Protein Family Classification Using Profile HMMs

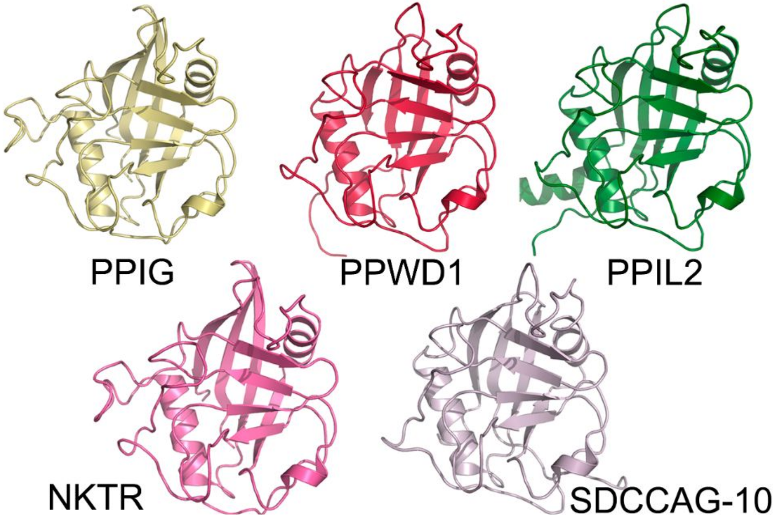
By: Levi Cavagnetto, Daniel Chang, and
Garrett Abou-Zeid

Motivation

- Why is protein family classification important?
    - If you know the family of a novel protein, you know its function!

- Why are profile HMMs useful for protein family classification?
    - A sequence that is a “distance cousin” of a family may **lack strong pairwise similarities** with **any singular sequence** within the family, but may have **subtle/weak pairwise similarities** with **many sequences** in the family
        - Thus, for diverse protein families, pairwise similarity is not a sufficient criteria for determining if a novel sequence is a family member

- Profile HMMs are more efficient to use than constructing an entirely new MSA whenever testing a novel sequence for family membership



Profile HMM

- Profile HMMs generate a probabilistic representation of a **given multiple sequence alignment**

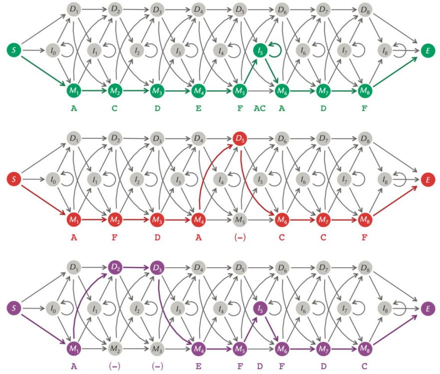
- Profile HMMs are **"trained"** by using each aligned sequence and computing their path through the HMM.
  - After all paths have been computed, transmission and emission probabilities are estimated

- Profile HMMs can then estimate **("testing")** how likely a novel sequence belongs to the family described by the MSA used for training
  - Using the viterbi algorithm, the most likely path (and subsequent log likelihood) is estimated

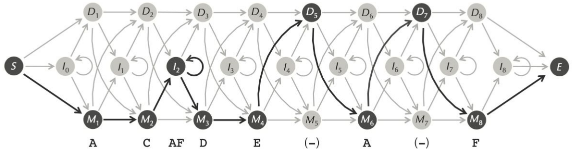
"training"

Alignment

```
A  C  D  E  F AC A  D  F
A  F  D  A  - -- C  C  F
A  -  -  E  F D- F  D  C
A  C  A  E  F -- A  -  C
A  D  D  E  F AA A  D  F
```



"testing"



