GenerateSH2sV5

November 22, 2024

This is a bash notebook for generating SH2 ancestral reconstructed sequences.
Version 4 makes a change to the 3rd pipeline which is that I'm going to use both the longer and shorter (gapless) reconstructed sequences.
Version 5 updates the markdown. This is the final version.

TJE 2022 03 28 (V4) TJE 2024 11 22 (V5)

0.0.1 The main workflow:

1. Start with the PFAM database in full for SH2 domains, ~23,100 deduplicated seqs.

2. Filter this for Human sequences only.
3. Compute the pairwise sequence identity for Btk vs. these human sequences.
4. Filter the sequences for a pairwise identity with Btk greater than 25% or greater than 50% (tecs).
5. Search for ~20 sequences with high sequence identity to the tecs. Note:
    a. Any sequence with only 1 aa difference from Btk is removed.
    b. I skip every other sequence when taking the 20, ordered by seq ID with BTK. So they span a larger sequence range.
6. Perform ancestral sequence reconstruction on these tecs
    a. include reconstructed sequences from both gapless and with-gap method, but only from method #1
7. Align the reconstructed sequences, tecs, and additional human seqs to the PFAM database
8. Remove any seq over 90 aa (because of ordering issues, they can't be that long). Remove extra gaps
9. Add in control sequences
10. Create nucleotide seqs

[1]:

```
#step 1.
python ../source/DeduplicateFasta.py ../sh2_pfam_full/PF00017_full.txt␣
 ↪PF00017_full_dedup.fasta
wc -l PF00017_full_dedup.fasta
```

```
231210 PF00017_full_dedup.fasta
```

[2]:

```
#step 2.
ggrep -A 9 --no-group-separator HUMAN PF00017_full_dedup.fasta > PF00017_HUMAN.
 ↪txt
wc -l PF00017_HUMAN.txt
```

```
    1550 PF00017_HUMAN.txt
```

[3]:

```
#step 3.
python ../source/Compare_to_BTK_FullIds.py PF00017_HUMAN.txt PF00017_HUMAN_IDs.
 ↪txt␣
 ↪"---------------------------------------------WYS-K--H-M---T--R---------------------SQ--A-E
wc -l PF00017_HUMAN_IDs.txt
```

```
     156 PF00017_HUMAN_IDs.txt
```

[4]:

```
#step 4.
#note that there are most likely some duplicates in this, see U3NG26 with looks␣
 ↪like another name for human BTK
#update that most of these duplicates should be removed.
#I'm not sure whether to removed Q5JY90_HUMAN, which is a splice isoform of BTK.
 ↪
awk 'BEGIN{FS=OFS="\t"}($2 >= 0.25){print $0}' PF00017_HUMAN_IDs.txt >␣
 ↪PF00017_HUMAN_IDs_filtered_25.txt
wc -l PF00017_HUMAN_IDs_filtered_25.txt
sort -rn -k 2,2 PF00017_HUMAN_IDs_filtered_25.txt

awk 'BEGIN{FS=OFS="\t"}($2 >= 0.50){print $0}' PF00017_HUMAN_IDs.txt >␣
 ↪PF00017_HUMAN_IDs_filtered_50.txt
wc -l PF00017_HUMAN_IDs_filtered_50.txt
sort -rn -k 2,2 PF00017_HUMAN_IDs_filtered_50.txt
```

```
      67 PF00017_HUMAN_IDs_filtered_25.txt
BTK_HUMAN/281-362       1.0
Q5JY90_HUMAN/281-355    0.8170731707317073
TEC_HUMAN/247-330       0.573170731707317
ITK_HUMAN/239-323       0.524390243902439
TXK_HUMAN/150-231       0.5121951219512195
BMX_HUMAN/296-377       0.5
FER_HUMAN/460-531       0.34146341463414637
SRMS_HUMAN/120-197      0.32926829268292684
A0A0A0MRF9_HUMAN/532-617        0.32926829268292684
NCK1_HUMAN/282-356      0.3170731707317073
F6TDL0_HUMAN/404-478    0.3170731707317073
NCK2_HUMAN/285-359      0.3048780487804878
LCP2_HUMAN/422-505      0.3048780487804878
DAPP1_HUMAN/35-109      0.3048780487804878
ABL2_HUMAN/173-248      0.3048780487804878
SLAP1_HUMAN/84-160      0.2926829268292683
PTN11_HUMAN/6-81        0.2926829268292683
PTK6_HUMAN/78-155       0.2926829268292683
P85A_HUMAN/624-698      0.2926829268292683
LYN_HUMAN/129-211       0.2926829268292683
I3L297_HUMAN/13-104     0.2926829268292683
```

```
E5RJ69_HUMAN/84-158     0.2926829268292683
CRKL_HUMAN/14-88        0.2926829268292683
YES_HUMAN/158-240       0.2804878048780488
SRC_HUMAN/151-233       0.2804878048780488
SHIP2_HUMAN/21-102      0.2804878048780488
SHE_HUMAN/395-471       0.2804878048780488
SHC2_HUMAN/487-558      0.2804878048780488
SHB_HUMAN/410-485       0.2804878048780488
SH2D7_HUMAN/51-126      0.2804878048780488
SH21B_HUMAN/5-86        0.2804878048780488
RASA1_HUMAN/181-256     0.2804878048780488
PLCG1_HUMAN/668-741     0.2804878048780488
H0Y3C5_HUMAN/143-225    0.2804878048780488
FRK_HUMAN/116-193       0.2804878048780488
FES_HUMAN/460-530       0.2804878048780488
E9PF55_HUMAN/1450-1543  0.2804878048780488
ABL1_HUMAN/127-202      0.2804878048780488
A0A494C067_HUMAN/1555-1649      0.2804878048780488
TNS2_HUMAN/1140-1232    0.2682926829268293
TENS3_HUMAN/1172-1267   0.2682926829268293
SH23A_HUMAN/15-95       0.2682926829268293
SH21A_HUMAN/6-87        0.2682926829268293
PTN6_HUMAN/110-194      0.2682926829268293
GRAP_HUMAN/60-135       0.2682926829268293
F5H1Z8_HUMAN/4-79       0.2682926829268293
CSK_HUMAN/82-156        0.2682926829268293
BLNK_HUMAN/346-429      0.2682926829268293
ZAP70_HUMAN/163-239     0.25609756097560976
SLAP2_HUMAN/94-176      0.25609756097560976
SHIP1_HUMAN/5-86        0.25609756097560976
SHD_HUMAN/240-316       0.25609756097560976
SHC3_HUMAN/499-570      0.25609756097560976
SHC1_HUMAN/488-559      0.25609756097560976
SH2D3_HUMAN/220-300     0.25609756097560976
P85B_HUMAN/622-696      0.25609756097560976
H3BU69_HUMAN/3-68       0.25609756097560976
GRAP2_HUMAN/58-132      0.25609756097560976
F8W6V4_HUMAN/139-215    0.25609756097560976
F8VU91_HUMAN/48-129     0.25609756097560976
F5H5M1_HUMAN/240-311    0.25609756097560976
F5GY79_HUMAN/69-144     0.25609756097560976
E9PJX5_HUMAN/53-134     0.25609756097560976
E9PAP0_HUMAN/171-253    0.25609756097560976
BCAR3_HUMAN/154-234     0.25609756097560976
A0A2R8Y5Q0_HUMAN/327-402        0.25609756097560976
Seq_ID  pairwise_identity
       7 PF00017_HUMAN_IDs_filtered_50.txt
BTK_HUMAN/281-362       1.0
```

```
Q5JY90_HUMAN/281-355    0.8170731707317073
TEC_HUMAN/247-330       0.573170731707317
ITK_HUMAN/239-323       0.524390243902439
TXK_HUMAN/150-231       0.5121951219512195
BMX_HUMAN/296-377       0.5
Seq_ID  pairwise_identity
```

