity over some window-size along the sequence and assign the average value to the central residue in the sliding window. The resulting averaged hydropathicity along the sequence is called the hydropathic profile.
3. From the resulting sequence hydropathic profile, predict the TM segments following some strategy, or algorithm.

In step 2, hydropathic profiles were generated using a window size of tried 9, 11, 13, 15, 17, 19, 21 and 23. It should be pointed out that the typical width of lipid bilayers is about 30Å, and assuming ideal $\alpha$-helical geometry (1.5Å rise per residue), 19 residues should lie inside the lipid bilayer, if the helix and the membrane are perpendicular to each other.

In step 3, the simplest algorithm for predicting residues whose score is less than some threshold value. Following Wimley and White we use 0.0 as the threshold for the KD scale. Nevertheless, we vary this threshold value in this work to judge how prediction accuracy changes as a function of different threshold values.

The algorithm used for predicting TM segments follow that of Jayasinghe et al. (2001):

1. Identify all minima in the hydropathy profile
2. Retain only those minima that are below the threshold (0.0).
3. If two minima occur within 9 residues, retain the lower minima.
4. Mark a segment of 19 residues centered on the minima retained in step 3
5. Combine all overlapping 19-residue segments in step 4 into predicted TM segments.

### *2.2.3 Accuracy*

The accuracy of TM helix prediction can be divided into two view points; one is by segment or topology accuracy, and the other is by residue accuracy. Due to two factors, it is said that the former is more important than the latter. One is that the topology of TM proteins strongly affects the function of themselves, and the other is that the edges of the TM segments may be moving and are still vague. Therefore, I judged the accuracy not only by residue but also by segment. When defining a segment to be predicted correctly, many publications did different ways. In this work, the definition of correct segments is as follows: a) the distance between the centers of observed segment and predicted segment is less than or equal 10 residues, and, b) one segment can be counted as correct segment only once.

Prediction accuracy were judged using the following scores:

$Q_{num}$. This score shows the percentage of the sequences for which all the predicted and observed TM segments match correctly. This can be the simplest measure of the segment prediction accuracy.

$$Q_{num} = 100 \times \frac{total \ \ number \ \ of \ \ proteins \ \ correctly \ \ predicted \ (number \ \ of \ \ TM \ \ helices)}{total \ \ number \ \ of \ \ proteins}$$

$Q_{topo}$. This score shows the percentage of the sequences for which the topology was predicted correctly. When both the number and positions of the transmembrane segments are predicted correctly, topology is said to be predicted correctly.

$$Q_{topo} = 100 \times \frac{total \ \ number \ \ of \ \ proteins \ \ correctly \ \ predicted \ (number \ \ and \ \ positions \ \ of \ \ all \ \ TM \ \ helices)}{total \ \ number \ \ of \ \ proteins}$$

$Q_{seg}^{\%obs}$. This score shows the ratio that how the observed segments are predicted correctly. The higher this score is, the less is the under-prediction, regard less of the over-predictions.

$$Q_{seg}^{\%obs} = 100 \times \frac{total \ \ number \ \ of \ \ TM \ \ helices \ \ correctly \ \ predicted \ (positions)}{total \ \ number \ \ of \ \ TM \ \ helices}$$

$Q_{seg}^{\%prd}$. This score shows the ratio that how the predicted segments are predicted correctly. The higher this score is, the less the over-predictions are, regard less of the under-predictions.

$$Q_{seg}^{\%prd} = 100 \times \frac{total\ \ number\ \ of\ \ TM\ \ helices\ \ correctly\ \ predicted\ \ (positions)}{total\ \ number\ \ of\ \ predicted\ \ TM\ \ helices}$$

$Q_2$. This measures the percentage of residues predicted correctly in either of the two states TM (transmembrane segment) or non-TM (not transmembrane segment).

$$Q_2 = 100 \times \frac{total\ \ number\ \ of\ \ residues\ \ correctly\ \ predicted\ \ (TM\ \ or\ \ non-TM)}{total\ \ number\ \ of\ \ residues}$$

Due to the existence of large globular regions, this score tends to dominant by such non-TM regions, and can be high. Therefore, I used following two scores as well.

$Q_{res}^{\%obs}$. This score shows the percentage of the observed transmembrane residues predicted correctly. As same as the $Q_{seg}^{\%obs}$, when this score is high, there is less number of under-predictions. Even if some method predicted all residue of the sequence as transmembrane segment, this score is to be 100.

$$Q_{res}^{\%obs} = 100 \times \frac{number\ \ of\ \ residues\ \ correctly\ \ predicted\ \ as\ \ TM}{number\ \ of\ \ residues\ \ observed\ \ as\ \ TM}$$

$Q_{res}^{\%prd}$. This score shows the percentage of the predicted transmembrane residues were correct prediction. As same as $Q_{seg}^{\%prd}$, when this score is high, there is less number or over-predictions.

$$Q_{res}^{\%prd} = 100 \times \frac{number\ \ of\ \ residues\ \ correctly\ \ predicted\ \ as\ \ TM}{number\ \ of\ \ residues\ \ predicted\ \ as\ \ TM}$$

## 2.3 Results & Discussion

### *2.3.1 TM helix prediction of TMPDB-2D dataset*

First we used a set of standard prediction programs for predicting TM helices in the TMPDB-2D dataset. The prediction programs used were: HMMTOP (Tusnady and Simon, 1998), TMHMM (Sonnhammer et al., 1998) and MEMSAT (Jones et al., 1994). In addition, we also used a simple KD prediction for TM prediction. The results are shown in Table 2.1.

**Table 2.1. Prediction accuracy of TMPDB-2D dataset.**

| METHOD | $Q_{seg}^{\%obs}$ | $Q_{seg}^{\%prd}$ | $Q_{num}$ | $Q_{topo}$ | $Q_2$ | $Q_{res}^{\%obs}$ | $Q_{res}^{\%prd}$ |
|---|---|---|---|---|---|---|---|
| HMMTOP 2.0 | 90.1 | 88.2 | 66.9 | 58.6 | 98.5 | 80.8 | 84.5 |
| KD (21-res) | 58 | 66.9 | 22.1 | 17.7 | 94.8 | 75.6 | 45.2 |
| MEMSAT 2 | 84.9 | 89.4 | 61.3 | 55.8 | 98.2 | 73.9 | 84.3 |
| TMHMM 2.0 | 85.9 | 92.3 | 56.9 | 51.4 | 98.2 | 78.3 | 80.6 |

As is clear from the table, simple hydropathy-based methods show much poorer performance when compared to the three best methods. We tried to improve the KD-prediction (data not shown) by changing the various parameters associated with the prediction algorithm but did not achieve any substantial improvement. The other three methods show quite impressive performance, especially on a per-segment basis. However, as can be seen, the $Q_{num}$ (the number of proteins for which all TM helices predicted correctly) or $Q_{topo}$ (the number of proteins for which the total number of TM helices, even when the actual positions are wrong, are predicted correctly) values are still not very high.

### *2.3.2 Length distribution of TMPDB-2D dataset*

Our aim is to understand the success and failures of the prediction methods. In order to do so, we wanted to first characterize the TMPDB-2D dataset. We do it in two ways. First, we show the helix length distribution of all TM helices in the dataset. This is shown in Fig. 2.3. The average length is about 21-residue, slightly longer than the 19-residue corresponding to the number of TM residues in an ideal TM α-helix, normal to the membrane plane. The majority of helices are in the range 18-25 residues.

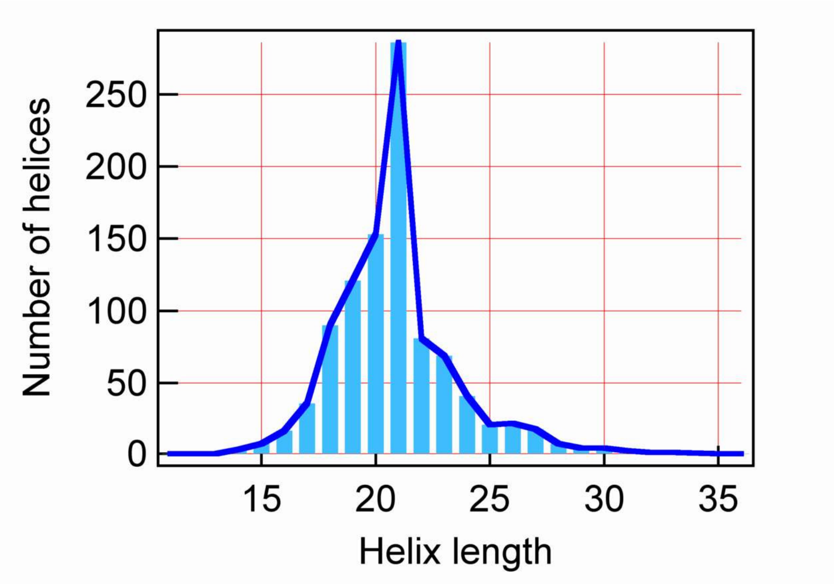

Fig. 2.3. Length distribution of TM helices in TMPDB-2D dataset.

### *2.3.3 Average hydropathy values of helices in the TMPDB-2D dataset*

Then we show how the average KD hydropathies of TM helices are distributed within the TMPDB-2D dataset in Fig. 2.4.