Alignment

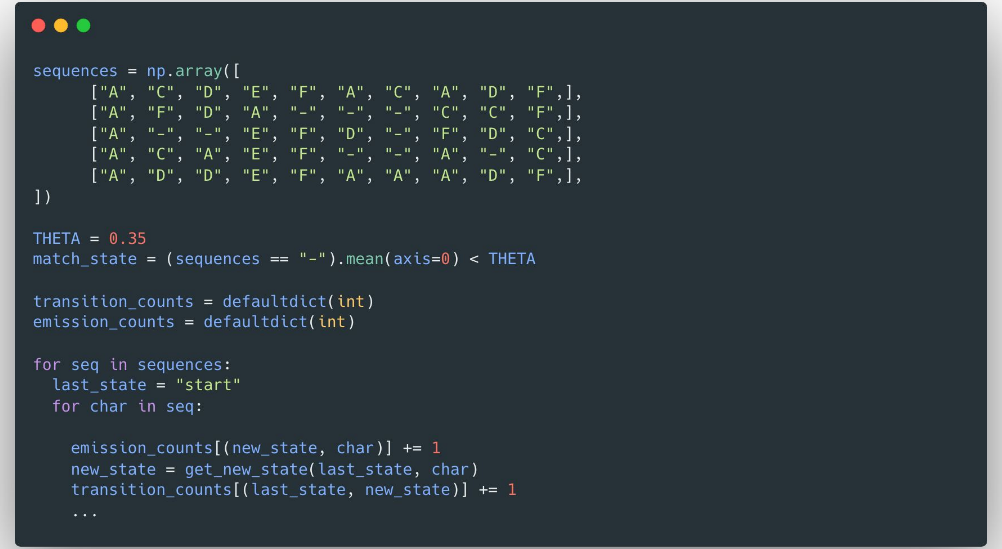
```
A  C -- D  E  F AC A  D  F
A  F -- D  A  - -- C  C  F
A  - -- -  E  F D- F  D  C
A  C -- A  E  F -- A  -  C
A  D -- D  E  F AA A  D  F
```

Text

```
A  C AF D  E  - -- A  -  F
```

Slide borrowed from Bioinformatics Algorithms: an Active Learning Approach

Profile HMM (our implementation)

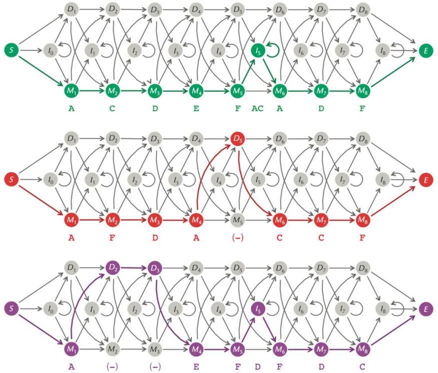


```
sequences = np.array([
      ["A", "C", "D", "E", "F", "A", "C", "A", "D", "F",],
      ["A", "F", "D", "A", "-", "-", "-", "C", "C", "F",],
      ["A", "-", "-", "E", "F", "D", "-", "F", "D", "C",],
      ["A", "C", "A", "E", "F", "-", "-", "A", "-", "C",],
      ["A", "D", "D", "E", "F", "A", "A", "A", "D", "F",],
])

THETA = 0.35
match_state = (sequences == "-").mean(axis=0) < THETA

transition_counts = defaultdict(int)
emission_counts = defaultdict(int)

for seq in sequences:
  last_state = "start"
  for char in seq:

    emission_counts[(new_state, char)] += 1
    new_state = get_new_state(last_state, char)
    transition_counts[(last_state, new_state)] += 1
    ...
```

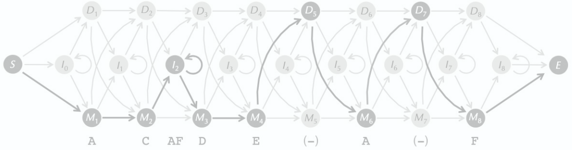
“training”

Alignment

```
A    C    D    E    F AC A    D    F
A    F    D    A    - -- C    C    F
A    -    -    E    F D- F    D    C
A    C    A    E    F -- A    -    C
A    D    D    E    F AA A    D    F
```



“testing”