[5]:
```
#step 4, continued
tail -n +2 PF00017_HUMAN_IDs_filtered_50.txt | cut -f 1 | ggrep -A 9␣
 ↪--no-group-separator -f - PF00017_HUMAN.txt > PF00017_HUMAN_filtered_50.txt
wc -l PF00017_HUMAN_filtered_50.txt
```

```
      60 PF00017_HUMAN_filtered_50.txt
```

[319]:
```
#new step 5., 2022 03 16
python ../source/ChooseSimilarSeq.py PF00017_HUMAN_filtered_50.txt␣
 ↪PF00017_full_dedup.fasta PF00017_HUMAN_WithSimilarSeq_50.txt 14
```

```
seqs processed:
10000
seqs processed:
20000
```

[320]:
```
#step 5, construct the fasta file and check sequence numbers.
#REMOVE PROBLEMATIC SEQUENCE HERE.
tail -n +2 PF00017_HUMAN_WithSimilarSeq_50.txt | cut -f 1-2 | awk␣
 ↪'BEGIN{FS="\t";OFS=""}{print $1, "\n", $2}' | sort | uniq | grep -v␣
 ↪'A0A452S617_URSAM/281-366' | wc -l
tail -n +2 PF00017_HUMAN_WithSimilarSeq_50.txt | cut -f 1-2 | awk␣
 ↪'BEGIN{FS="\t";OFS=""}{print $1, "\n", $2}' | sort | uniq | grep -v␣
 ↪'A0A452S617_URSAM/281-366' | ggrep -A 9 --no-group-separator -f -␣
 ↪PF00017_full_dedup.fasta > PF00017_HUMAN_WithSimilarSeq_50.fasta
# python ../source/FilterSeqsByLength.py PF00017_HUMAN_WithSimilarSeq_50.fasta␣
 ↪PF00017_HUMAN_WithSimilarSeq_50_Short.fasta 88
grep -c ">" PF00017_HUMAN_WithSimilarSeq_50.fasta
```

```
      89
89
```

[321]:
```
#rerun the pairwise analysis.
python ../source/Compare_to_BTK_FullIds.py PF00017_HUMAN_WithSimilarSeq_50.
 ↪fasta PF00017_HUMAN_WithSimilarSeq_IDs_50.txt␣
 ↪"---------------------------------------------WYS-K--H-M---T--R---------------------SQ--A-E
wc -l PF00017_HUMAN_WithSimilarSeq_IDs_50.txt
```

```
      90 PF00017_HUMAN_WithSimilarSeq_IDs_50.txt
```

[322]:
```
#step 5., continued
```

```
#note that paml doesn't allow names more than 30 characters, so I need to check␣
 ↪my name length:
cut -f 1 PF00017_HUMAN_WithSimilarSeq_50.txt | awk 'BEGIN{FS=OFS=""}{print NF}'␣
 ↪| sort | uniq | sort -rn
```

```
20
17
5
```

[323]:
```
#step 5.
rm -r PF00017_HUMAN_WithSimilarSeq_Reconstruction/ #so that I can run this again
python /Users/timeisen/Applications/Pasta/pasta/run_pasta.py -a -i␣
 ↪PF00017_HUMAN_WithSimilarSeq_50.fasta -d protein -j PreReconstruct␣
 ↪--temporaries ./ -o PF00017_HUMAN_WithSimilarSeq_Reconstruction/
```

```
PASTA INFO: Reading input sequences from
'PF00017_HUMAN_WithSimilarSeq_50.fasta'…
PASTA INFO: Masking alignment sites with less than 7 sites before running the
tree step
PASTA INFO: Configuration written to "/Users/timeisen/Dropbox (Personal)/Kuriyan
Lab/Sequences/AncestralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reco
nstruction_4/PF00017_HUMAN_WithSimilarSeq_Reconstruction/PreReconstruct_temp_pas
ta_config.txt".

PASTA INFO: Directory for temporary files created at /Users/timeisen/Dropbox (Pe
rsonal)/KuriyanLab/Sequences/AncestralSequenceReconstruction/SH2_Domain_Reconstr
uction/sh2_reconstruction_4/PreReconstruct/tempbpbr0cj6
PASTA INFO: Name translation information saved to /Users/timeisen/Dropbox (Perso
nal)/KuriyanLab/Sequences/AncestralSequenceReconstruction/SH2_Domain_Reconstruct
ion/sh2_reconstruction_4/PF00017_HUMAN_WithSimilarSeq_Reconstruction/PreReconstr
uct_temp_name_translation.txt as safe name, original name, blank line format.
PASTA INFO: Creating a starting tree for the PASTA algorithm…
PASTA INFO: Input sequences assumed to be aligned (based on sequence lengths).
PASTA INFO: Performing initial tree search to get starting tree…
PASTA INFO: Starting PASTA algorithm on initial tree…
PASTA INFO: Max subproblem set to 45
PASTA INFO: Step 0. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 0. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted despite the score not improving.
PASTA INFO: current score: -1904.161, best score: -1904.161
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Step 1. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 1. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted despite the score not improving.
PASTA INFO: current score: -1911.192, best score: -1904.161
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Step 2. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 2. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted despite the score not improving.
PASTA INFO: current score: -1911.195, best score: -1904.161
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Writing resulting alignment to PF00017_HUMAN_WithSimilarSeq_Reconstr
uction/PreReconstruct.marker001.PF00017_HUMAN_WithSimilarSeq_50.aln
PASTA INFO: Writing resulting tree to
PF00017_HUMAN_WithSimilarSeq_Reconstruction/PreReconstruct.tre
PASTA INFO: Writing resulting likelihood score to
PF00017_HUMAN_WithSimilarSeq_Reconstruction/PreReconstruct.score.txt
PASTA INFO: The resulting alignment (with the names in a "safe" form) was first
written as the file "/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Anc
estralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PF00
017_HUMAN_WithSimilarSeq_Reconstruction/PreReconstruct_temp_iteration_2_seq_alig
nment.txt"
PASTA INFO: The resulting tree (with the names in a "safe" form) was first
written as the file "/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Anc
estralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PF00
017_HUMAN_WithSimilarSeq_Reconstruction/PreReconstruct_temp_iteration_2_tree.tre
"
Refused to clean '/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Ancest
ralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PreReco
nstruct/tempbpbr0cj6/step2/mincluster/pw/tempopal3v5bj7o_': not created by PASTA
'/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/AncestralSequenceRecons
truction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PreReconstruct/tempbpbr0
cj6/step2/mincluster/pw/tempopal3v5bj7o_' is not registered as a temporary
directory that was created by this process!
Refused to clean '/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Ancest
ralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PreReco
nstruct/tempbpbr0cj6/step0/mincluster/pw/tempopal44if_er9': not created by PASTA
'/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/AncestralSequenceRecons
truction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PreReconstruct/tempbpbr0
cj6/step0/mincluster/pw/tempopal44if_er9' is not registered as a temporary
directory that was created by this process!
Refused to clean '/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Ancest
ralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PreReco
nstruct/tempbpbr0cj6/step1/mincluster/pw/tempopalih6xhdb1': not created by PASTA
'/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/AncestralSequenceRecons
truction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PreReconstruct/tempbpbr0
cj6/step1/mincluster/pw/tempopalih6xhdb1' is not registered as a temporary
directory that was created by this process!
PASTA INFO: Total time spent: 9.835088729858398s
```