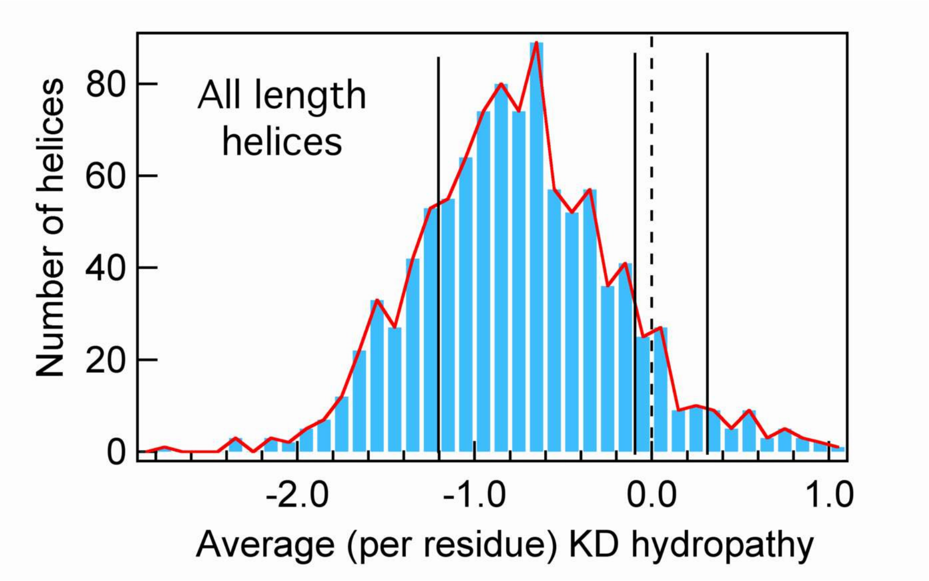

Fig. 2.4. Distribution of average KD hydropathy of TM helices in TMPDB-2D dataset. The average hydropathy value was computed by summing over all the KD hydropathy values in the helix and diving the sum by the total number of residues in the helix.

The average KD hydropathy value of a TM helix represents the overall hydrophobicity of the helix. If the average value is negative, the helix is overall hydrophobic, on the other hand if the average value is positive, the helix can be considered to be overall hydrophilic. Out of a total of 997 helices, 83 show positive values for average KD hydropathy. In other words, about 8.3% of TM helices in the dataset are hydrophilic. This fact clearly indicates that using hydropathy of TM segments as the sole criterion for TM helix prediction, about 10% TM helices will be missed.

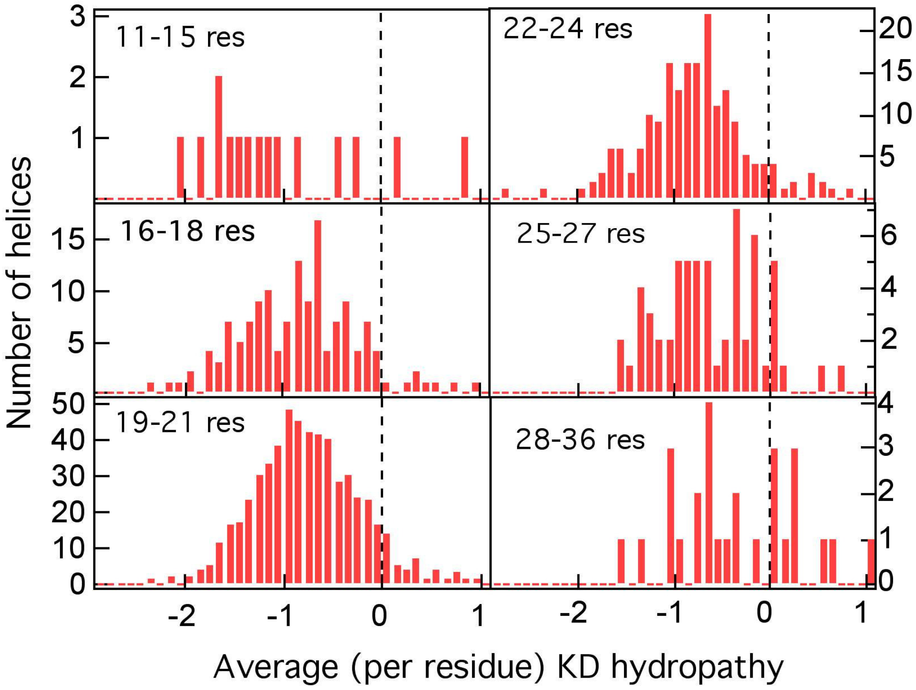

**Fig. 2.5. Distribution of average KD hydropathy of TM helices in TMPDB-2D dataset as a function of helix length.**

Prediction algorithms typically fix a window size when predicting TM segments. The data set in Fig. 2.4 correspond to a variety of helix length (see Fig. 2.3). So we asked a simple question: Is there a correlation between helix length and average hydropathy values. In Fig. 2.5 we show the distribution of average KD hydropathy values as a function of helix lengths. As can be seen from Fig. 2.5, there seems no clear correlation, although the longest helices show a higher percent of hydrophilicity. However, the dataset is too small for any statistical significance.

### *2.3.4 Biases in difficult-to-predict helices in the TMPDB-2D dataset*

Having characterized the dataset in terms of average KD hydropathy, we wanted to investigate if there was any bias in difficult-to-predict helices. In Fig. 2.6 we show the same distribution as in Fig. 2.4., however, four new distributions are now superimposed. These correspond to predicted helices using HMMTOP, TMHMM, MEMSAT and a simple KD hydropathy based method with a window size of 21 residues. We used a 21-residue window size because the average length of TM helices in the TMPDB-2D dataset was close to this value.

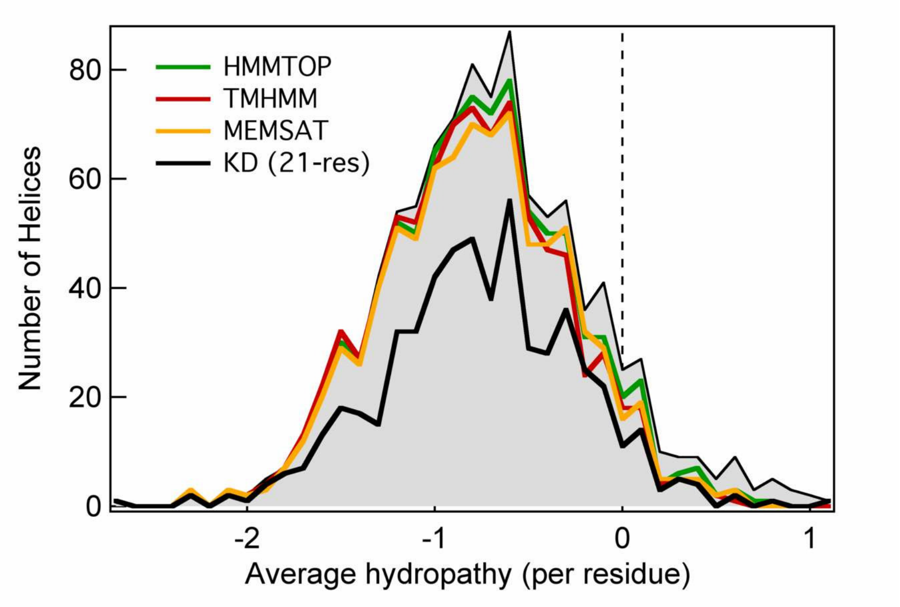

**Fig. 2.6. Distribution of average KD hydropathy of TM helices in TMPDB-2D dataset as a function of prediction by four methods.**

As seen from Fig. 2.6, all three knowledge-based methods, HMMTOP, TMHMM and MEMSAT, perform much better than the *ab initio* hydropathy-based method. Among the best three, it is difficult to decide which one is the best since their overall performance as well as their performance as a function of average helix KD hydropathy is very similar. For more hydrophobic helices, the performance is near to 100%. As the helices become less hydrophobic, and especially, as the helices become hydrophilic, the performances of all methods decline. Clearly, prediction of more hydrophilic helices is most difficult, even by the best methods.

This point is illustrated more clearly in Fig. 2.7 where $Q_{seg}^{\%obs}$ values for two predictions (HMMTOP and KD hydropathy based prediction) are shown as a function of average TM helix hydropathy. For both methods, up to about average hydropathy of -0.1, the $Q_{seg}^{\%obs}$ remains constant (~0.9 for HMMTOP and ~0.6 for KD prediction). However, as the average hydropathy increases beyond this threshold, there is a sharp decrease in $Q_{seg}^{\%obs}$. In other words, the hydrophilic helices are the most difficult to predict.

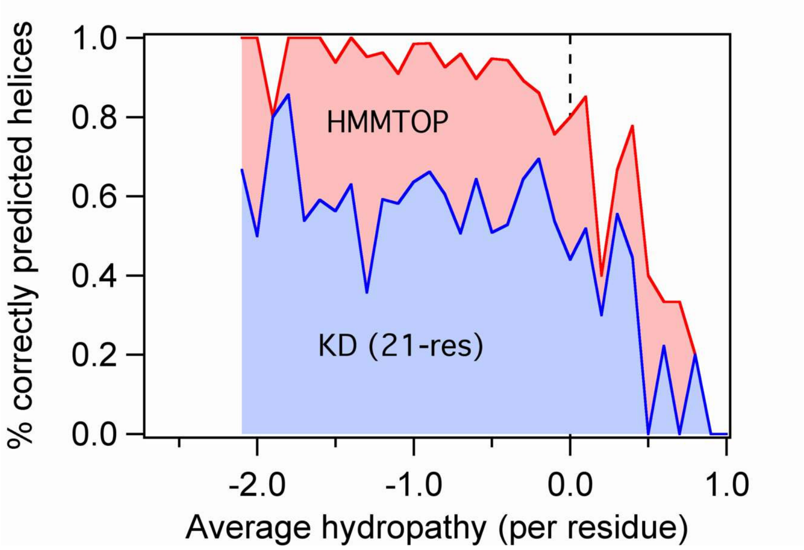

**Fig. 2.7. Fraction of corrected predicted helices ($Q_{seg}^{\%obs}$) in TMPDB-2D dataset by KD method and HMMTOP as a function of average KD hydropathy of TM helices.**

Also, because the performance of KD method, even when helices are hydrophobic, is only moderate (60%); one can conclude that in addition to simple KD hydropathy values, amino acid sequences of TM segments must also exhibit some special characteristics. One way to investigate any such special characteristic is to look at amino acid propensities. In fact, one of the three best performing methods, MEMSAT, uses such propensities. In the following sections we will now discuss amino acid propensities of TM helices and loops that connect them.