Alignment

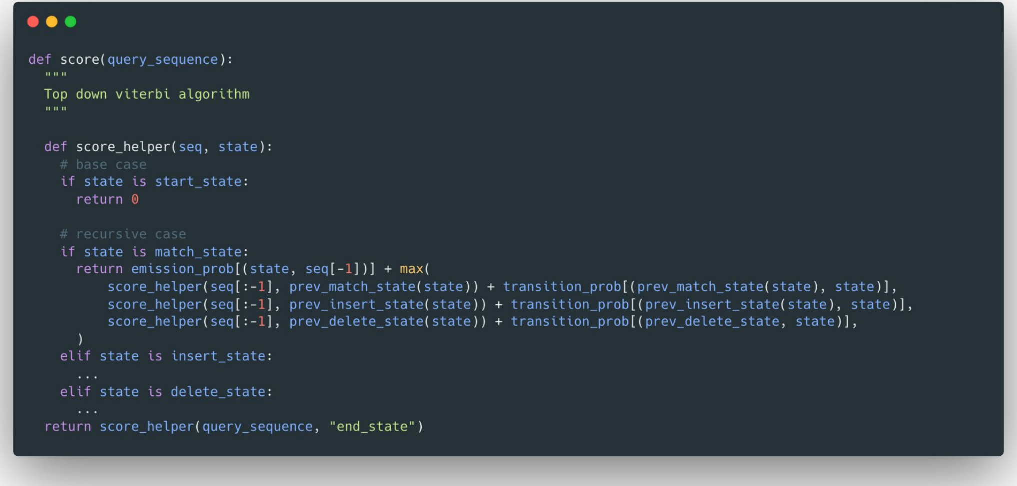
```
A    C -- D    E    F AC A    D    F
A    F -- D    A    - -- C    C    F
A    - -- -    E    F D- F    D    C
A    C -- A    E    F -- A    -    C
A    D -- D    E    F AA A    D    F
```

Text

```
A    C AF D    E    - -- A    -    F
```

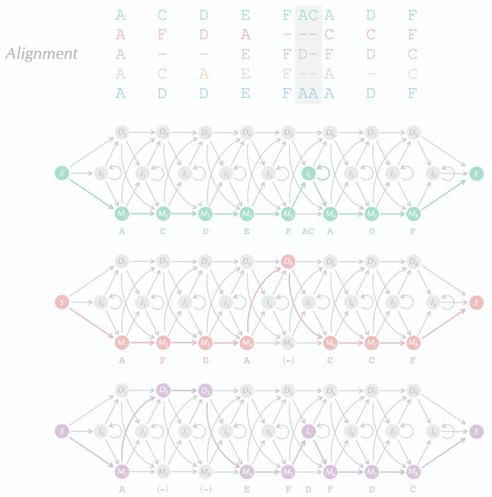
Slide borrowed from Bioinformatics Algorithms: an Active Learning Approach

Profile HMM
(our implementation)

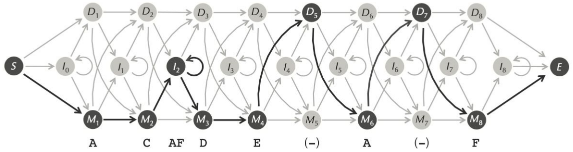


```
def score(query_sequence):
  """
  Top down viterbi algorithm
  """

  def score_helper(seq, state):
    # base case
    if state is start_state:
      return 0

    # recursive case
    if state is match_state:
      return emission_prob[(state, seq[-1])] + max(
          score_helper(seq[:-1], prev_match_state(state)) + transition_prob[(prev_match_state(state), state)],
          score_helper(seq[:-1], prev_insert_state(state)) + transition_prob[(prev_insert_state(state), state)],
          score_helper(seq[:-1], prev_delete_state(state)) + transition_prob[(prev_delete_state, state)],
      )
    elif state is insert_state:
      ...
    elif state is delete_state:
      ...
  return score_helper(query_sequence, "end_state")
```

"training"



"testing"



	A	C	--	D	E	F	AC	A	D	F
Alignment	A	F	--	D	A	-	--	C	C	F
	A	-	--	-	E	F	D-	F	D	C
	A	C	--	A	E	F	--	A	-	C
	A	D	--	D	E	F	AA	A	D	F
Text	A	C	AF	D	E	-	--	A	-	F