[324]:

```
#step 6., continued
#at this point I had to go into the figtree program and change the format of␣
 ↪the tree from nexus to the newick format.
#then I had to change the codeml.ctl to point to the correct files
#then I had to copy the jones.dat file to this directory
```

```
#note that this codeml has been changed to use initial branch lengths and leave␣
 ↪the ambiguity.
cd PF00017_HUMAN_WithSimilarSeq_Reconstruction
cp /Users/timeisen/Dropbox\ \(Personal\)/KuriyanLab/Sequences/
 ↪AncestralSequenceReconstruction/SH2_Domain_Reconstruction/codeml.ctl ./
 ↪codeml_pre.ctl
cp /Users/timeisen/Dropbox\ \(Personal\)/KuriyanLab/Sequences/
 ↪AncestralSequenceReconstruction/SH2_Domain_Reconstruction/
 ↪sh2_reconstruction_2/PF00017_HUMAN_WithSimilarSeq_Reconstruction/jones.dat ./
sed "s/'//g" PreReconstruct.newick > PreReconstruct.newick_cleaned
sed 's/noisy = 9/noisy = 1/' codeml_pre.ctl | sed 's/cleandata = 1/cleandata =␣
 ↪0/' | sed 's/fix_blength = -1/fix_blength = 1/' > codeml.ctl
head codeml.ctl
head jones.dat
```

```
      seqfile = PreReconstruct.marker001.PF00017_HUMAN_WithSimilarSeq_50.aln *
sequence data filename
     treefile = PreReconstruct.newick_cleaned      * tree structure file name
      outfile = mlc           * main result file name

        noisy = 1  * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1  * 0: concise; 1: detailed, 2: too much
      runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 2  * 1:codons; 2:AAs; 3:codons-->AAs

 58
 54  45
 81  16 528
 56 113  34  10
 57 310  86  49   9
105  29  58 767   5 323
179 137  81 130  59  26 119
 27 328 391 112  69 597  26  23
 36  22  47  11  17   9  12   6  16
```