# APPENDIX 2.1 Kyte and Doolittle Hydropathy Scale

Kyte and Doolittle (KD) hydrophobicity index is a well known hydrophobicity scale which was first proposed by Kyte and Doolittle (Kyte and Doolittle, 1982). Since we use the KD scale extensively in this work, here the origin and basis of the KD scale is outlined.

A good hydrophobicity scale should reflect the partition of an amino residue between water and a non-polar medium, like the interior of membrane or interior of proteins. Kyte and Doolittle considered three kinds of experimental values of transfer free energy: a) water / condensed vapor, b) water / ethanol, and, c) ethanol / condensed vapor. Because the ideal non-polar medium should not show any specific interactions (like hydrogen bonds) with the amino acid residue, and ethanol is known to show specific interactions, Kyte and Doolittle selectively chose the first set of experimental transfer free energy values.

In addition to the experimental numbers, Kyte and Doolittle referred to the tendency that a given side-chain prefers the interior of a protein or the exterior, which is reflected in tabulation of residue accessibilities as was available then from a limited number of protein structures by Chothia, considering that the average of the actual locations of a side-chain should be a direct evaluation of its hydropathy when it is in a protein. Two sets of values were presented by Chothia, the fraction of the total number of a given residue that is more than 95% buried in the native structures, and the fraction that is 100% buried.

Both, the water-vapor transfer free energies and the interior-exterior distribution of amino acid side-chains (both 95%, and 100%) were finally combined in assigning the final hydropathy values. First they normalize the score, and then they mixed the values subjectively, some were simple averages, and some were respective score, and so on.

# Chapter 3

# Analysis of Transmembrane Helices: Amino Acid Propensities

## 3.1 Introduction

The basis of any successful prediction of TM helices lie in the fact that the amino acid sequences in TM are biased when compared to amino acid sequences of non-TM segments.

One way to quantify this bias is to compile the average hydropathy index of TM helices. As shown before, the TM helices are indeed strongly biased towards being hydrophobic. However, there is no one unique hydropathy value for an amino acid. It is highly dependent on experiments that are performed to measure them. As shown in Fig. 2.1, the widely used KD values and the more recent (and probably more relevant) WW values are not the same.

Another way to quantify the bias is to use amino acid propensity values. Amino acid propensities are obtained from a large dataset of known TM and non-TM residues. Unlike hydropathy scales, the actual values of propensities are more robust. In this chapter we compute and analyze several different kinds of residue propensities associated with TM helices.

## 3.2 Materials & Methods

### *3.2.1 Dataset*

The dataset used is the TMPDB-2D data set, identical to that in chapter 2.

### *3.2.2 Propensities*

Propensities of amino acid residue aa in some region of sequence X (TM helix, loop etc.) $p_{aa}^{X}$ were calculated as:

$$p_{aa}^{X} = \frac{N_{aa}^{X}}{\sum\limits_{aa=1,20} N_{aa}^{X}} \Bigg/ \frac{N_{aa}^{all}}{\sum\limits_{aa=1,20} N_{aa}^{all}}$$

where $N_{aa}^{X}$ is the total number of amino acid aa in a specific region of sequence X and $N_{aa}^{all}$ is the total number of amino acid aa in any region of sequence (the entire dataset).

## 3.3 Results & Discussion

### *3.3.1 Short loop propensities*

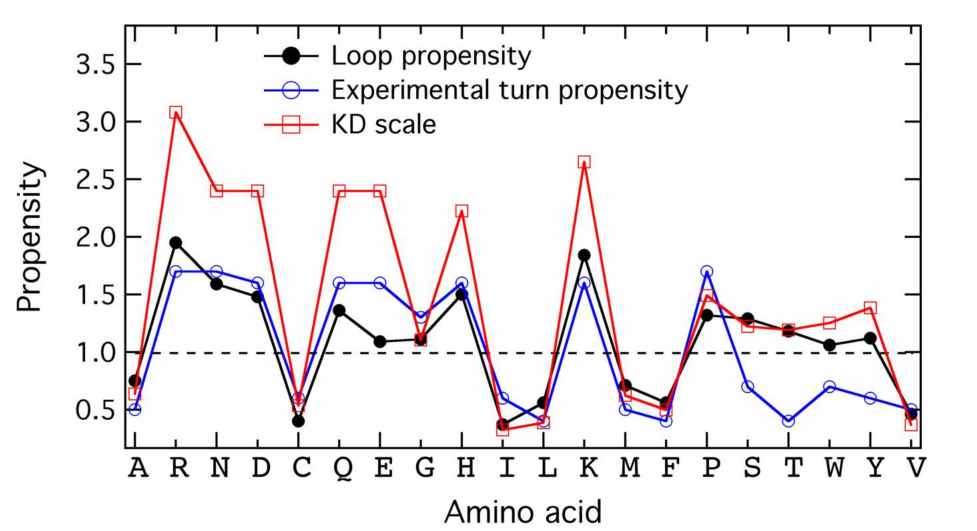

**Fig. 3.1 Loop (1-5 residues) propensities of 20 amino acids (black). The red line indicates ln(KD hydropathy values) and the blue line indicates experimental turn propensity (Monne et al., 1999).**

The first propensity we looked at is the propensity of amino acids in short (1-5 residues) loops joining two TM helices. We were interested in loops since the prediction of short loops was found to be rather difficult when simple KD hydropathy based predictions are performed (data not shown). The loop propensities are shown in Fig. 2.8 along with KD hydropathy values and an experimental turn propensity for single residue turns in between two TM helices (Monne et al., 1999).

The loop propensities correlate very well with the experimental turn propensities, except for Ser, Thr, Trp and Tyr. On the other hand, loop propensities and KD hydropathy values seem to be strongly correlated for all amino acids. In other words, using loop propensity based log odd score, averaged over a small window size (say 5 resdiues) and using KD hydropathy based scores, averaged over a similarly small window size, will yield almost identical results.

### *3.3.2 Propensities of TM helix and flanking regions*

We divided TM helix segments and the flanking regions into five sections as shown in Fig. 3.2.

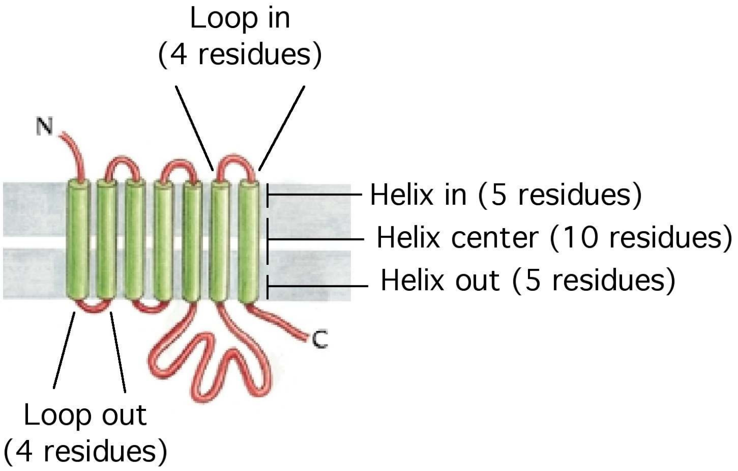

**Fig. 3.2 Definition of helix (in), helix (out), helix (center), loop (in) and loop (out).**

### *3.3.3 Propensities of loop residues (four residues flanking the TM helix)*

In Fig. 3.3 the propensities of loop residues (both in and out, see Fig. 3.2) are shown as a function amino acid residues. In Fig. 3.1 only very short loops connecting two TM helices were considered. Here we consider loops of all length but restrict ourselves to only four contiguous residues from the TM helix. The main difference in loop propensity of Fig. 3.1 and Fig. 3.3 is that residues in 1-5 length loops show considerably lower propensity for Phe, Val, Ile and Leu compared to residues in flanking positions of TM helices. On the other hand, for Arg, Asn, Asp, Gln, His, Lys, Pro and Ser the loop propensities of residues in 1-5 length loops are higher. This tendency can be justified from the fact that the former set is too hydrophobic to be present in short turns outside the membrane. The latter set, on the other hand, is more polar or turn-inducing (like Pro). Experimental turn propensities or hydropathy values (KD and WW) show much less correlation with the propensity values (when compared to Fig. 3.1)