Slide borrowed from Bioinformatics Algorithms: an Active Learning Approach

Methodology

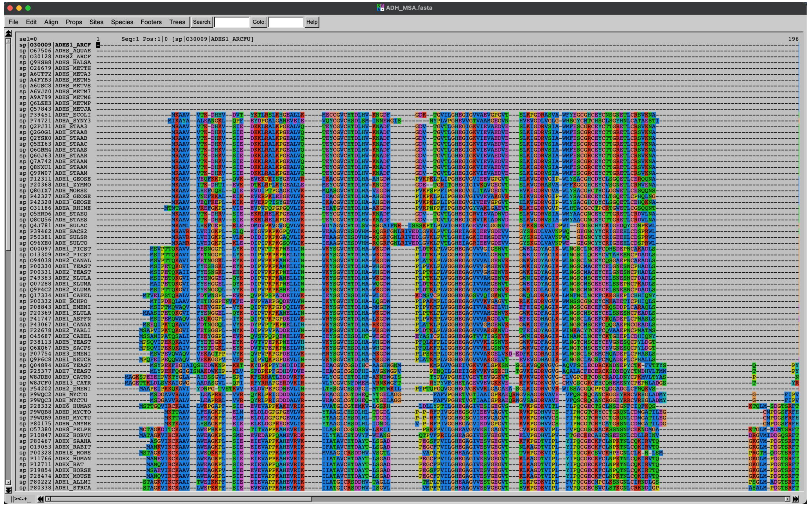
- We focused on building profile HMMs for two genes: **alcohol dehydrogenase (ADH)** and **acetaldehyde dehydrogenase (ACDH)**.

- ~300 Fasta files of the two genes were downloaded from Uniprot
  - To avoid outliers, sequences were limited to those of length 200-400
  - We created a hold out set for testing purposes

- SeaView MSA tool was used to generate MSAs
  - Obviously mismatching sequences were discarded

- Created a profile HMM implementation using python, and used an online implementation of profile HMMs for comparison

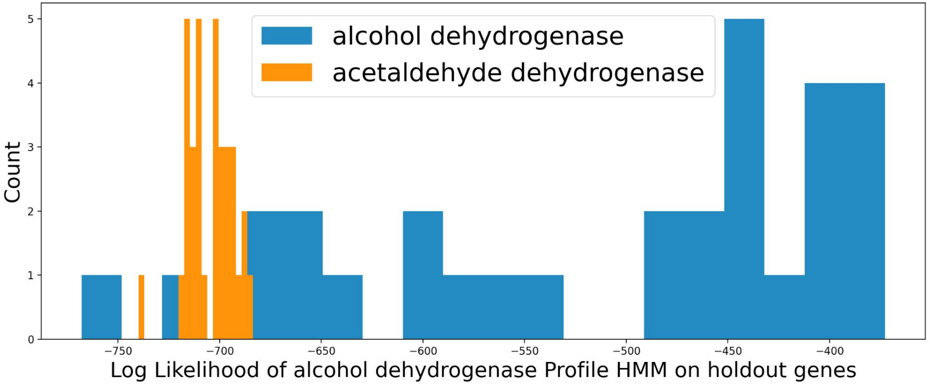
- Generated log likelihoods of the trained HMMs when run on the holdout sequences