[325]:
```
pwd
codeml
```

```
/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/AncestralSequenceReconst
ruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/PF00017_HUMAN_WithSimilar
Seq_Reconstruction

AAML in paml version 4.8a, August 2014

processing fasta file
```

```
reading seq# 1 BTK_HUMAN/281-362                                       88 sites
reading seq# 2 I3LN58_PIG/281-362                                      88 sites
reading seq# 3 A0A2Y9FBB9_PHYMC/281-362                                88 sites
reading seq# 4 A0A5N4C153_CAMDR/307-388                                88 sites
reading seq# 5 F7BKT0_ORNAN/318-399                                    88 sites
reading seq# 6 A0A340Y2U3_LIPVE/281-355                                88 sites
reading seq# 7 F1S1L0_PIG/281-355                                      88 sites
reading seq# 8 A0A2K6TV44_SAIBB/265-338                                88 sites
reading seq# 9 M3WMI5_FELCA/280-354                                    88 sites
reading seq#10 Q5JY90_HUMAN/281-355                                    88 sites
reading seq#11 A0A671DKT5_RHIFE/280-354                                88 sites
reading seq#12 A0A6I8NCY8_ORNAN/279-353                                88 sites
reading seq#13 A0A094NE02_ANTCR/235-317                                88 sites
reading seq#14 A0A093CY13_TAUER/241-323                                88 sites
reading seq#15 A0A226MPG1_CALSU/83-165                                 88 sites
reading seq#16 A0A091QH50_MERNU/256-338                                88 sites
reading seq#17 A0A087RFM3_APTFO/241-323                                88 sites
reading seq#18 A0A6J2H0P0_9PASS/360-442                                88 sites
reading seq#19 A0A2K6EYQ1_PROCO/241-324                                88 sites
reading seq#20 A0A2Y9JLY6_ENHLU/268-351                                88 sites
reading seq#21 A0A2Y9I079_NEOSC/246-329                                88 sites
reading seq#22 Q8CFK4_MOUSE/224-307                                    88 sites
reading seq#23 A0A6J0DL91_PERMB/246-329                                88 sites
reading seq#24 H0VD88_CAVPO/225-308                                    88 sites
reading seq#25 G5B8D6_HETGA/248-331                                    88 sites
reading seq#26 L9L9I7_TUPCH/246-329                                    88 sites
reading seq#27 A0A673TJP9_SURSU/247-330                                88 sites
reading seq#28 F1SEA4_PIG/248-331                                      88 sites
reading seq#29 A0A5N3VZ71_MUNRE/239-322                                88 sites
reading seq#30 G1STI6_RABIT/247-330                                    88 sites
reading seq#31 A0A6P6CKB3_PTEVA/247-330                                88 sites
reading seq#32 A0A1U7U4B3_CARSF/247-330                                88 sites
reading seq#33 TEC_HUMAN/247-330                                       88 sites
reading seq#34 R4GB22_ANOCA/281-363                                    88 sites
reading seq#35 K7F777_PELSI/277-359                                    88 sites
reading seq#36 A0A6P5IJN4_PHACI/297-378                                88 sites
reading seq#37 F6YES1_MONDO/281-362                                    88 sites
reading seq#38 F7IHG0_CALJA/281-362                                    88 sites
reading seq#39 A0A7E6D215_9CHIR/309-390                                88 sites
reading seq#40 A0A671DKQ7_RHIFE/280-361                                88 sites
reading seq#41 A0A6P3RL81_PTEVA/281-362                                88 sites
reading seq#42 A0A6J0E543_PERMB/298-379                                88 sites
reading seq#43 H0XW86_OTOGA/284-365                                    88 sites
reading seq#44 A0A1S3GNU6_DIPOR/281-362                                88 sites
reading seq#45 A0A2K6SA55_SAIBB/296-377                                88 sites
reading seq#46 A0A3Q2HYV1_HORSE/281-362                                88 sites
reading seq#47 A0A673V1S0_SURSU/273-354                                88 sites
reading seq#48 G1MBR3_AILME/276-357                                    88 sites
reading seq#49 A0A2Y9GG10_NEOSC/276-357                                88 sites
reading seq#50 A0A384CIC9_URSMA/276-357                                88 sites
reading seq#51 A0A2Y9KLX9_ENHLU/274-355                                88 sites
reading seq#52 M3XU75_MUSPF/276-357                                    88 sites
reading seq#53 BMX_HUMAN/296-377                                       88 sites
reading seq#54 A0A1U7TS48_CARSF/280-361                                88 sites
reading seq#55 A0A6P3QJB8_PTEVA/273-354                                88 sites
reading seq#56 A0A2Y9E4Q9_TRIMA/276-357                                88 sites
reading seq#57 A0A2U4AJT4_TURTR/276-357                                88 sites
reading seq#58 A0A5N4C2L1_CAMDR/288-369                                88 sites
reading seq#59 A0A340Y4G0_LIPVE/272-353                                88 sites
reading seq#60 A0A1U7U634_CARSF/148-229                                88 sites
reading seq#61 A0A5N4EHK4_CAMDR/126-207                                88 sites
reading seq#62 A0A6P3QP21_PTEVA/150-231                                88 sites
reading seq#63 W5Q5S4_SHEEP/150-231                                    88 sites
reading seq#64 A0A6P3J1L6_BISBI/150-231                                88 sites
reading seq#65 A0A673TTW7_SURSU/174-255                                88 sites
reading seq#66 A0A2Y9JM55_ENHLU/174-255                                88 sites
reading seq#67 E2RBA0_CANLF/174-255                                    88 sites
reading seq#68 A0A485N2L8_LYNPA/150-231                                88 sites
reading seq#69 A0A2Y9RMH4_TRIMA/152-233                                88 sites
reading seq#70 TXK_HUMAN/150-231                                       88 sites
reading seq#71 A0A2K5PHS5_CEBIM/150-231                                88 sites
reading seq#72 A0A096NDG3_PAPAN/150-231                                88 sites
reading seq#73 A0A5F7ZHT0_MACMU/150-231                                88 sites
reading seq#74 A0A2K5E184_AOTNA/150-231                                88 sites
reading seq#75 F7FTR6_MONDO/239-323                                    88 sites
reading seq#76 A0A2K6AVK0_MACNE/239-323                                88 sites
reading seq#77 G3UBJ2_LOXAF/247-331                                    88 sites
reading seq#78 ITK_HUMAN/239-323                                       88 sites
reading seq#79 A0A452SNA4_URSAM/250-331                                88 sites
reading seq#80 G1LHK5_AILME/239-323                                    88 sites
reading seq#81 A0A286ZPK2_PIG/217-301                                  88 sites
reading seq#82 S7N0E0_MYOBR/242-326                                    88 sites
reading seq#83 I3MD63_ICTTR/239-323                                    88 sites
reading seq#84 L5L1K8_PTEAL/239-323                                    88 sites
reading seq#85 A0A6P6HUD7_PUMCO/239-323                                88 sites
reading seq#86 A0A671FJF7_RHIFE/239-323                                88 sites
reading seq#87 A0A452GAH1_CAPHI/239-323                                88 sites
reading seq#88 W5PNG3_SHEEP/239-323                                    88 sites
reading seq#89 D4A7W7_RAT/239-323                                      88 sites
ns = 89         ls = 88
Reading sequences, sequential format..
Counting site patterns..  0:00
          84 patterns at       88 /       88 sites (100.0%),  0:00
Counting frequencies..

    31328 bytes for distance
    26880 bytes for conP
        0 bytes for fhK
  5000000 bytes for space

TREE #  1

  1169280 bytes for conP, adjusted

1 node(s) used for scaling (Yang 2000 J Mol Evol 51:423-432):
 112

ntime & nrate & np:   175     0   175

np =   175
lnL0 = -2538.774206
Out..
lnL  = -1954.145209
46520 lfun, 0 eigenQcodon, 8141000 P(t)

Reconstructed ancestral states go into file rst.

lnL = -1954.145209 from ProbSitePattern.
Marginal reconstruction.
        Node  90: lnL = -1954.145209
        Node  91: lnL = -1954.145209
        Node  92: lnL = -1954.145209
        Node  93: lnL = -1954.145209
        Node  94: lnL = -1954.145209
        Node  95: lnL = -1954.145209
        Node  96: lnL = -1954.145209
        Node  97: lnL = -1954.145209
        Node  98: lnL = -1954.145209
        Node  99: lnL = -1954.145209
        Node 100: lnL = -1954.145209
        Node 101: lnL = -1954.145209
        Node 102: lnL = -1954.145209
        Node 103: lnL = -1954.145209
        Node 104: lnL = -1954.145209
        Node 105: lnL = -1954.145209
        Node 106: lnL = -1954.145209
        Node 107: lnL = -1954.145209
        Node 108: lnL = -1954.145209
        Node 109: lnL = -1954.145209
        Node 110: lnL = -1954.145209
        Node 111: lnL = -1954.145209
        Node 112: lnL = -1954.145209
        Node 113: lnL = -1954.145209
        Node 114: lnL = -1954.145209
        Node 115: lnL = -1954.145209
        Node 116: lnL = -1954.145209
        Node 117: lnL = -1954.145209
        Node 118: lnL = -1954.145209
        Node 119: lnL = -1954.145209
        Node 120: lnL = -1954.145209
        Node 121: lnL = -1954.145209
        Node 122: lnL = -1954.145209
        Node 123: lnL = -1954.145209
        Node 124: lnL = -1954.145209
        Node 125: lnL = -1954.145209
        Node 126: lnL = -1954.145209
        Node 127: lnL = -1954.145209
        Node 128: lnL = -1954.145209
        Node 129: lnL = -1954.145209
        Node 130: lnL = -1954.145209
        Node 131: lnL = -1954.145209
        Node 132: lnL = -1954.145209
        Node 133: lnL = -1954.145209
        Node 134: lnL = -1954.145209
        Node 135: lnL = -1954.145209
        Node 136: lnL = -1954.145209
        Node 137: lnL = -1954.145209
        Node 138: lnL = -1954.145209
        Node 139: lnL = -1954.145209
        Node 140: lnL = -1954.145209
        Node 141: lnL = -1954.145209
        Node 142: lnL = -1954.145209
        Node 143: lnL = -1954.145209
        Node 144: lnL = -1954.145209
        Node 145: lnL = -1954.145209
        Node 146: lnL = -1954.145209
        Node 147: lnL = -1954.145209
        Node 148: lnL = -1954.145209
        Node 149: lnL = -1954.145209
        Node 150: lnL = -1954.145209
        Node 151: lnL = -1954.145209
        Node 152: lnL = -1954.145209
        Node 153: lnL = -1954.145209
        Node 154: lnL = -1954.145209
        Node 155: lnL = -1954.145209
        Node 156: lnL = -1954.145209
        Node 157: lnL = -1954.145209
        Node 158: lnL = -1954.145209
        Node 159: lnL = -1954.145209
        Node 160: lnL = -1954.145209
        Node 161: lnL = -1954.145209
        Node 162: lnL = -1954.145209
```

```
        Node 163: lnL = -1954.145209
        Node 164: lnL = -1954.145209
        Node 165: lnL = -1954.145209
        Node 166: lnL = -1954.145209
        Node 167: lnL = -1954.145209
        Node 168: lnL = -1954.145209
        Node 169: lnL = -1954.145209
        Node 170: lnL = -1954.145209
        Node 171: lnL = -1954.145209
        Node 172: lnL = -1954.145209
        Node 173: lnL = -1954.145209
        Node 174: lnL = -1954.145209
        Node 175: lnL = -1954.145209
        Node 176: lnL = -1954.145209

lnL = -1954.145209 from ProbSitePattern.
Joint reconstruction.

  4677120 bytes for conP, adjusted
end of tree file.

Time used:  1:57
```

[326]:

```
#step 6., continued
#reformat the data
python ../../source/parse_rst.py rst out1.txt out2.txt
```

```
output 1: (1) Marginal reconstruction of ancestral sequences
(eqn. 4 in Yang et al. 1995 Genetics 141:1641-1650).

output 2: (2) Joint reconstruction of ancestral sequences
(eqn. 2 in Yang et al. 1995 Genetics 141:1641-1650),
using the algorithm of Pupko et al. (2000 Mol Biol Evol 17:890-896),
modified to generate sub-optimal reconstructions.
```

[327]:

```
#step 6., continued
#reconstructed tree
grep -A 1 'Rod' rst | tail -n +2 > output_tree.tre
```

[328]:

```
#step 6., continued
#reconstructed sequences
python ../../source/ReformatNodes.py out1.txt out1.fasta
python ../../source/ReformatNodes.py out2.txt out2.fasta
```

This out1.fasta has multiple duplicated sequences. I'll dedpulicate it as follows, first by combining out1 and out2, then deduplicating.