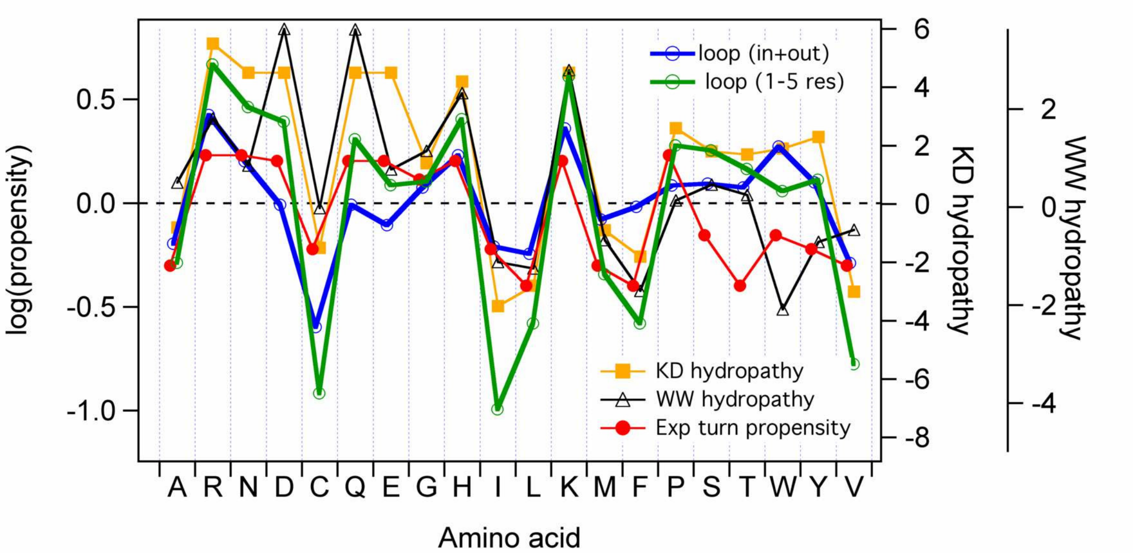

Fig. 3.3 Log loop (4 flanking residues in all length loops) propensities of 20 amino acids (blue). The yellow line indicates KD hydropathy values, the black line indicates the WW hydropathy values and the red line indicates experimental turn propensity (Monne et al., 1999).

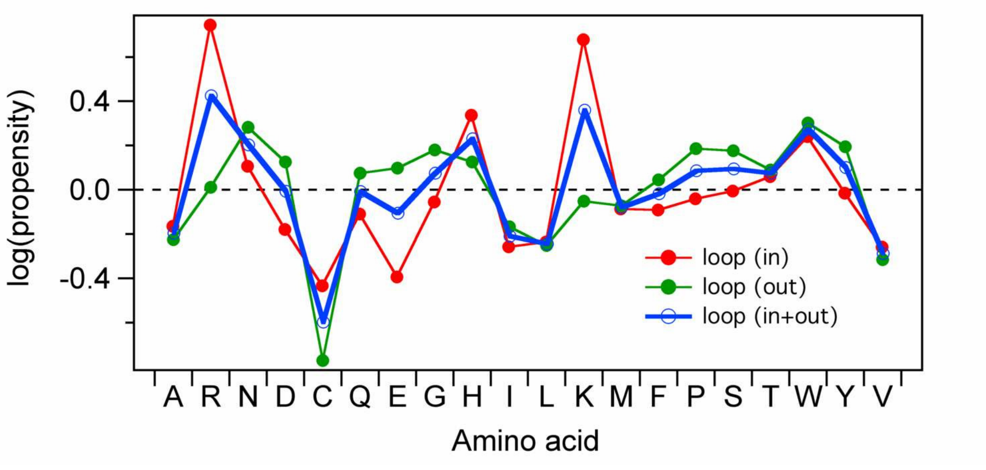

**Fig. 3.4 Log loop (4 flanking residues in all length loops) propensities of 20 amino acids (blue). The red and the green line correspond to in and out loops.**

In Fig. 3.4 we show the difference in loop propensities of in and out segments of four residues flanking TM helices. Positively charged residues (Arg, Lys and His) clearly are more populated in the 'in' region while negatively charged residues (Asp and Glu) are present more towards the 'out' region. This reflects the widely known 'positive-inside' rule of von Heijne (1986).

### *3.3.4 Propensities of residues in the TM helix*

The propensities of amino acids in TM helix are shown in Fig. 3.5. The propensities match quite well with two hydropathy scales, KD and WW. However, the main difference is for Trp and Tyr. According to the KD scale, both Trp and Tyr are disfavored in the helix. In the WW scale, on the other hand, the two residues are favored, especially Trp. According to the propensity values, Tyr is neutral while Trp is slightly favored. Thus, compared to the KD scale, the WW scale tends to reflect the propensities better.

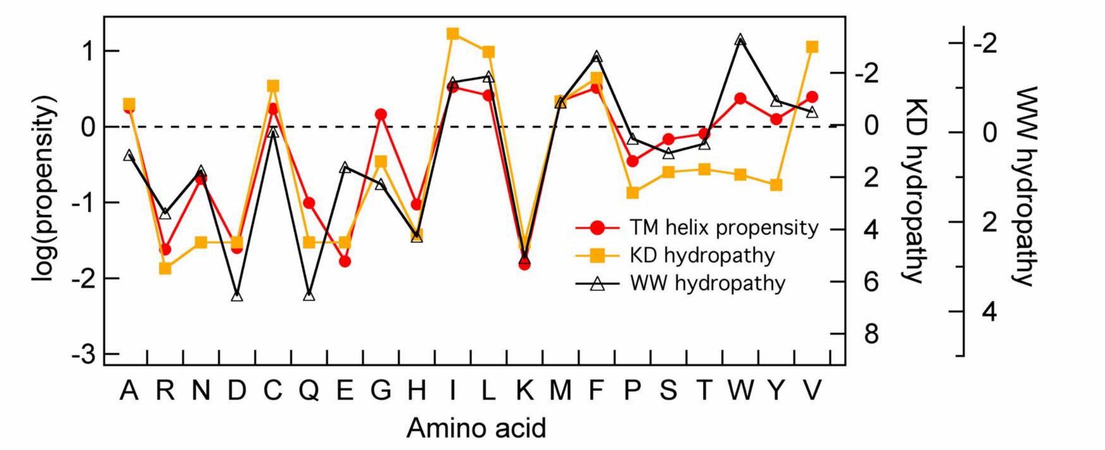

**Fig. 3.5 Log Helix (all residues) propensities of 20 amino acids (red). The yellow and the black lines correspond to KD and WW hydropathy values.**

In Fig. 3.6 the helix propensities are shown separately as helix (in), helix (out) and helix (center). Interestingly, the difference between helix (center) and helix ends become clear for Tyr and Trp, While both residues are under represented in helix (center), they are over represented at the helix ends, matching quite well with the WW scale. For these two amino acids, the KD scale correlates better with the helix (center) propensities. In other words, the WW scale is more suited for helix ends while the KD scale is more suited for helix (center). Apart from Tyr and Trp, charged residues and His are slightly less under represented in helix ends when compared to helix (center).

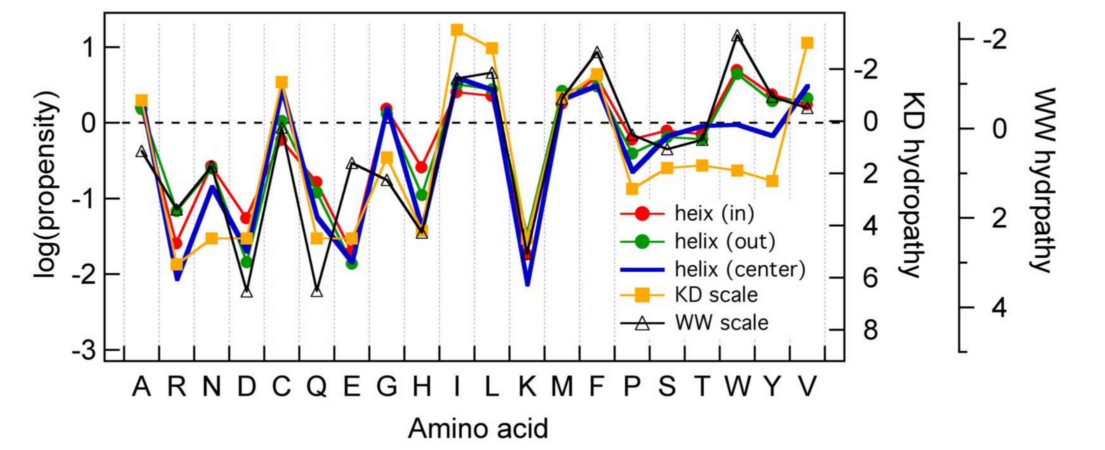

Fig. 3.6 Log Helix (all residues) propensities of 20 amino acids (red). The yellow and the black lines correspond to KD and WW hydropathy values.

### *3.3.5 Propensities of residues in the TM helix as a function of average sequence KD hydropathy*

Having analyzed TM helix propensities for the entire data set, we analyzed the dataset as a function of average KD hydropathy values (see Fig. 2.4), as shown in Fig. 3.7.