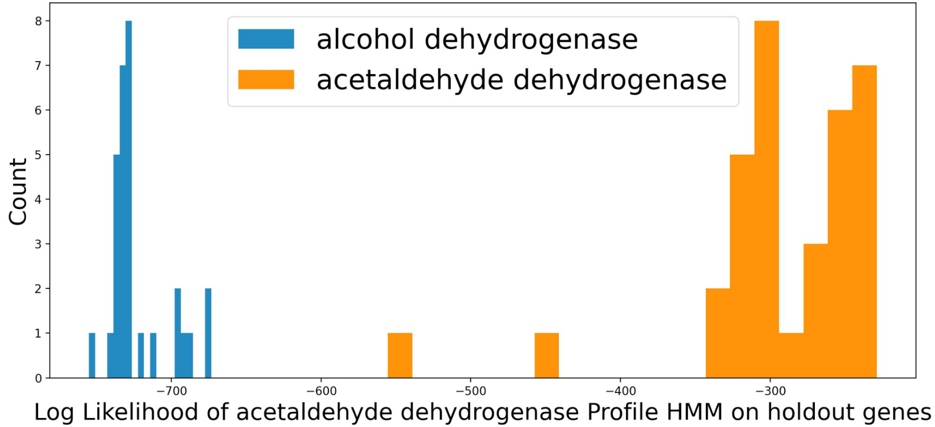


SeaView MSA

Results: our implementation

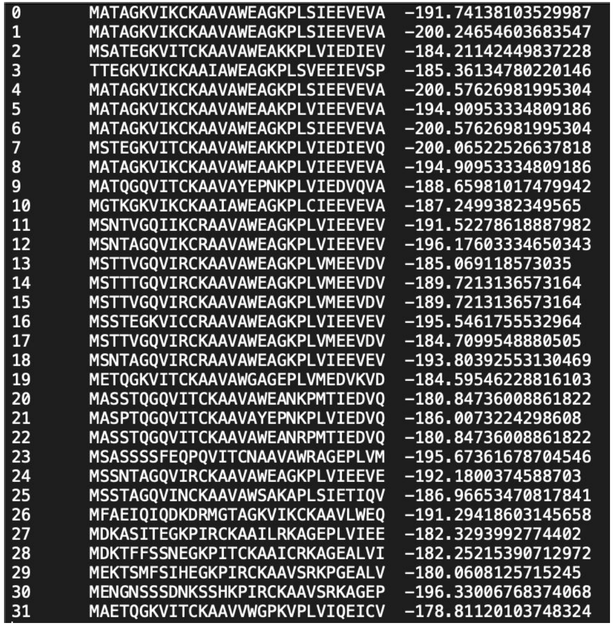
- Using our profile HMM implementation, we trained a profile HMM for both genes
- Log likelihood scores of the test set were generated using the two trained profile HMMs
- The pHMM was able to stratify the two genes fairly well
  - Alcohol dehydrogenase seems to have more heterogeneity
- Performance was highly dependent on smoothing factor (LaPlace) and the maximum proportion of insertions per column for match states





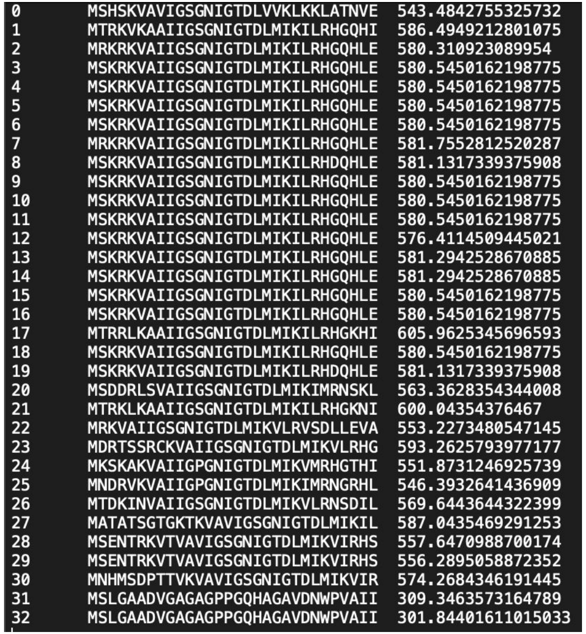
Results: online implementation

- A profile HMM was trained using **acetaldehyde** dehydrogenase (ACDH) MSAs
- Log odds scores of the hold out genes are displayed



```
0       MATAGKVIKCKAAVAWEAGKPLSIEEVEVA  -191.74138103529987
1       MATAGKVIKCKAAVAWEAGKPLSIEEVEVA  -200.24654603683547
2       MSATEGKVITCKAAVAWEAKKPLVIEDIEV  -184.21142449837228
3       TTEGKVIKCKAAIAWEAGKPLSVEEIEVSP  -185.36134780220146
4       MATAGKVIKCKAAVAWEAGKPLSIEEVEVA  -200.57626981995304
5       MATAGKVIKCKAAVAWEAAKPLVIEEVEVA  -194.90953334809186
6       MATAGKVIKCKAAVAWEAGKPLSIEEVEVA  -200.57626981995304
7       MSTEGKVITCKAAVAWEAKKPLVIEDIEVQ  -200.06522526637818
8       MATAGKVIKCKAAVAWEAAKPLVIEEVEVA  -194.90953334809186
9       MATQGQVITCKAAVAYEPNKPLVIEDVQVA  -188.65981017479942
10      MGTKGKVIKCKAAIAWEAGKPLCIEEVEVA  -187.2499382349565
11      MSNTVGQIIKCRAAVAWEAGKPLVIEEVEV  -191.52278618887982
12      MSNTAGQVIKCRAAVAWEAGKPLVIEEVEV  -196.17603334650343
13      MSTTVGQVIRCKAAVAWEAGKPLVMEEVDV  -185.069118573035
14      MSTTTGQVIRCKAAVAWEAGKPLVMEEVDV  -189.7213136573164
15      MSTTVGQVIRCKAAVAWEAGKPLVMEEVDV  -189.7213136573164
16      MSSTEGKVICCRAAVAWEAGKPLVIEEVEV  -195.5461755532964
17      MSTTVGQVIRCKAAVAWEAGKPLVMEEVDV  -184.7099548880505
18      MSNTAGQVIRCRAAVAWEAGKPLVIEEVEV  -193.80392553130469
19      METQGKVITCKAAVAWGAGEPLVMEDVKVD  -184.59546228816103
20      MASSTQGQVITCKAAVAWEANKPMTIEDVQ  -180.84736008861822
21      MASPTQGQVITCKAAVAYEPNKPLVIEDVQ  -186.0073224298608
22      MASSTQGQVITCKAAVAWEANRPMTIEDVQ  -180.84736008861822
23      MSASSSSFEQPQVITCNAAVAWRAGEPLVM  -195.67361678704546
24      MSSNTAGQVIRCKAAVAWEAGKPLVIEEVE  -192.1800374588703
25      MSSTAGQVINCKAAVAWSAKAPLSIETIQV  -186.96653470817841
26      MFAEIQIQDKDRMGTAGKVIKCKAAVLWEQ  -191.29418603145658
27      MDKASITEGKPIRCKAAILRKAGEPLVIEE  -182.3293992774402
28      MDKTFFSSNEGKPITCKAAICRKAGEALVI  -182.25215390712972
29      MEKTSMFSIHEGKPIRCKAAVSRKPGEALV  -180.0608125715245
30      MENGNSSSDNKSSHKPIRCKAAVSRKAGEP  -196.33006768374068
31      MAETQGKVITCKAAVVWGPKVPLVIQEICV  -178.81120103748324
```

