Predicting protein trans-membranal($\alpha - helix$) domain using HMM

Oren Sultan, Maria Tseytlin, Roee Lieberman, 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 $\alpha - helix$ transmembrane proteins by using HMM.

2 Research Question & Objective

Is it possible to predict the $\alpha - helix$ trans membranal domain in proteins using HMM? Our objective is creating a reliable model representing protein, capable of determining the existence and identifying the $\alpha - helix$ trans membranal regions of a given protein sequence, using the Baum-Welsh algorithm and existing public data.

3 Data

Our source database is in format of XML. The data is taken from: pdbtm.enzim.hu/

Figure 1: Dataset's structure

```
<CHAIN CHAINID="A" NUM_TM="7" TYPE="alpha">
        <SEQ>
                AVRENALLSS SLWVHVALAG 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="1" pdb_beg="22"  seq_end="2" pdb_end="23"  type="U"/>
        <REGION seq_beg="3" pdb_beg="24"  seq_end="9" pdb_end="30"  type="1"/>
        <REGION seq_beg="10" pdb_beg="31"  seq_end="31" pdb_end="52"  type="H"/>
        <REGION seq_beg="32" pdb_beg="53"  seq_end="38" pdb_end="59"  type="2"/>
        <REGION seq_beg="39" pdb_beg="60"  seq_end="59" pdb_end="80"  type="H"/>
        <REGION seq_beg="60" pdb_beg="81"  seq_end="84" pdb_end="105"  type="1"/>
        <REGION seq_beg="85" pdb_beg="106"  seq_end="104" pdb_end="125"  type="H"/>
        <REGION seq_beg="105" pdb_beg="126"  seq_end="109" pdb_end="130"  type="2"/>
        <REGION seq_beg="110" pdb_beg="131"  seq_end="131" pdb_end="152"  type="H"/>
        <REGION seq_beg="132" pdb_beg="153"  seq_end="138" pdb_end="159"  type="1"/>
        <REGION seq_beg="139" pdb_beg="160"  seq_end="160" pdb_end="181"  type="H"/>
        <REGION seq_beg="161" pdb_beg="182"  seq_end="176" pdb_end="197"  type="2"/>
        <REGION seq_beg="177" pdb_beg="198"  seq_end="196" pdb_end="217"  type="H"/>
        <REGION seq_beg="197" pdb_beg="218"  seq_end="208" pdb_end="229"  type="1"/>
        <REGION seq_beg="209" pdb_beg="230"  seq_end="230" pdb_end="251"  type="H"/>
        <REGION seq_beg="231" pdb_beg="252"  seq_end="241" pdb_end="262"  type="2"/>
        <REGION seq_beg="242" pdb_beg="263"  seq_end="253" pdb_end="274"  type="U"/>
</CHAIN>
```

The primary element in the XML that we are looking for is the "CHAIN" which contains the following attributes:

- **CHAINID:** the chain identifier
- **NUM_TM:** the number of transmembrane segments
- **TYPE:** the type of transmembrane segments (alpha, beta or coil (i.e. non alpha and non beta)) 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 are looking only on "alpha" segments which are chains of proteins that cross the membrance with the type of $\alpha - helix$.

Each Chain contains the following data:

- **SEQ:** the sequence of the protein
- **REGION:** locates the chain segment in the space relative to the membrane.
  - The REGION can be one of many types, we are looking only on TYPE="H" which is $\alpha - helix$.
  - We take the seq_beg and seq_end to know the starting position and ending position of the region in the sequence.

4 Model

5 Tools and Algorithms

6 References