[329]:

```
ls
```

```
PreReconstruct.err.txt
PreReconstruct.marker001.PF00017_HUMAN_WithSimilarSeq_50.aln
PreReconstruct.newick
PreReconstruct.newick_cleaned
PreReconstruct.out.txt
PreReconstruct.score.txt
PreReconstruct.tre
PreReconstruct_temp_iteration_0_seq_alignment.txt
PreReconstruct_temp_iteration_0_seq_unmasked_alignment.gz
PreReconstruct_temp_iteration_0_tree.tre
PreReconstruct_temp_iteration_1_seq_alignment.txt
PreReconstruct_temp_iteration_1_seq_unmasked_alignment.gz
PreReconstruct_temp_iteration_1_tree.tre
PreReconstruct_temp_iteration_2_seq_alignment.txt
PreReconstruct_temp_iteration_2_seq_unmasked_alignment.gz
PreReconstruct_temp_iteration_2_tree.tre
PreReconstruct_temp_iteration_initialsearch_seq_alignment.txt
PreReconstruct_temp_iteration_initialsearch_seq_unmasked_alignment.gz
PreReconstruct_temp_iteration_initialsearch_tree.tre
PreReconstruct_temp_name_translation.txt
PreReconstruct_temp_pasta_config.txt
codeml.ctl
codeml_pre.ctl
jones.dat
lnf
mlc
out1.fasta
out1.txt
out2.fasta
out2.txt
output_tree.tre
rst
rst1
rub
```