ADH Test

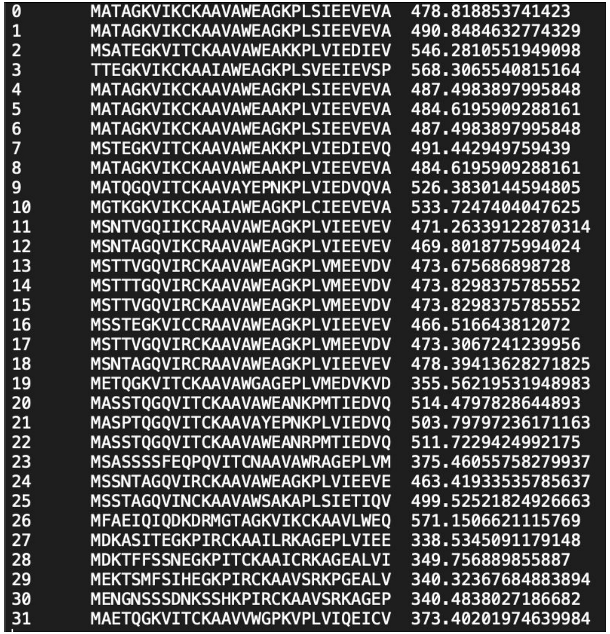


```
0       MSHSKVAVIGSGNIGTDLVVKLKKLATNVE  543.4842755325732
1       MTRKVKAAIIGSGNIGTDLMIKILRHGQHI  586.4949212801075
2       MRKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.310923089954
3       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
4       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
5       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
6       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
7       MRKRKVAIIGSGNIGTDLMIKILRHGQHLE  581.7552812520287
8       MSKRKVAIIGSGNIGTDLMIKILRHDQHLE  581.1317339375908
9       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
10      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
11      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
12      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  576.4114509445021
13      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  581.2942528670885
14      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  581.2942528670885
15      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
16      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
17      MTRRLKAAIIGSGNIGTDLMIKILRHGKHI  605.9625345696593
18      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  580.5450162198775
19      MSKRKVAIIGSGNIGTDLMIKILRHDQHLE  581.1317339375908
20      MSDDRLSVAIIGSGNIGTDLMIKIMRNSKL  563.3628354344008
21      MTRKLKAAIIGSGNIGTDLMIKILRHGKNI  600.04354376467
22      MRKVAIIGSGNIGTDLMIKVLRVSDLLEVA  553.2273480547145
23      MDRTSSRCKVAIIGSGNIGTDLMIKVLRHG  593.2625793977177
24      MKSKAKVAIIGPGNIGTDLMIKVMRHGTHI  551.8731246925739
25      MNDRVKVAIIGPGNIGTDLMIKIMRNGRHL  546.3932641436909
26      MTDKINVAIIGSGNIGTDLMIKVLRNSDIL  569.6443644322399
27      MATATSGTGKTKVAVIGSGNIGTDLMIKIL  587.0435469291253
28      MSENTRKVTVAVIGSGNIGTDLMIKVIRHS  557.6470988700174
29      MSENTRKVTVAVIGSGNIGTDLMIKVIRHS  556.2895058872352
30      MNHMSDPTTVKVAVIGSGNIGTDLMIKVIR  574.2684346191445
31      MSLGAADVGAGAGPPGQHAGAVDNWPVAII  309.3463573164789
32      MSLGAADVGAGAGPPGQHAGAVDNWPVAII  301.84401611015033
```

