Research Proposal – Predicting Protein trans membranal Domain with Hidden Markov Model

# Background and Motivation

trans membranal proteins are among the most crucial proteins for the survival of the cell. Estimated to comprise 27% of the proteins found in humans, transmembrane proteins perform crucial roles such as transport of nutrients and cellular communication.

As such, the ability to identify and predict the transmembrane domain of proteins is very attractive, as it is helpful in predicting a proteins function.

An amino acid in a trans membranal protein have two distinct hidden phases that it can be in – either it is inside the membrane or outside of it. The acid also has multiple known phases it can be in (which type of amino acid) which can be considered dependent on the hidden phase. These qualities make it ideal to predict using a hidden Markov model.

# Research Question and Objective

Is it possible to predict the trans membranal domain in proteins using hidden Markov model?

Our objective is creating a reliable model representing protein, capable of determining the existence and identifying the trans membranal regions of a given protein sequence, using the Baum-Welsh algorithm and existing public data.

# Data

The data is comprised of a large number of sequences representing trans membranal proteins, represented by their canonical single letter code.

Text

Description automatically generated  
Each sequence is represented by a CHAIN tag, which has SEQ and REGION inside.  
From the SEQ part we extracted the sequence that represents the protein, the chain id, number of transmembrane segments and the type of transmembrane segments (alpha, beta or coil). From the REGION part we extracted the type or the region (the location of the chain segment relative to the membrane), and the beginning and end of each region in the sequence. We decided to concentrate only on segments of type alpha.

The data is taken from [*PDBTM: Protein Data Bank of Transmembrane Proteins*](http://pdbtm.enzim.hu/).

# The Model

The model we will use to analyze the data is the TCM model – Two Component Mixture.  
This model is using the EM algorithm with HMMs to find motifs is biological sequences.  
As we saw in class, the TCM model is one of three models: OOPS – One Occurrence per Sequence, ZOOPS – Zero or One Occurrence per Sequence and TCM, which can find zero, one or multiple occurrences in a sequence.

1

1

1

p

…

p-1

A picture containing rectangle

Description automatically generatedThis model is the most fitting for our data, as there could be many trans membranal regions is a protein (2), one region (1), or not at all.

The parameters of the model are transitions - and emissions -   
Those parameters will be learned using Baum – Welch algorithm, which uses MLE:

- the sufficient statistics for the MLE

s.t

d

We will use the algorithm that finds a that maximizes the log-likelihood estimator: