Remote Homology Detection With Augmented Hidden Markov Models

Alexander Berezovsky, Shehjar Sadhu, Prof. Noah Daniels

Department of Computer Science and Statistics.

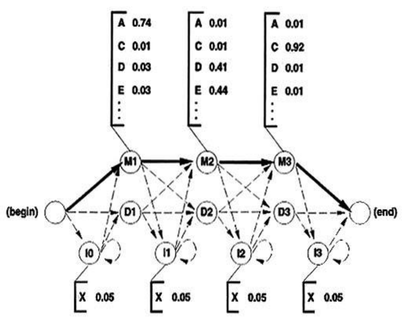
Introduction

In this project we aim to classify proteins into their respective superfamilies given their amino acid sequence. This is a form of hierarchical classification problem. Once we are able to classify them their structure can be inferred and given their structure we can infer their function. It is a central problem of bioinformatics to do so because once we know the function of a protein we know what it does.

When classifying unknown amino acid sequences a problem that occurs is dealing with sparse datasets. One solution that was proposed was to use simulated evolution implicitly changing the emission probabilities by Kumar and Cowen.

In our solution Instead, we explicitly change the emission probabilities by using the BLOSUM62 substitution Matrix by using a widely used model called Hidden Markov Models in classification of proteins.

Profile Hidden Markov Models



- Extract the emission probabilities from the hmm file
- Take the column sum from the BLOSUM62 matrix in probability space
- Take the natural log of the match state probabilities
- Take the sum of the BLOSUM62 matrix and emission probabilities from the hmm file.
- Convert the resulting vector back into probability space by taking the exponential
- Normalize the vector by taking the row sum of each row and dividing each element in that row by the sum

Software

The software we trying to modify is HMMER (http://hmmer.org). It is a freely available software developed by Sean Eddy and Travis Wheeler.

HMMER is widely used for searching sequence databases for homologs (sequences with similar structure to the query sequence)

Software Archeology

We performed software archeology on HMMER code base. There are about 100 thousand lines of code in the HMMER software. Our job was to find the right place to put the modifications.

Testing our implementation

Taking random 10 Superfamily sequences from CATH

```
>cath|current|3kowE02/45-114
TTPSIERSVLLRMGFSSLEAKAIVDKTMDRGLMGKGAGHIVYKIAKEKNISVREAGLALSEGKYWDDAIQIFKGGVK
>cath|current|3kowF02/45-114
TTPSIERSVLLRMGFSSLEAKAIVDKTMDRGLMGKGAGHIVYKIAKEKNISVREAGLALSEGKYWDDAIQIFKGGVK
>cath|current|3kowG02/45-114
TTPSIERSVLLRMGFSSLEAKAIVDKTMDRGLMGKGAGHIVYKIAKEKNISVREAGLALSEGKYWDDAIQIFKGGVK
```

Generating MSA using MUSCLE

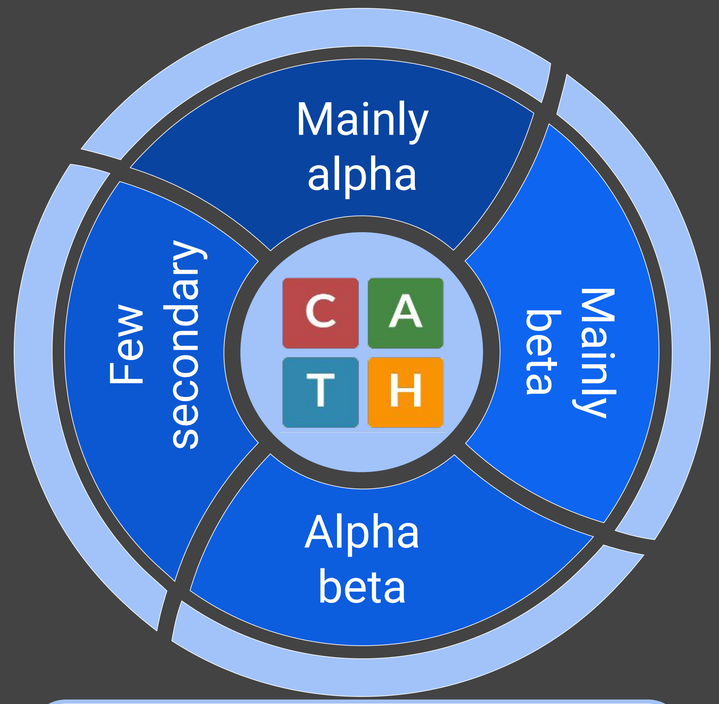
```
>cath|current|4nooB00/28-122
--------SENCNDTSGVHQKILVCIQNEIAKSETQIRNNISSKSIDYGFPDDFYSKQRL
AIHEKCXLYINVGGQRGELLXNQCELSXLQGLDIYIQQYIEDVDNS---------
>cath|current|4nooD00/28-122
--------SENCNDTSGVHQKILVCIQNEIAKSETQIRNNISSKSIDYGFPDDFYSKQRL
AIHEKCXLYINVGGQRGELLXNQCELSXLQGLDIYIQQYIEDVDNS---------
>cath|current|4nsoB00/31-124
```

Running HMM build command to generate .hmm files using newly implemented solution.