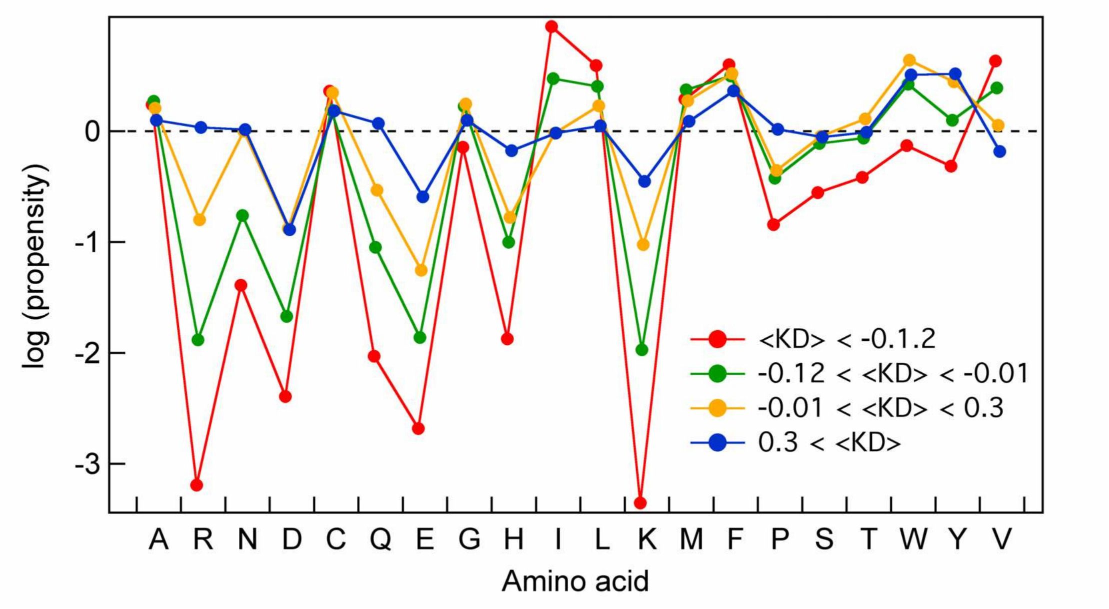

Fig. 3.7 Log Helix (all residues) propensities of 20 amino acids as a function of average hydropathy values (see Fig. 2.4 for the corresponding population distributions).

The entire dataset was divided into four groups according to the average KD hydropathy value: 1) strongly hydrophobic: less than –1.2, 2) hydrophobic: between –0.12 and –0.01, 3) neutral: between –0,01 and 0.3, and, 4) hydrophilic: more than 0.3.

For Ala, Cys, Gly, Met and Phe, the propensities are invariant among the four groups. For charged and polar amino acids (Arg, Asn, Asp, Gln, Glu, His, Lys) the propensities increase steadily as the average hydrophobicity decreases. For non-polar residues (Ile, Leu, Val), the propensities decrease steadily as the average hydrophobicity decreases. However, for Ser, Thr, Trp and to some extent Tyr, the strongly hydrophobic helices stand out from the rest. Interestingly, for these residues, the difference between the KD scale and the WW scale is maximum. The KD scale correlates more with the strongly hydrophobic helices while the rest, especially the neutral and the hydrophilic group correlate better with the WW scale.

The helices that were found to be difficult-to-predict (<KD> > 0.0), show unique amino acid propensity. All aromatic amino acid residues (Trp, Tyr and Phe) are over-represented. Three charged residues (Glu, Asp and Lys) are under-represented. Most surprisingly, the three hydrophobic residues that are typically over-represented in TM helices (Val, Ile and Leu), show almost neutral preference in under-represented in difficult-to-predict TM helices. This is an important observation. If current TM helix prediction methods are to be improved, special attention must be paid to understand how to capture this special feature in the prediction algorithm.

# Chapter 4

## Analysis of Transmembrane Helices: Three Dimensional Features

### 4.1 Introduction

In chapter 2 and chapter 3 we looked at the overall hydropathy and amino acid propensities of TM helices. The dataset used was simple amino acid sequences annotated as either TM helices or not. Although such 2D dataset gives valuable information, it is also important that the 3D structures of transmembrane proteins be considered.

Although only a few 3D structures of integral membrane proteins are known, the total number of helices in such a dataset is large enough (over 100) for statistical analysis. In this chapter we attempt to analyze TM helices in proteins whose 3D structures are known. Specifically we ask two questions.

The first is to examine the sequence annotation of TM helices in these proteins. Typically, 3D structures of integral membrane proteins comes from X-ray crystallography and so the crystal structure doesn't have a clear demarcation of where the membrane boundaries would lie when the protein is membrane bound. The standard way to annotate TM helices in proteins with known 3D structure is to use secondary structural features (alpha-helix). We wanted to examine this point.

An additional information associated with membrane proteins whose 3D structures are known is that it contains information about how the TM helices are packed. In simple sequence based methods, either hydropathy value of each residue or the amino acid kind at each sequence position is used for any prediction. However, each amino acid, independent of its kind or its hydropathy value, may contribute differently depending on if the residue is exposed to the lipid molecules in the membrane or if it is buried inside the packed helices. We try to address this point in this chapter and try to identify any trend that lipid exposure of residues may show with the predictability of the helix.

## 4.2 Materials & Methods

### *4.2.1 Dataset*

The dataset used is a select set (1L7V, 1JSQ,1FQY, 1A91, 1BGY, 1OCC, 1AR1, 1EHK, 1EUL, 1L0V, 1QLA, 1E12, 1BL8, 1KZU, 1LGH, 1MSL, 1FX8) of TM proteins from the MPTopo dataset ( http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html ). The total number of helices in the dataset is 139.

### *4.2.2 Accessible surface area (ASA)*

Solvent accessible surface area (ASA) were calculated using the DSSP program (Kabsch and Sander, 1983). Although, strictly speaking this area corresponds to as accessible to solvent (water) molecules, in this work, TM helix residues with high ASA is considered to be accessible to lipid molecules. The absolute ASA was divided by a maximum ASA value for all residues (Ahmad et al., 2003) to get the fraction ASA. For 1jsq, the PDB file only has C-alpha atoms and so no ASA were computed (6 helices).

### *4.2.3 Annotation of TM helices*

Two types of annotations were used to identify TM helices. The first was obtained from the MPTopo database. It identifies a continuous stretch of helical residues traversing the membrane as TM helical stretches. The other annotation was done by using OM program (Basu, G. unpublished). The basic algorithm of OM program is shown in Fig. 4.1. For annotations in this work, the protein was not tilted or rotated. Only translation along the z-direction was considered. The WW hydropathy scale was used for calculating mismatch energies.

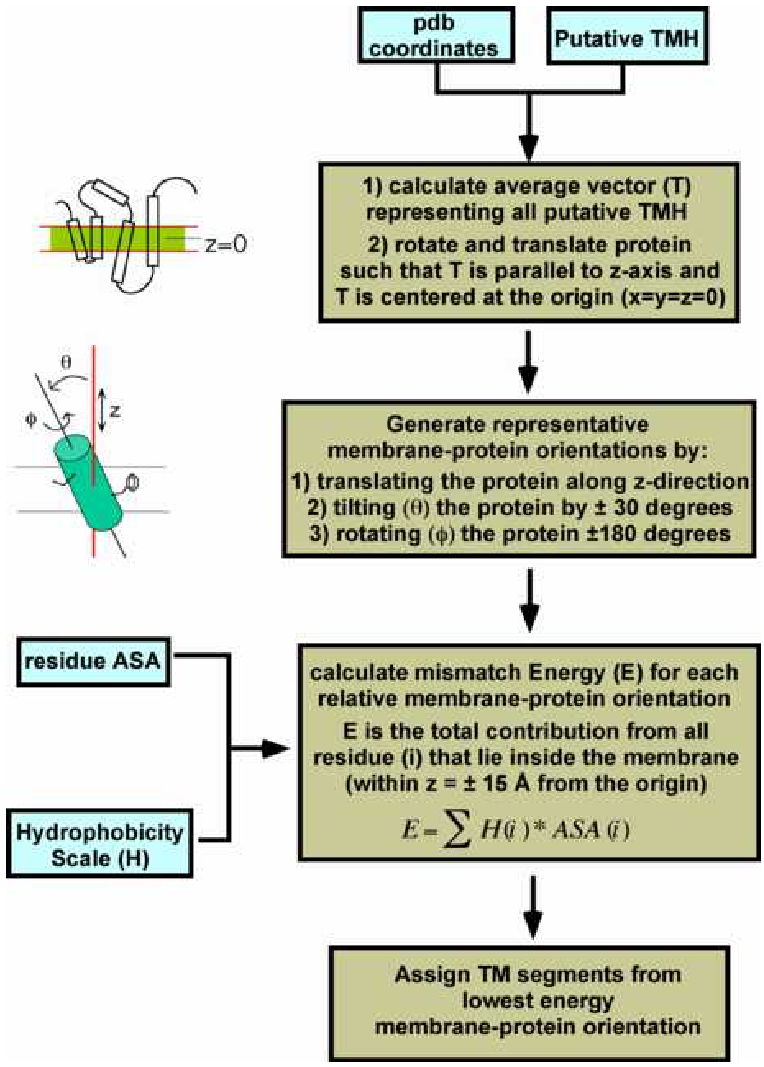
pdb coordinates
Putative TMH
1) calculate average vector (T) representing all putative TMH
2) rotate and translate protein such that T is parallel to z-axis and T is centered at the origin (x=y=z=0)
z=0
θ
φ
z
Generate representative membrane-protein orientations by:
1) translating the protein along z-direction
2) tilting (θ) the protein by ± 30 degrees
3) rotating (φ) the protein ±180 degrees
residue ASA
Hydrophobicity Scale (H)
calculate mismatch Energy (E) for each relative membrane-protein orientation
E is the total contribution from all residue (i) that lie inside the membrane (within z = ± 15 Å from the origin)
$E = \sum H(i) * ASA(i)$
Assign TM segments from lowest energy membrane-protein orientation

Fig. 4.1. The OM algorithm

## 4.3 Results & Discussion

### *4.3.1 Length distribution of TM helices*

The length distribution of TM helices in the 3D dataset is shown in Fig. 4.2. The TM helices were identified according to DSSP or OM definition. As can be seen from Fig. 4.2, the OM defined TM helix lengths are typically shorter than DSSP defined TM helix lengths. Histograms of TM helix length distributions are shown in Fig. 4.3. Again, the difference in the two assignments is very clear in this figure. In Fig. 4.4 we explicitly show the difference in TM helix assignment by the two methods for cytochrome c oxidase. Consistently, the DSSP-assigned helices are longer.

