Hackathon Research Proposal - Predicting protein trans-membranal($\alpha - helical$) domain using HMM

Oren Sultan, Roee Lieberman, Maria Tseytlin, Roei Zucker, Lee Lankri

26-28/02/21

1 Background

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. In this work we focused on predicting $\alpha - helical$ transmembrane proteins by using HMM.

2 Research Question & Objective

Is it possible to predict the $\alpha - helical$ trans membranal domain in proteins using HMM? Our objective is creating a reliable model representing protein, capable of identifying the $\alpha - helical$ trans membranal regions of a given protein sequence.

3 Data

Our source database is in format of XML. The data is taken from PDBTM(7). We are looking in the scope of chains which include: **CHAINID:** the chain identifier **NUM_TM:** the number of transmembrane segments **TYPE:** the type of transmembrane segments or the type of the chain if it does not cross the membrane (non_tm) or if it is not a protein chain (lipid), we take only "alpha" segments. Every chain contains the following data: **SEQ:** the sequence of the protein **REGION:** locates the chain segment in the space relative to the membrane. we are looking only on TYPE="H" which is $\alpha - helical$ region. We take the "seq_beg" and "seq_end" to know the starting position and ending position of the region in the sequence. In the next figure we can see the structure of the CHAIN element with an example:

```
<CHAIN CHAINID="A" NUM_TM="7" TYPE="alpha">
    <SEQ>
        AVRENALLSS SLWVNVALAG IAILVFVYMG RTIRPGRPRL IWGATLMIPL
        VSISSYLGLL SGLTVGMIEM PAGHALAGEM VRSQWGRYLT WALSTPMILL
        ALGLLADVDL GSLFTVIAAD IGMCVTGLAA AMTTSALLFR WAFYAISCAF
        FVVVLSALVT DWAASASSAG TAEIFDTLRV LTVVLWLGYP IVWAVGVEGL
        ALVQSVGATS WAYSVLDVFA KYVFAFILLR WVANNERTVA VAGQTLGTMS
        SDD
    </SEQ>
    <REGION seq_beg="10" pdb_beg="31"  seq_end="31" pdb_end="52"  type="H"/>
    <REGION seq_beg="39" pdb_beg="60"  seq_end="59" pdb_end="80"  type="H"/>
    <REGION seq_beg="85" pdb_beg="106"  seq_end="104" pdb_end="125"  type="H"/>
    <REGION seq_beg="110" pdb_beg="131"  seq_end="131" pdb_end="152"  type="H"/>
    <REGION seq_beg="139" pdb_beg="160"  seq_end="160" pdb_end="181"  type="H"/>
    <REGION seq_beg="177" pdb_beg="198"  seq_end="196" pdb_end="217"  type="H"/>
    <REGION seq_beg="209" pdb_beg="230"  seq_end="230" pdb_end="251"  type="H"/>
</CHAIN>
```

Figure 1: Dataset's structure

4 Model

We will use HMM. Our model is trying to differentiate between two major stages, an 'in' stage where the protein is inside the membrane and an 'out' state(background stage) where the protein can be either inside the cell or outside the cell. In both stages the sequence length isn't constant, that is why we used two different emissions for those stages. In the actual model the $'B'$ stage model the $'out'$ state and all the other $'SM'$ and $'LM'$ stages model the $\alpha - helical$ that go through the cell membrane. The $'SM'$ stands for short motif and those stages are for the more likely motif length as seen in the data. Each $'SM\{i\}'$ node has the option for transition, going back to the $'B'$ stage (end of current motif) or keeping to $'SM\{i+1\}'$, the aim of this architecture is to give the model better control in those motifs' length. The $'LM'$ are for the rest and longer possible motifs each $'LM\{i\}'$ can transit only to $'LM\{i+1\}'$ except the last $'LM'$ which can go to $'B'$ or to keep going to itself, that makes the model able to sample any length of motif but with less control in those lengths.

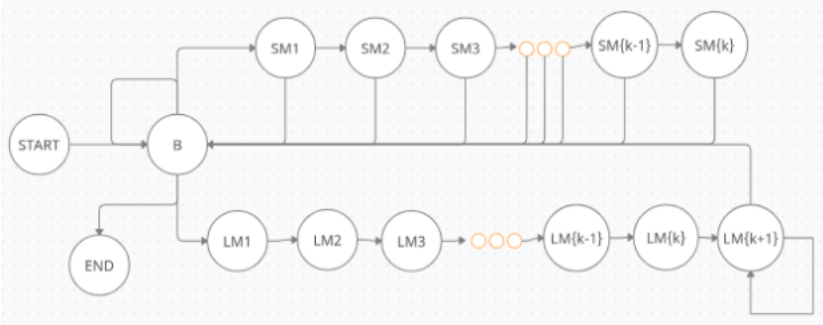

Figure 2: The HMM Model. $'B'$-Background, $'SM'$-short motif, $'LM'$-long motif. In our case $k = 30$.

5 Algorithm - HMM Supervised Learning

We will use supervised training to train and initiate the HMM. We will generate labels for the regions in the protein sequences according to the model architecture we explained in section 4, by using the data of the regions in the sequences gathered from the PDBTM as explained in section 2. In the next figure we can see an example of observation and the label:

```
$,M,E,V,N,Q,L,G,F,I,A,T,A,L,F,V,L,V,P,S,V,F,L,I,I,L,Y,V,Q,T,E,S,Q,Q,K,S,S,^
start,B,B,B,B,B,B,B,B,B,B,SM1,SM2,SM3,SM4,SM5,SM6,SM7,SM8,SM9,SM10,SM11,SM12,SM13,B,B,B,B,B,B,B,B,B,B,end
```

Figure 3: Example for observation(the first row) and label(the second row) from the generated training data

If we are trying to parameterize our HMM model using simple discrete distributions, we can simply apply the MLE to compute the transition and emission distributions by counting the number of transitions from any given state to another state. Similarly, we can compute the emission distribution by counting the output states from different hidden states. Therefore the transition and emission probabilities can be computed using the MLE.