[330]:

```
cd ../
rm -r PF00017_HUMAN_WithSimilarSeq_Reconstruction_Gapless
mkdir PF00017_HUMAN_WithSimilarSeq_Reconstruction_Gapless
cd PF00017_HUMAN_WithSimilarSeq_Reconstruction_Gapless
cp ../PF00017_HUMAN_WithSimilarSeq_Reconstruction/PreReconstruct.marker001.
 ↪PF00017_HUMAN_WithSimilarSeq_50.aln ./
cp ../PF00017_HUMAN_WithSimilarSeq_Reconstruction/PreReconstruct.newick_cleaned␣
 ↪./

cp /Users/timeisen/Dropbox\ \(Personal\)/KuriyanLab/Sequences/
 ↪AncestralSequenceReconstruction/SH2_Domain_Reconstruction/codeml.ctl ./
 ↪codeml_pre.ctl
```

```
cp /Users/timeisen/Dropbox\ \(Personal\)/KuriyanLab/Sequences/
 ↪AncestralSequenceReconstruction/SH2_Domain_Reconstruction/
 ↪sh2_reconstruction_2/PF00017_HUMAN_WithSimilarSeq_Reconstruction/jones.dat ./

sed 's/noisy = 9/noisy = 1/' codeml_pre.ctl | sed 's/fix_blength = -1/
 ↪fix_blength = 1/' > codeml.ctl
head codeml.ctl
head jones.dat
codeml
```

```
      seqfile = PreReconstruct.marker001.PF00017_HUMAN_WithSimilarSeq_50.aln *
sequence data filename
     treefile = PreReconstruct.newick_cleaned      * tree structure file name
      outfile = mlc           * main result file name

        noisy = 1  * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1  * 0: concise; 1: detailed, 2: too much
      runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 2  * 1:codons; 2:AAs; 3:codons-->AAs

 58
 54  45
 81  16 528
 56 113  34  10
 57 310  86  49   9
105  29  58 767   5 323
179 137  81 130  59  26 119
 27 328 391 112  69 597  26  23
 36  22  47  11  17   9  12   6  16

AAML in paml version 4.8a, August 2014

processing fasta file
reading seq# 1 BTK_HUMAN/281-362                                       88 sites
reading seq# 2 I3LN58_PIG/281-362                                      88 sites
reading seq# 3 A0A2Y9FBB9_PHYMC/281-362                                88 sites
reading seq# 4 A0A5N4C153_CAMDR/307-388                                88 sites
reading seq# 5 F7BKT0_ORNAN/318-399                                    88 sites
reading seq# 6 A0A340Y2U3_LIPVE/281-355                                88 sites
reading seq# 7 F1S1L0_PIG/281-355                                      88 sites
reading seq# 8 A0A2K6TV44_SAIBB/265-338                                88 sites
reading seq# 9 M3WMI5_FELCA/280-354                                    88 sites
reading seq#10 Q5JY90_HUMAN/281-355                                    88 sites
reading seq#11 A0A671DKT5_RHIFE/280-354                                88 sites
reading seq#12 A0A6I8NCY8_ORNAN/279-353                                88 sites
reading seq#13 A0A094NE02_ANTCR/235-317                                88 sites
reading seq#14 A0A093CY13_TAUER/241-323                                88 sites
reading seq#15 A0A226MPG1_CALSU/83-165                                 88 sites
reading seq#16 A0A091QH50_MERNU/256-338                                88 sites
reading seq#17 A0A087RFM3_APTFO/241-323                                88 sites
reading seq#18 A0A6J2H0P0_9PASS/360-442                                88 sites
reading seq#19 A0A2K6EYQ1_PROCO/241-324                                88 sites
reading seq#20 A0A2Y9JLY6_ENHLU/268-351                                88 sites
reading seq#21 A0A2Y9I079_NEOSC/246-329                                88 sites
reading seq#22 Q8CFK4_MOUSE/224-307                                    88 sites
reading seq#23 A0A6J0DL91_PERMB/246-329                                88 sites
reading seq#24 H0VD88_CAVPO/225-308                                    88 sites
reading seq#25 G5B8D6_HETGA/248-331                                    88 sites
reading seq#26 L9L9I7_TUPCH/246-329                                    88 sites
reading seq#27 A0A673TJP9_SURSU/247-330                                88 sites
reading seq#28 F1SEA4_PIG/248-331                                      88 sites
reading seq#29 A0A5N3VZ71_MUNRE/239-322                                88 sites
reading seq#30 G1STI6_RABIT/247-330                                    88 sites
reading seq#31 A0A6P6CKB3_PTEVA/247-330                                88 sites
reading seq#32 A0A1U7U4B3_CARSF/247-330                                88 sites
reading seq#33 TEC_HUMAN/247-330                                       88 sites
reading seq#34 R4GB22_ANOCA/281-363                                    88 sites
reading seq#35 K7F777_PELSI/277-359                                    88 sites
reading seq#36 A0A6P5IJN4_PHACI/297-378                                88 sites
reading seq#37 F6YES1_MONDO/281-362                                    88 sites
reading seq#38 F7IHG0_CALJA/281-362                                    88 sites
reading seq#39 A0A7E6D215_9CHIR/309-390                                88 sites
reading seq#40 A0A671DKQ7_RHIFE/280-361                                88 sites
reading seq#41 A0A6P3RL81_PTEVA/281-362                                88 sites
reading seq#42 A0A6J0E543_PERMB/298-379                                88 sites
reading seq#43 H0XW86_OTOGA/284-365                                    88 sites
reading seq#44 A0A1S3GNU6_DIPOR/281-362                                88 sites
reading seq#45 A0A2K6SA55_SAIBB/296-377                                88 sites
reading seq#46 A0A3Q2HYV1_HORSE/281-362                                88 sites
reading seq#47 A0A673V1S0_SURSU/273-354                                88 sites
reading seq#48 G1MBR3_AILME/276-357                                    88 sites
reading seq#49 A0A2Y9GG10_NEOSC/276-357                                88 sites
reading seq#50 A0A384CIC9_URSMA/276-357                                88 sites
reading seq#51 A0A2Y9KLX9_ENHLU/274-355                                88 sites
reading seq#52 M3XU75_MUSPF/276-357                                    88 sites
reading seq#53 BMX_HUMAN/296-377                                       88 sites
reading seq#54 A0A1U7TS48_CARSF/280-361                                88 sites
reading seq#55 A0A6P3QJB8_PTEVA/273-354                                88 sites
reading seq#56 A0A2Y9E4Q9_TRIMA/276-357                                88 sites
reading seq#57 A0A2U4AJT4_TURTR/276-357                                88 sites
reading seq#58 A0A5N4C2L1_CAMDR/288-369                                88 sites
reading seq#59 A0A340Y4G0_LIPVE/272-353                                88 sites
reading seq#60 A0A1U7U634_CARSF/148-229                                88 sites
reading seq#61 A0A5N4EHK4_CAMDR/126-207                                88 sites
reading seq#62 A0A6P3QP21_PTEVA/150-231                                88 sites
reading seq#63 W5Q5S4_SHEEP/150-231                                    88 sites
reading seq#64 A0A6P3J1L6_BISBI/150-231                                88 sites
reading seq#65 A0A673TTW7_SURSU/174-255                                88 sites
reading seq#66 A0A2Y9JM55_ENHLU/174-255                                88 sites
reading seq#67 E2RBA0_CANLF/174-255                                    88 sites
reading seq#68 A0A485N2L8_LYNPA/150-231                                88 sites
reading seq#69 A0A2Y9RMH4_TRIMA/152-233                                88 sites
reading seq#70 TXK_HUMAN/150-231                                       88 sites
reading seq#71 A0A2K5PHS5_CEBIM/150-231                                88 sites
reading seq#72 A0A096NDG3_PAPAN/150-231                                88 sites
reading seq#73 A0A5F7ZHT0_MACMU/150-231                                88 sites
reading seq#74 A0A2K5E184_AOTNA/150-231                                88 sites
reading seq#75 F7FTR6_MONDO/239-323                                    88 sites
reading seq#76 A0A2K6AVK0_MACNE/239-323                                88 sites
reading seq#77 G3UBJ2_LOXAF/247-331                                    88 sites
reading seq#78 ITK_HUMAN/239-323                                       88 sites
reading seq#79 A0A452SNA4_URSAM/250-331                                88 sites
reading seq#80 G1LHK5_AILME/239-323                                    88 sites
reading seq#81 A0A286ZPK2_PIG/217-301                                  88 sites
reading seq#82 S7N0E0_MYOBR/242-326                                    88 sites
reading seq#83 I3MD63_ICTTR/239-323                                    88 sites
reading seq#84 L5L1K8_PTEAL/239-323                                    88 sites
reading seq#85 A0A6P6HUD7_PUMCO/239-323                                88 sites
reading seq#86 A0A671FJF7_RHIFE/239-323                                88 sites
reading seq#87 A0A452GAH1_CAPHI/239-323                                88 sites
reading seq#88 W5PNG3_SHEEP/239-323                                    88 sites
reading seq#89 D4A7W7_RAT/239-323                                      88 sites
ns = 89         ls = 88
Reading sequences, sequential format..
Counting site patterns..  0:00
          68 patterns at       72 /       72 sites (100.0%),  0:00
Counting frequencies..

    31328 bytes for distance
    21760 bytes for conP
        0 bytes for fhK
  5000000 bytes for space

TREE #  1

   946560 bytes for conP, adjusted

1 node(s) used for scaling (Yang 2000 J Mol Evol 51:423-432):
 112

ntime & nrate & np:   175     0   175
```

```
np =   175
lnL0 = -2211.211358
Out..
lnL  = -1678.245544
46034 lfun, 0 eigenQcodon, 8055950 P(t)

Reconstructed ancestral states go into file rst.

lnL = -1678.245544 from ProbSitePattern.
Marginal reconstruction.
        Node  90: lnL = -1678.245544
        Node  91: lnL = -1678.245544
        Node  92: lnL = -1678.245544
        Node  93: lnL = -1678.245544
        Node  94: lnL = -1678.245544
        Node  95: lnL = -1678.245544
        Node  96: lnL = -1678.245544
        Node  97: lnL = -1678.245544
        Node  98: lnL = -1678.245544
        Node  99: lnL = -1678.245544
        Node 100: lnL = -1678.245544
        Node 101: lnL = -1678.245544
        Node 102: lnL = -1678.245544
        Node 103: lnL = -1678.245544
        Node 104: lnL = -1678.245544
        Node 105: lnL = -1678.245544
        Node 106: lnL = -1678.245544
        Node 107: lnL = -1678.245544
        Node 108: lnL = -1678.245544
        Node 109: lnL = -1678.245544
        Node 110: lnL = -1678.245544
        Node 111: lnL = -1678.245544
        Node 112: lnL = -1678.245544
        Node 113: lnL = -1678.245544
        Node 114: lnL = -1678.245544
        Node 115: lnL = -1678.245544
        Node 116: lnL = -1678.245544
        Node 117: lnL = -1678.245544
        Node 118: lnL = -1678.245544
        Node 119: lnL = -1678.245544
        Node 120: lnL = -1678.245544
        Node 121: lnL = -1678.245544
        Node 122: lnL = -1678.245544
        Node 123: lnL = -1678.245544
        Node 124: lnL = -1678.245544
        Node 125: lnL = -1678.245544
        Node 126: lnL = -1678.245544
        Node 127: lnL = -1678.245544
        Node 128: lnL = -1678.245544
        Node 129: lnL = -1678.245544
        Node 130: lnL = -1678.245544
        Node 131: lnL = -1678.245544
        Node 132: lnL = -1678.245544
        Node 133: lnL = -1678.245544
        Node 134: lnL = -1678.245544
        Node 135: lnL = -1678.245544
        Node 136: lnL = -1678.245544
        Node 137: lnL = -1678.245544
        Node 138: lnL = -1678.245544
        Node 139: lnL = -1678.245544
        Node 140: lnL = -1678.245544
        Node 141: lnL = -1678.245544
        Node 142: lnL = -1678.245544
        Node 143: lnL = -1678.245544
        Node 144: lnL = -1678.245544
        Node 145: lnL = -1678.245544
        Node 146: lnL = -1678.245544
        Node 147: lnL = -1678.245544
        Node 148: lnL = -1678.245544
        Node 149: lnL = -1678.245544
        Node 150: lnL = -1678.245544
        Node 151: lnL = -1678.245544
        Node 152: lnL = -1678.245544
        Node 153: lnL = -1678.245544
        Node 154: lnL = -1678.245544
        Node 155: lnL = -1678.245544
        Node 156: lnL = -1678.245544
        Node 157: lnL = -1678.245544
        Node 158: lnL = -1678.245544
        Node 159: lnL = -1678.245544
        Node 160: lnL = -1678.245544
        Node 161: lnL = -1678.245544
        Node 162: lnL = -1678.245544
        Node 163: lnL = -1678.245544
        Node 164: lnL = -1678.245544
        Node 165: lnL = -1678.245544
        Node 166: lnL = -1678.245544
        Node 167: lnL = -1678.245544
        Node 168: lnL = -1678.245544
        Node 169: lnL = -1678.245544
        Node 170: lnL = -1678.245544
        Node 171: lnL = -1678.245544
        Node 172: lnL = -1678.245544
        Node 173: lnL = -1678.245544
        Node 174: lnL = -1678.245544
```

```
        Node 175: lnL = -1678.245544
        Node 176: lnL = -1678.245544

lnL = -1678.245544 from ProbSitePattern.
Joint reconstruction.

  3786240 bytes for conP, adjusted
end of tree file.

Time used:  1:22
```