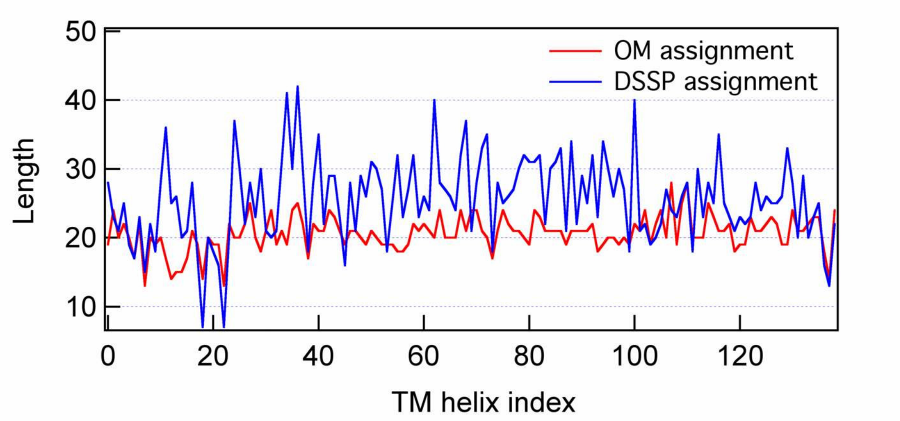

Fig. 4.2 TM lengths for all TM helices in the 3D dataset

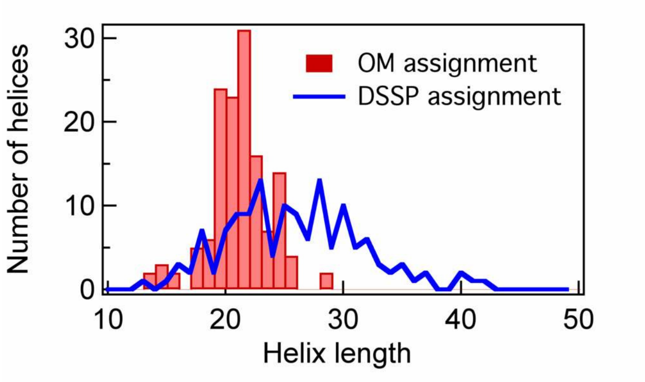

Fig. 4.3 TM length distributions in the 3D dataset.

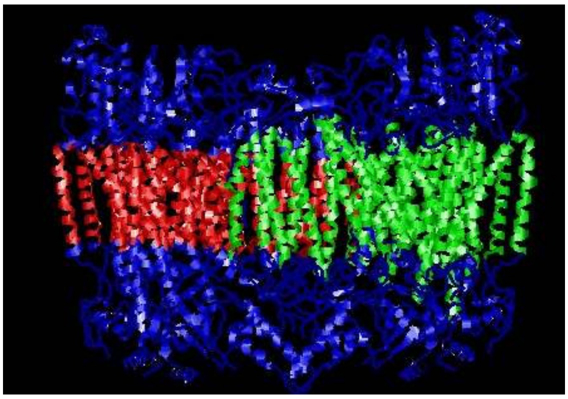

**Fig. 4.4 Dimeric cytochrome c oxidase (PDB code: 1occ). In one monomer the DSSP-assigned TM helices are shown (green) and in the other monomer the OM-assigned TM segments are shown (red).**

The meaning of the DSSP-assigned TM helices is as follows: they are contiguous stretches of helical residues in TM helices. The sequence contiguity is considered even when the helices are no more inside the membrane. On the other hand, OM-assigned TM helices only consider helical (or non-helical) residues that are within the membrane.

Both assignments are correct, in that they have different meanings. So, when predictions are made, it is important that the meaning of the two assignments are kept in mind when assessing the prediction accuracy.

It is also important to note that the OM-assigned TM helix length distributions match the TMPDB-2D data set (Fig. 2.3), indicating that OM-assignments reflect simple experimental methods that are used to identify TM segments in the 2D dataset.

### *4.3.2 Prediction accuracy*

The prediction accuracy of TM helices in the 3D dataset is shown in Table 4.1 (compared to DSSP assignment) and Table 4.2 (compared to OM assignment).

Table 4.1. Prediction accuracy of 3D dataset (DSSP assignment).

| METHOD | $Q_{seg}^{\%obs}$ | $Q_{seg}^{\%prd}$ | $Q_{num}$ | $Q_{topo}$ | $Q_2$ | $Q_{res}^{\%obs}$ | $Q_{res}^{\%prd}$ |
|---|---|---|---|---|---|---|---|
| HMMTOP 2.0 | 90 | 94 | 75.7 | 70.3 | 93.9 | 63.3 | 89.4 |
| KD (19-res) | 75 | 81.4 | 59.5 | 54.1 | 91.9 | 69.8 | 71.2 |
| DSSP | 97.1 | 97.8 | 97.3 | 91.9 | 95.3 | 70.4 | 94.4 |
| TMHMM 2.0 | 87.9 | 96.1 | 67.6 | 64.9 | 94.1 | 62.4 | 92.8 |

Table 4.2. Prediction accuracy of 3D dataset (OM assignment).

| METHOD | $Q_{seg}^{\%obs}$ | $Q_{seg}^{\%prd}$ | $Q_{num}$ | $Q_{topo}$ | $Q_2$ | $Q_{res}^{\%obs}$ | $Q_{res}^{\%prd}$ |
|---|---|---|---|---|---|---|---|
| HMMTOP 2.0 | 89.9 | 93.3 | 75.7 | 73 | 96.1 | 78.4 | 82.7 |
| KD (19-res) | 77 | 82.9 | 59.5 | 56.8 | 93.1 | 82.2 | 62.5 |
| OM | 97.8 | 97.1 | 97.3 | 91.9 | 95.3 | 94.4 | 70.4 |
| TMHMM 2.0 | 88.5 | 96.1 | 67.6 | 67.6 | 95.9 | 75.1 | 83.3 |

When compared to prediction accuracies of the 2D dataset (for HMMTOP prediction), given in Table 2.1, the 3D dataset shows slightly better overall accuracy ($Q_{num}$). However, this may be due to a smaller 3D dataset. However, $Q_{res}^{\%obs}$ values are consistently lower, indicating that at the residue level there is more under prediction in the 3D dataset, when DSSP-assigned helices are used for calculating accuracies. However, when OM-assigned helices are used for calculating prediction accuracies, residue under prediction in 3D dataset and 2D dataset are comparable. When Tables 4.1 and 4.2 are compared, at the segment level, the accuracies are comparable. At the residue level, DSSP-assigned helix annotation show more under prediction while the over prediction in OM-assigned helices is slightly higher. In summary, we showed that OM-assigned helices are more suitable for calculating residue-based accuracies when 3D dataset is tested.

### *4.3.3 Average KD hydropathy of TM helices in the 3D dataset*

Similar to our analysis of the 2D dataset (see chapter 2), we calculated the average KD hydropathy of TM helices in the 3D dataset and the distribution of the average hydropathy values are given in Fig. 4.5. The distribution is very similar to the 2D dataset (Fig. 2.4).

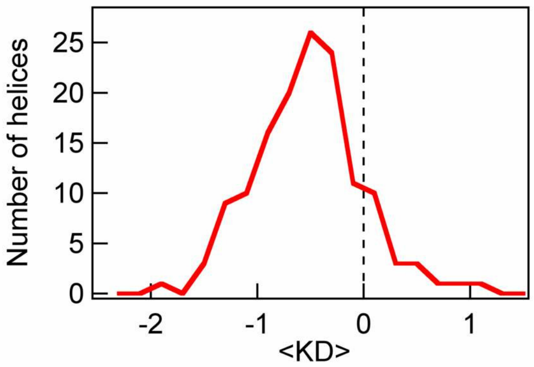

Fig. 4.5 Distribution of average KD hydropathy of TM helices in the 3D dataset.

### *4.3.4 ASA of residues in TM helices*

Using the DSSP program we calculated the ASA of all residues in the TM helices in the 3D dataset. For each helix the ASA values were summed and divided by the total number of residues to obtain the average ASA. In Fig. 4.6 the distribution of average ASA of all TM helices is shown. As can be seen from Fig. 4.6, there is a broad distribution of ASA from being buried (<ASA> < 0.1 to moderately exposed (0.4> <ASA> > 0.2).