Running HMM build command to generate .hmm files using old HMM build.

```
HMM     A        C        D        E        F        G        H        I        K        L        M        N        P
        m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO  2.68448  3.85718  2.80541  2.66246  3.38685  2.84114  3.37679  2.67324  2.73892  2.42823  3.70995  2.95183  3.69363
         2.68627  4.42142  2.77511  2.73115  3.46362  2.48509  3.72583  3.29362  2.67749  2.69363  4.24546  2.98316  2.73748
         0.34471  1.93203  1.91925  1.36850  0.29643  0.00000        *
  1  4.14899  4.94814  4.79975  4.77242  5.36352  3.79610  5.26231  5.55676  4.83800  5.49522  5.00307  4.74786  4.45024
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
         0.03918  3.65494  4.37729  0.61958  0.77255  0.59507  0.80190
  2  4.77254  5.64388  4.38015  3.40449  5.41952  4.95108  5.84391  5.56186  4.51852  5.45822  5.10143  4.57052  4.64412
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
         0.03918  3.65494  4.37729  0.61958  0.77255  0.59507  0.80190
  3  4.54771  5.35942  4.49704  4.51632  5.14189  4.85492  5.08970  5.54925  4.68302  5.46386  5.07565  5.33763  4.59650
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
         0.03918  3.65494  4.37729  0.61958  0.77255  0.59507  0.80190
  4  4.36770  2.35856  5.69480  5.54372  4.87141  4.84073  5.53570  4.72079  5.28720  4.87727  4.64101  5.29340  4.62436
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
         0.03918  3.65494  4.37729  0.61958  0.77255  0.59507  0.80190
  5  4.54771  5.35942  4.49704  4.51632  5.14189  4.85492  5.08970  5.54925  4.68302  5.46386  5.07565  5.33763  4.59650
```

```
STATS LOCAL FORWARD   -4.3390  0.71680
HMM     A        C        D        E        F        G        H        I        K        L        M        N        P
        m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO  2.68448  3.85718  2.80541  2.66246  3.38685  2.84114  3.37679  2.67324  2.73892  2.42823  3.70995  2.95183  3.69363
         2.68627  4.42142  2.77511  2.73115  3.46362  2.48509  3.72583  3.29362  2.67749  2.69363  4.24546  2.98316  2.73748
         0.34471  1.93203  1.91925  1.36850  0.29643  0.00000        *
  1  2.15991  4.11238  2.93416  2.75780  4.07492  1.98990  3.95288  3.53653  2.98768  3.23843  4.10289  3.02815  3.55955
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
         0.03918  3.65494  4.37729  0.61958  0.77255  0.59507  0.80190
  2  2.78346  4.00804  2.44256  1.38988  4.13091  3.14552  3.73448  3.54083  2.58821  3.20143  4.20126  2.85161  3.74442
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
         0.03918  3.65494  4.37729  0.61958  0.77255  0.59507  0.80190
  3  2.55864  4.52366  2.63145  2.58170  3.85328  3.04863  3.78019  3.52902  2.67271  3.28707  4.17548  1.61872  3.69681
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
         0.03918  3.65494  4.37729  0.61958  0.77255  0.59507  0.80190
  4  2.37862  1.52280  3.82842  3.52910  3.58281  3.03445  4.22619  2.70056  3.35689  2.62848  3.74164  3.57449  3.72466
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
         0.03918  3.65494  4.37729  0.61958  0.77255  0.59507  0.80190
  5  2.55864  4.52366  2.63145  2.58170  3.85328  3.04863  3.78019  3.52902  2.67271  3.28707  4.17548  1.61872  3.69681
         2.68618  4.42225  2.77519  2.73123  3.46354  2.48513  3.72494  3.29354  2.67741  2.69355  4.24690  2.98347  2.73739
```



Discussion

Originally we thought of converting the BLOSUM 62 matrix in probability space into log space and working in log space from there which gave us erroneous output.

We looked back at our algorithm and we realized that if we implemented the procedure above then the end result will be in log space which is not what we want because HMMER and HMMs works with probability space.

By explicitly modifying the match state probabilities of the Hidden Markov Model using the BLOSUM62 Matrix in probability space, we get similar results to the current hmmer implementation.

Acknowledgements

- We extend our greatest thanks to Prof. Noah Daniels who helped us through this project without whos guidance this won't be possible.
- We also thank College of Arts and Sciences for funding this fellowship.