[331]:

```
python ../../source/parse_rst.py rst out1.txt out2.txt
grep -A 1 'Rod' rst | tail -n +2 > output_tree.tre
#step 6., continued
#reconstructed sequences
python ../../source/ReformatNodes.py out1.txt out1.fasta
python ../../source/ReformatNodes.py out2.txt out2.fasta
```

```
output 1: (1) Marginal reconstruction of ancestral sequences
(eqn. 4 in Yang et al. 1995 Genetics 141:1641-1650).

output 2: (2) Joint reconstruction of ancestral sequences
(eqn. 2 in Yang et al. 1995 Genetics 141:1641-1650),
using the algorithm of Pupko et al. (2000 Mol Biol Evol 17:890-896),
modified to generate sub-optimal reconstructions.
```

[332]:

```
awk 'BEGIN{FS=OFS=""}($1 == ">"){split($0, array, "#"); print array[1], "#", array[2] + 243}($1 != ">"){print $0}' out1.fasta > out1_rename.fasta
```

construct the c-term helix, which is removed by the ancestral reconstruction for some reason:

[370]:

```
awk 'BEGIN{FS=OFS=""}($1 == ">"){print $0}($1 != ">"){print $0,"IPELINYH"}' out1_rename.fasta > out1_rename_append.fasta
```

[371]:

```
cat ../PF00017_HUMAN_WithSimilarSeq_Reconstruction/out1.fasta out1_rename_append.fasta > ancestral_seqs.fasta
python ../../source/DeduplicateFasta.py ancestral_seqs.fasta ancestral_seqs_dedup.fasta
```

[372]:

```
#step 6 continued
grep -c ">" ancestral_seqs_dedup.fasta
```

```
114
```

[373]:

```
#step 7.
cd ../
hmmbuild PF00017_HMM_Dedup PF00017_full_dedup.fasta #build the HMM for the PFAM database from the SH2 domain
```

```
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             PF00017_full_dedup.fasta
# output HMM file:                  PF00017_HMM_Dedup
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                  nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     PF00017_full_dedup   23121   529   211  1621.75  0.590

# CPU time: 0.51u 0.04s 00:00:00.55 Elapsed: 00:00:00.59
```

[374]:
```
#step 7., continued
#align the new sequences with the old seqs
hmmalign --informat fasta --mapali PF00017_full_dedup.fasta --amino PF00017_HMM_Dedup PF00017_HUMAN_WithSimilarSeq_Reconstruction_Gapless/ancestral_seqs_dedup.fasta > PF00017_Dedup_with_ARseqs.txt
```

[375]:
```
#step 7., continued
#now PF00017_full_with_ARseqs is an alignment file in a stockholm format.
python ../source/StockholmToFasta.py PF00017_Dedup_with_ARseqs.txt PF00017_Dedup_with_ARseqs.fasta
```

```
Converted 23235 records
```

[376]:
```
#step 6., continued
#original_seqs and similar
tail -n +2 PF00017_HUMAN_WithSimilarSeq_50.txt | cut -f 1-2 | awk 'BEGIN{FS="\t";OFS=""}{print $1, "\n", $2}' | sort | uniq |  ggrep -A 9 --no-group-separator -f - PF00017_Dedup_with_ARseqs.fasta > HumanAndSimilarAndAncestral.fasta
#add in ancestral
grep ">" PF00017_HUMAN_WithSimilarSeq_Reconstruction_Gapless/ancestral_seqs_dedup.fasta | cut -c 2- | ggrep -A 9 --no-group-separator -f - PF00017_Dedup_with_ARseqs.fasta >> HumanAndSimilarAndAncestral.fasta
```

[377]:
```
#step 7., continued
wc -l HumanAndSimilarAndAncestral.fasta
grep -c ">" HumanAndSimilarAndAncestral.fasta
head -11 HumanAndSimilarAndAncestral.fasta
# tail -11 HumanAndSimilarAndAncestral.fasta
```

```
    2040 HumanAndSimilarAndAncestral.fasta
204
>BTK_HUMAN/281-362
---------------------------------------------WYS-K--H-M---T-
```

```
-R---------------------SQ--A-E-Q-L-LKQ---------------------e
GK--E--G-G--FI-------------V--R-----------D------S--------S-
-------------K----------A--------G-----------K---------Y----
---T----V---S--VFA--KStgd-----------------------------------
---pqgVIR-H----Y--V----V--C-----S---T---P--QS---------------
-----------------Q-Y-Y---L------A-----E------K-----HL-------
F--S--T--I-P-ELINYH-----------------------------------------
--------------------------------------------------
>I3LN58_PIG/281-362
```

[378]:

```
#step 7., continued
python ../source/Compare_to_BTK_FullIds.py HumanAndSimilarAndAncestral.fasta HumanAndSimilarAndAncestral_IDs.txt "---------------------------------------------WYS-K--H-M---T--R---------------------SQ--A-E
tail HumanAndSimilarAndAncestral_IDs.txt
```

```
node#399        0.5121951219512195
node#400        0.5365853658536586
node#401        0.524390243902439
node#408        0.524390243902439
node#410        0.524390243902439
node#411        0.5121951219512195
node#413        0.524390243902439
node#414        0.9390243902439024
node#417        0.9512195121951219
node#418        0.9390243902439024
```

[379]:

```
#step 8., continued
#remake the tree
rm -r HumanAndSimilarAndAncestralChecking
python /Users/timeisen/Applications/Pasta/pasta/run_pasta.py -a -i HumanAndSimilarAndAncestral.fasta -d protein -o HumanAndSimilarAndAncestralChecking/
```

```
PASTA INFO: Reading input sequences from 'HumanAndSimilarAndAncestral.fasta'…
PASTA INFO: Masking alignment sites with less than 41 sites before running the
tree step
PASTA INFO: Configuration written to "/Users/timeisen/Dropbox (Personal)/Kuriyan
Lab/Sequences/AncestralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reco
nstruction_4/HumanAndSimilarAndAncestralChecking/pastajob_temp_pasta_config.txt"
.

PASTA INFO: Directory for temporary files created at
/Users/timeisen/.pasta/pastajob/temprxv6jsvt
PASTA INFO: Name translation information saved to /Users/timeisen/Dropbox (Perso
nal)/KuriyanLab/Sequences/AncestralSequenceReconstruction/SH2_Domain_Reconstruct
ion/sh2_reconstruction_4/HumanAndSimilarAndAncestralChecking/pastajob_temp_name_
translation.txt as safe name, original name, blank line format.
PASTA INFO: Creating a starting tree for the PASTA algorithm…
PASTA INFO: Input sequences assumed to be aligned (based on sequence lengths).
PASTA INFO: Performing initial tree search to get starting tree…
PASTA INFO: Starting PASTA algorithm on initial tree…
PASTA INFO: Max subproblem set to 102
PASTA INFO: Step 0. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 0. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted and score improved.
PASTA INFO: current score: -3258.76, best score: -3258.76
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Step 1. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 1. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted despite the score not improving.
PASTA INFO: current score: -2303.017, best score: -2303.017
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Step 2. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 2. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted despite the score not improving.
PASTA INFO: current score: -2343.459, best score: -2303.017
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Writing resulting alignment to HumanAndSimilarAndAncestralChecking/p
astajob.marker001.HumanAndSimilarAndAncestral.aln
PASTA INFO: Writing resulting tree to
HumanAndSimilarAndAncestralChecking/pastajob.tre
PASTA INFO: Writing resulting likelihood score to
HumanAndSimilarAndAncestralChecking/pastajob.score.txt
PASTA INFO: The resulting alignment (with the names in a "safe" form) was first
written as the file "/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Anc
estralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/Huma
nAndSimilarAndAncestralChecking/pastajob_temp_iteration_2_seq_alignment.txt"
PASTA INFO: The resulting tree (with the names in a "safe" form) was first
written as the file "/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Anc
estralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/Huma
nAndSimilarAndAncestralChecking/pastajob_temp_iteration_2_tree.tre"
Refused to clean '/Users/timeisen/.pasta/pastajob/temprxv6jsvt/step2/mincluster/
pw/tempopalu_ha1zfp': not created by PASTA
'/Users/timeisen/.pasta/pastajob/temprxv6jsvt/step2/mincluster/pw/tempopalu_ha1z
fp' is not registered as a temporary directory that was created by this process!
Refused to clean '/Users/timeisen/.pasta/pastajob/temprxv6jsvt/step0/mincluster/
pw/tempopal3t9h0m4r': not created by PASTA
'/Users/timeisen/.pasta/pastajob/temprxv6jsvt/step0/mincluster/pw/tempopal3t9h0m
4r' is not registered as a temporary directory that was created by this process!
Refused to clean '/Users/timeisen/.pasta/pastajob/temprxv6jsvt/step1/mincluster/
pw/tempopal0d5th_mc': not created by PASTA
'/Users/timeisen/.pasta/pastajob/temprxv6jsvt/step1/mincluster/pw/tempopal0d5th_
mc' is not registered as a temporary directory that was created by this process!
PASTA INFO: Total time spent: 20.94556999206543s
```