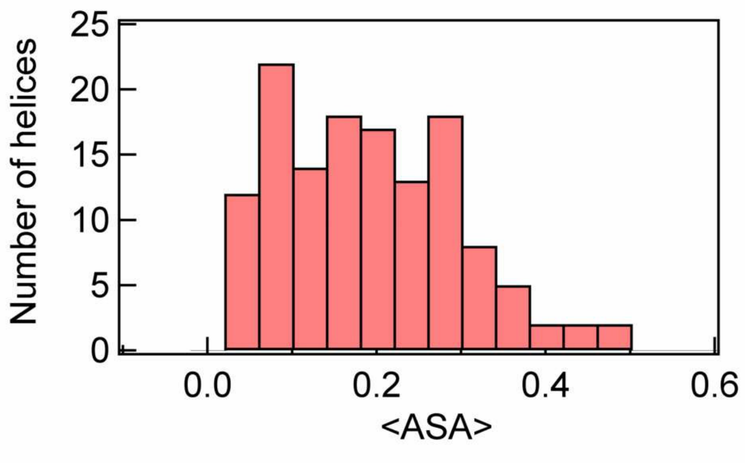

Fig. 4.6 average ASA distribution of TM helices in the 3D dataset

In order to see if the exposed (or partially exposed) residues are mostly hydrophobic or hydrophilic, we calculated average values of the product of KD hydropathy and ASA for the dataset. The distribution of this average, <KD.ASA> is shown in Fig. 4.7. The average of the product is mostly negative indicating that overall, the exposed residues are hydrophobic. However, a small but finite fraction of helices (about 10%) are hydrophilic when their exposed residues are considered. In addition, a large fraction of helices are only mildly hydrophobic.

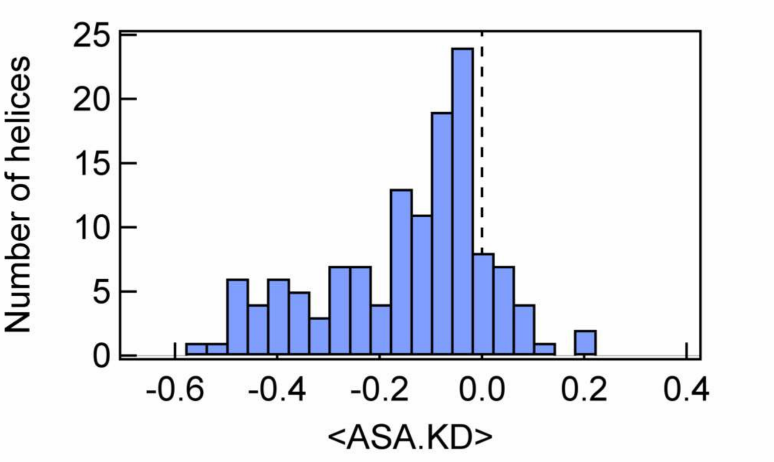

**Fig. 4.7 Average ASA.KD distribution of TM helices in the 3D dataset**

The correlation between <ASA> and <KD> is shown in Fig. 4.8. The correlation between the two is poor indicating that helices that are overall hydrophilic (positive <KD>), are not necessarily buried.

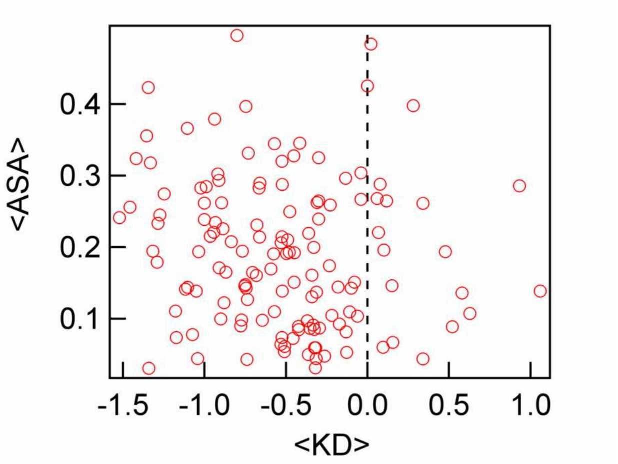

**Fig. 4.8 Average ASA vs. average ASA.KD for TM helices in the 3D dataset**

The correlation between <ASA> and <KD> is shown in Fig. 4.9. In general, as the overall ASA of helices increases, the overall hydrophobicity also increases. However, for a small fraction, the hydrophobicity remains invariant. And for some helices, exposed helices are actually hydrophilic. Fig. 4.8 and 4.9, together, indicate that the small number of helices that are hydrophilic or mildly hydrophobic can be lipid exposed as well as buried.

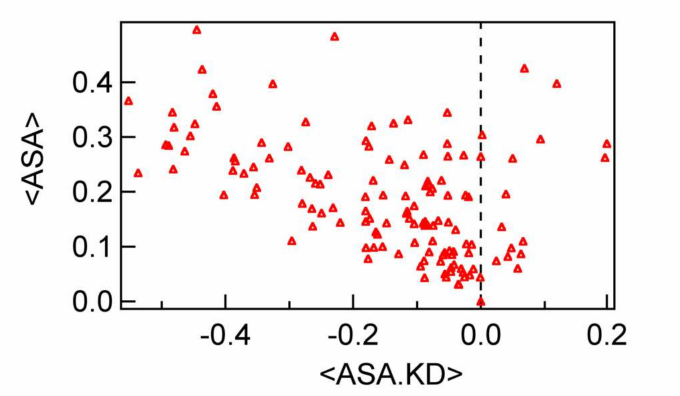

Fig. 4.9 Average ASA vs. average ASA.KD for TM helices in the 3D dataset

The correlation between <KD.ASA> and <KD> is shown in Fig. 4.10. The two quantities are well correlated. However, it should be noted that a small fraction of helices with positive <KD> show negative <KD.ASA> while another small fraction with negative <KD> show positive <KD.ASA>. This indicates that the effect of ASA in modifying <KD> values is not simple.

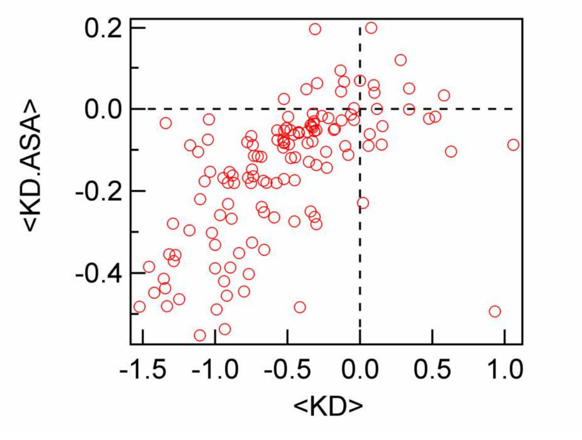

Fig. 4.10 Average ASA vs. average ASA.KD for TM helices in the 3D dataset

### *4.3.5 TM prediction as a function of <KD>, <ASA> and <ASA.KD>*

Correctly predicted helices, as a function of <KD>, are shown in fig. 4.11. Like the 2D dataset, helices with near-zero or positive values of <KD> are predicted less correctly for all methods. In addition we see that the KD method cannot predict a fraction of helices even when the <KD> is positive. These difficult-to-predict helices mostly are contiguous with one correctly predicted helix. The problem is with the algorithm that translates the hydropathy profile into distinct TM segments.

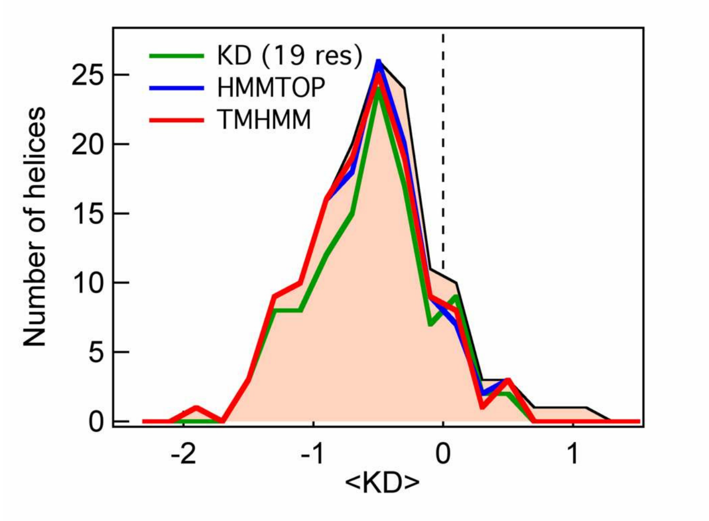

**Fig. 4.11 Prediction of TM helices in 3D dataset as a function of <KD>**

Correctly predicted helices, as a function of <ASA>, are shown in fig. 4.12. The trend with prediction and <ASA> is not very clear. However, the KD method tends to predict all helices with <ASA> greater than 0.3 (exposed).