ACDH Test

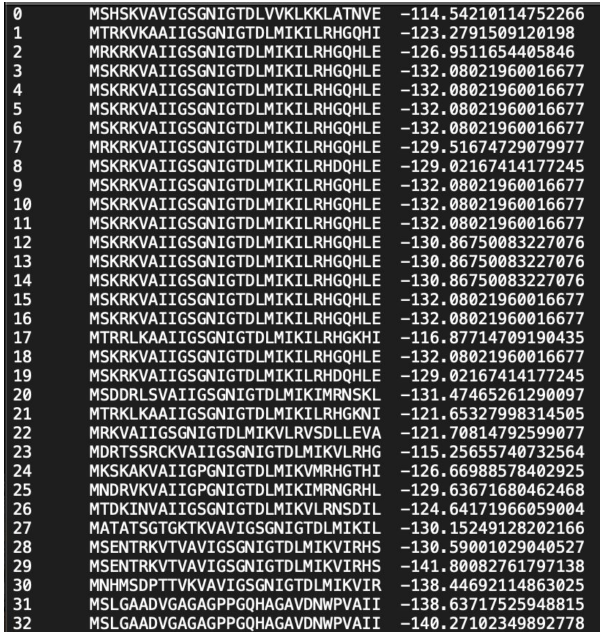
Results: online implementation

- A profile HMM was trained using **alcohol** dehydrogenase (ADH) MSAs
- Log odds scores of the hold out genes are displayed



```
0       MATAGKVIKCKAAVAWEAGKPLSIEEVEVA  478.818853741423
1       MATAGKVIKCKAAVAWEAGKPLSIEEVEVA  490.8484632774329
2       MSATEGKVITCKAAVAWEAKKPLVIEDIEV  546.2810551949098
3       TTEGKVIKCKAAIAWEAGKPLSVEEIEVSP  568.3065540815164
4       MATAGKVIKCKAAVAWEAGKPLSIEEVEVA  487.4983897995848
5       MATAGKVIKCKAAVAWEAAKPLVIEEVEVA  484.6195909288161
6       MATAGKVIKCKAAVAWEAGKPLSIEEVEVA  487.4983897995848
7       MSTEGKVITCKAAVAWEAKKPLVIEDIEVQ  491.442949759439
8       MATAGKVIKCKAAVAWEAAKPLVIEEVEVA  484.6195909288161
9       MATQGQVITCKAAVAYEPNKPLVIEDVQVA  526.3830144594805
10      MGTKGKVIKCKAAIAWEAGKPLCIEEVEVA  533.7247404047625
11      MSNTVGQIIKCRAAVAWEAGKPLVIEEVEV  471.26339122870314
12      MSNTAGQVIKCRAAVAWEAGKPLVIEEVEV  469.8018775994024
13      MSTTVGQVIRCKAAVAWEAGKPLVMEEVDV  473.675686898728
14      MSTTTGQVIRCKAAVAWEAGKPLVMEEVDV  473.8298375785552
15      MSTTVGQVIRCKAAVAWEAGKPLVMEEVDV  473.8298375785552
16      MSSTEGKVICCRAAVAWEAGKPLVIEEVEV  466.516643812072
17      MSTTVGQVIRCKAAVAWEAGKPLVMEEVDV  473.3067241239956
18      MSNTAGQVIRCRAAVAWEAGKPLVIEEVEV  478.39413628271825
19      METQGKVITCKAAVAWGAGEPLVMEDVKVD  355.56219531948983
20      MASSTQGQVITCKAAVAWEANKPMTIEDVQ  514.4797828644893
21      MASPTQGQVITCKAAVAYEPNKPLVIEDVQ  503.79797236171163
22      MASSTQGQVITCKAAVAWEANRPMTIEDVQ  511.7229424992175
23      MSASSSSFEQPQVITCNAAVAWRAGEPLVM  375.46055758279937
24      MSSNTAGQVIRCKAAVAWEAGKPLVIEEVE  463.41933535785637
25      MSSTAGQVINCKAAVAWSAKAPLSIETIQV  499.52521824926663
26      MFAEIQIQDKDRMGTAGKVIKCKAAVLWEQ  571.1506621115769
27      MDKASITEGKPIRCKAAILRKAGEPLVIEE  338.5345091179148
28      MDKTFFSSNEGKPITCKAAICRKAGEALVI  349.756889855887
29      MEKTSMFSIHEGKPIRCKAAVSRKPGEALV  340.32367684883894
30      MENGNSSSDNKSSHKPIRCKAAVSRKAGEP  340.4838027186682
31      MAETQGKVITCKAAVVWGPKVPLVIQEICV  373.40201974639984
```

ADH Test



```
0       MSHSKVAVIGSGNIGTDLVVKLKKLATNVE  −114.54210114752266
1       MTRKVKAAIIGSGNIGTDLMIKILRHGQHI  −123.2791509120198
2       MRKRKVAIIGSGNIGTDLMIKILRHGQHLE  −126.9511654405846
3       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
4       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
5       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
6       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
7       MRKRKVAIIGSGNIGTDLMIKILRHGQHLE  −129.51674729079977
8       MSKRKVAIIGSGNIGTDLMIKILRHDQHLE  −129.02167414177245
9       MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
10      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
11      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
12      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −130.86750083227076
13      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −130.86750083227076
14      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −130.86750083227076
15      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
16      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
17      MTRRLKAAIIGSGNIGTDLMIKILRHGKHI  −116.87714709190435
18      MSKRKVAIIGSGNIGTDLMIKILRHGQHLE  −132.08021960016677
19      MSKRKVAIIGSGNIGTDLMIKILRHDQHLE  −129.02167414177245
20      MSDDRLSVAIIGSGNIGTDLMIKIMRNSKL  −131.47465261290097
21      MTRKLKAAIIGSGNIGTDLMIKILRHGKNI  −121.65327998314505
22      MRKVAIIGSGNIGTDLMIKVLRVSDLLEVA  −121.70814792599077
23      MDRTSSRCKVAIIGSGNIGTDLMIKVLRHG  −115.25655740732564
24      MKSKAKVAIIGPGNIGTDLMIKVMRHGTHI  −126.66988578402925
25      MNDRVKVAIIGPGNIGTDLMIKIMRNGRHL  −129.63671680462468
26      MTDKINVAIIGSGNIGTDLMIKVLRNSDIL  −124.64171966059004
27      MATATSGTGKTKVAVIGSGNIGTDLMIKIL  −130.15249128202166
28      MSENTRKVTVAVIGSGNIGTDLMIKVIRHS  −130.59001029040527
29      MSENTRKVTVAVIGSGNIGTDLMIKVIRHS  −141.80082761797138
30      MNHMSDPTTVKVAVIGSGNIGTDLMIKVIR  −138.44692114863025
31      MSLGAADVGAGAGPPGQHAGAVDNWPVAII  −138.63717525948815
32      MSLGAADVGAGAGPPGQHAGAVDNWPVAII  −140.27102349892778
```

ACDH Test

Questions?

Our implementation: https://github.com/danielchang2002/5481_final