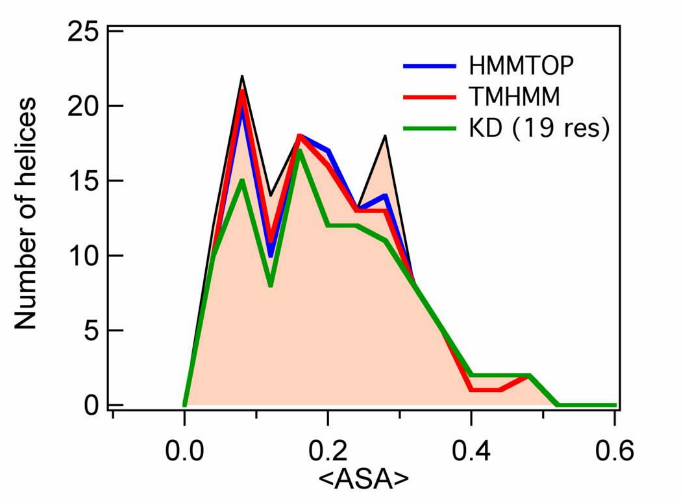

**Fig. 4.12 Prediction of TM helices in 3D dataset as a function of <ASA>**

Correctly predicted helices, as a function of <ASA.KD>, are shown in fig. 4.13. The best methods tend to underpredict helices with high <ASA.KD>. For KD prediction, the trend is not so clear

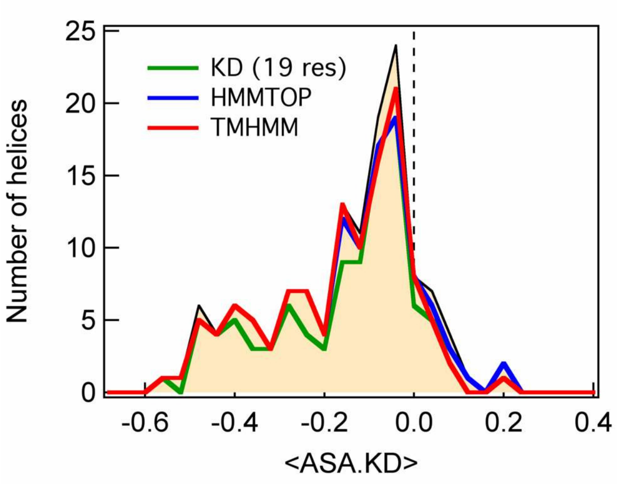

**Fig. 4.13 Prediction of TM helices in 3D dataset as a function of <ASA.KD>**

One of our expectations was that the results of KD prediction, although far less than satisfactory, would preferentially predict exposed TM helices. If that was the case then we wanted to ascribe a clear meaning to KD predictions. However, no such clear correlation was found. This is also clear from Fig. 4.14 where as a function of <KD> and <ASA>, all correctly and incorrectly predicted TM helices (by KD method) are shown.

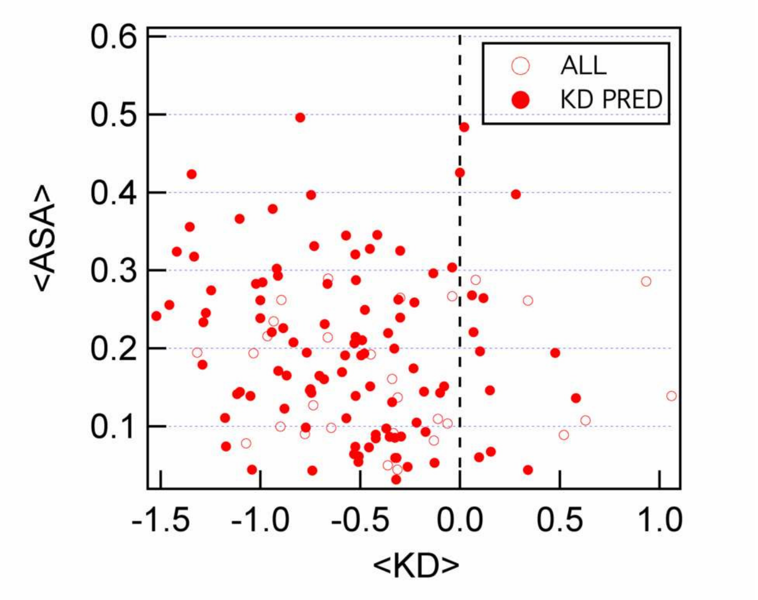

**Fig. 4.14 Correctly and incorrectly predicted helices by KD method.**

Fig. 4.14 shows that the KD method correctly predicts helices that are exposed (<ASA> more than 0.3). However there is no clear trend in prediction of buried helices. One problem of with the KD method is that contiguous helices with negative <KD> values are predicted as one long helix. This automatically makes it impossible to predict all helices with negative <KD> values. The problem is with the algorithm that translates hydrophobic segments into distinct helices.

In other words the story is not as simple as we first thought it might be. A clear meaning to KD predictions is not so straightforward. Also there was no clear correlation between difficult-to-predict helices and their lipid accessibility.

# Chapter 5
# Conclusion

Here we summarize the main results of this work.

The main goal of this work was to understand the nature of TM helices in proteins. The ultimate aim was to analyze the nature of correctly predicted and the incorrectly predicted helices so that an understanding about limitations of TM helix predictions can be achieved.

In chapter 2 we analyzed TM helices in terms of their average hydropathy values. Our analysis showed the following:

1. TM helices are mostly hydrophobic, however, about 10% of all helices are overall hydrophilic.
2. The distribution of average hydropathy of TM helices do not correlate with TM helix length
3. The performance of simple hydrpathy-based prediction is poor when compared to more advanced methods.
4. The more hydrophilic helices are difficult to predict, by hydropathy-based as well as more advanced methods.

In chapter 3 we analyzed TM helices in terms of amino acid propensities. Our analysis showed the following:

1. Propensities of short loop residues correlated well with simple hydropathy values (negative correlation). However for Ser, Thr, Trp and Tyr, they didn't match with experimental turn propensities.
2. Propensities of four residues flanking the helix termini were slightly different from that of short loop residues. The propensities also confirmed the positive-inside rule.

3. Propensities of helical residues correlated well with hydropathy-based scales. Between KD and WW scales, the correlation with WW scale was better.
4. The helix (center) propensities matched better with the KD scale while the helix (end) propensities matched better with the WW scale.
5. When helix propensities were analyzed as a function of average KD hydropathy of helices, propensities of the less hydrophobic helices correlated better with the WW scale while the KD scale correlated better with more hydrophobic residues.
6. Helices that were difficult-to-predict showed unique amino acid preferences. Aromatic residues (Tyr, Trp and Phe) were over-represented, true charged residue (Asp, Glu and Lys) were under-represented, and all other residue (including Val, Ile and Leu) showed neutral preferences.

In chapter 4 we analyzed TM helices in terms of their 3D structure. Our analysis showed the following:

1. The average length of TM segments in the 3D dataset is longer than the 2D dataset when annotations are done on the basis of helical secondary structure of residues (DSSP).
2. The average length of TM segments in the 3D dataset is comparable to that in the 2D dataset when annotations are done on the basis of consistent stretches of 30 Å around the protein surface that is most hydrophobic (OM).
3. There was no strong correlation between ASA and the hydropathy of the TM helices.
4. The ASA (or the <KD.ASA>) values of predicted and difficult-to-predict helices showed no clear trend.

This work is very preliminary. In addition, the dataset (especially the 3D dataset) is too small to draw any clear conclusion. But overall, we showed that weakly hydrophobic and hydrophilic helices are the most difficult to predict. Future improvement in TM helix prediction methods must take this fact into account. Also, no clear correlation between difficult-to-predict helices and their average ASA indicates that the effect of lipid exposure in the final structure may not play an important role in deciding whether or not a segment will be TM or not.

# Acknowledgement

First of all, I would like to express my gratitude to Associate Prof. Gautam Basu and Associate Prof. Takeshi Kawabata, my supervisors. They introduced me to the world of bioinformatics during numerous extended discussion in or laboratory. I also would like to thank Prof. Nobuhiro Go, Kensuke Nakamura, Keiko Matsuda, Saharuddin Bin Mohamad, Jitender Jit Singh Cheema, for helpful comments on this thesis.

Also, I would like to thank the members of bioinformatics work group in our laboratory: Naoshi Fukuhara, Daisuke Ishibe, Masataka Oikawa, and Hiromichi Yamagiwa. My significant interest to the computational biology is founded during the continuous discussion with them. From this excellent work group, I learned a lot about biology, chemistry and bioinformatics from the view points of theoretical analysis as well as real-world phenomenon.

Finally, I would like to thank my wife Yasue, she always supported me while I was writing this thesis.

# References

Ahmad, S., Gromiha, M. M, and Sarai, A. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins: Struct. Funct. Genet.* 50, 629-635.

Branden, C. and Tooze, J. (1999) ***Introduction to Protein Structure***. Garland Publishing, New York.

Chen, C. P. and Rost, B. (2002) State-of-the-art in membrane protein prediction *App. Bioinformatics* 1, 21-35.

Crasto, C. J. and Feng, J. (2001) Sequence codes for extended conformations: A neighbor-dependent sequence analysis of loops in proteins. *Proteins: Structure Funct. Genet.* 42, 399-413.

Jayasinghe, S., Hristova, K. and White, S. (2001) Energetics, stability and prediction of transmembrane helices. *J. Mol. Biol.* 312, 927-934

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

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-2637.

Kyte J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein (1982) *J. Mol. Biol.* 157, 105-132.

Marchese, A. (1999) Novel GPCRs and their endogenous ligands: expanding the boundaries of physiology and pharmacology. *Trends Pharmacol Sci* 20, 370-375.

Monne M, Nilsson I, Elofsson A, von Heijne G. (1999) Turns in transmembrane helices: determination of the minimal length of a "helical hairpin" and derivation of a fine-grained turn propensity scale. *J. Mol. Biol.* 293, 807-14.

Nilsson. J., Persson, B., and Von Heijne, G. (2002) Prediction of partial membrane protein topologies using a consensus approach. *Protein Sci.* 11, 2974-80.

Sonnhammer, E. L. L., von Heijne, G. and Krogh, A. (1998) A hidden Mrkov model for predicting trans-membrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol.* 6,175-82.

Tusnasy, G. E. and Simon, I. (1988) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* 283, 489-506.

Von Heijne, G. (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with trans-membrane topology. *EMBO J.* 5, 3021-3027.

Wallin, E. and von Heijne, G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.
*Protein Sci.*7, 1029-38.

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