[380]:
```
#add in the sequences of the human that are less then 50% id
cat HumanAndSimilarAndAncestral.fasta > HumanAndSimilarAndAncestralWithLowerIDs.
↪fasta
awk 'BEGIN{FS=OFS="\t"}($2 >= 0.25 && $2 < 0.5){print $0}' PF00017_HUMAN_IDs.
↪txt > PF00017_HUMAN_IDs_filtered_25to50.txt
cut -f 1 PF00017_HUMAN_IDs_filtered_25to50.txt | ggrep -A 9␣
↪--no-group-separator -f - PF00017_Dedup_with_ARseqs.fasta >>␣
↪HumanAndSimilarAndAncestralWithLowerIDs.fasta
```

[384]:
```
#compare results to BTK
python ../source/Compare_to_BTK_FullIds.py␣
↪HumanAndSimilarAndAncestralWithLowerIDs.fasta␣
↪HumanAndSimilarAndAncestralWithLowerIDs.txt␣
↪"---------------------------------------------WYS-K--H-M---T--R---------------------SQ--A-E
```

[385]:
```
head HumanAndSimilarAndAncestralWithLowerIDs.txt
```

```
Seq_ID  pairwise_identity
BTK_HUMAN/281-362       1.0
I3LN58_PIG/281-362      0.975609756097561
A0A1U7U4B3_CARSF/247-330        0.573170731707317
TEC_HUMAN/247-330       0.573170731707317
W5Q5S4_SHEEP/150-231    0.5365853658536586
A0A2Y9GG10_NEOSC/276-357        0.47560975609756095
A0A2U4AJT4_TURTR/276-357        0.5
A0A2Y9JLY6_ENHLU/268-351        0.5609756097560976
A0A6J2H0P0_9PASS/360-442        0.8414634146341463
```

[386]:
```
##not run, I'll just use the pasta alignment
#ungap
sed 's/\-//g' HumanAndSimilarAndAncestralWithLowerIDs.fasta | awk '/^>/␣
↪{printf("\n%s\n",$0);next; } { printf("%s",$0);}  END {printf("\n");}' |␣
↪tail -n +2 > HumanAndSimilarAndAncestralWithLowerIDs_gapless.fasta

grep -v ">" HumanAndSimilarAndAncestralWithLowerIDs_gapless.fasta | awk␣
↪'BEGIN{FS=OFS=""}{print NF}' | sort | uniq -c
```

```
   1 66
   1 71
   5 72
   2 74
  16 75
  10 76
   5 77
   3 78
  53 80
   3 81
  51 82
  14 83
  17 84
  15 85
   2 86
  61 88
   1 92
   1 93
   1 94
   1 95
   1 96
```

There's only one sequence of 86 aa. Most of them have many fewer. BTK is 82 aa.

I need to get rid of some seqs. To do this, I'm going to remove anc nodes from the Q5JY90 seq.

[396]:
```
pwd
```

```
/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/AncestralSequenceReconst
ruction/SH2_Domain_Reconstruction/sh2_reconstruction_4
```

[409]:
```
echo \
"node#341
node#98
node#99
node#342
node#346
node#100
node#101
node#343
node#103" | cat - > NodeRemoval.txt

python ../source/FilterSeqsByName.py NodeRemoval.txt␣
 ↪HumanAndSimilarAndAncestralWithLowerIDs.fasta␣
 ↪HumanAndSimilarAndAncestralWithLowerIDs_Trim.fasta
```

[410]:
```
grep -c ">" HumanAndSimilarAndAncestralWithLowerIDs_Trim.fasta
```

```
255
```

[411]:
```
python ../source/DeduplicateFasta.py␣
 ↪HumanAndSimilarAndAncestralWithLowerIDs_Trim.fasta␣
 ↪HumanAndSimilarAndAncestralWithLowerIDsDedupAttempt.fasta
grep -c ">" HumanAndSimilarAndAncestralWithLowerIDsDedupAttempt.fasta
```

```
255
```

[412]:
```
#These next steps begin the construction of controls and nt sequences.
#make a folder for this.
mkdir ConstructNtSeqs
cd ConstructNtSeqs
```

```
mkdir: ConstructNtSeqs: File exists
```

[413]:
```
#step 8. remove seqs longer than 90
python ../../source/FilterSeqsByLength.py ../
 ↪HumanAndSimilarAndAncestralWithLowerIDs_Trim.fasta␣
 ↪HumanAndSimilarAndAncestralWithLowerIDs_ShortSeqs.fasta 90
#HumanAndSimilarAndAncestralWithLowerIDs_gapless.fasta

#step 8.
#remove sites with 100% gaps
/Users/timeisen/Applications/Pasta/pasta/run_seqtools.py -infile␣
 ↪HumanAndSimilarAndAncestralWithLowerIDs_ShortSeqs.fasta -informat FASTA ␣
 ↪-outformat FASTA -outfile␣
 ↪HumanAndSimilarAndAncestralWithLowerIDs_Reduce_ShortSeqs.fasta -masksites 1
```

[414]:
```
#step 9. take the 30 sequences closest to BTK, and create R mutations for them.
python ../../source/RMutator.py␣
 ↪HumanAndSimilarAndAncestralWithLowerIDs_Reduce_ShortSeqs.fasta␣
 ↪SeqsWithKR_Mutant_Controls_Reduce.fasta ../HumanAndSimilarAndAncestral_IDs.
 ↪txt
```

```
seqs mutated: 62
```

[415]:
```
grep -c ">" "SeqsWithKR_Mutant_Controls_Reduce.fasta"
```

```
329
```

[432]:
```
#step 10. make the final nt seqs, with BsaI sites
python ../../source/BTK_SH2_ancestorsV2.py SeqsWithKR_Mutant_Controls_Reduce.
 ↪fasta BTK_SH2_seqs.json final_nt_output.txt
```

```
Sequences written:  1256
```

[434]:
```
rm -r FinalCheck
sed 's/\*/X/g' SeqsWithKR_Mutant_Controls_Reduce.fasta >␣
 ↪SeqsWithKR_Mutant_Controls_Reduce_PASTA_INPUT.fasta
python /Users/timeisen/Applications/Pasta/pasta/run_pasta.py -a -i␣
 ↪SeqsWithKR_Mutant_Controls_Reduce_PASTA_INPUT.fasta -d protein -o FinalCheck/
```

```
rm: FinalCheck: No such file or directory
PASTA INFO: Reading input sequences from
'SeqsWithKR_Mutant_Controls_Reduce_PASTA_INPUT.fasta'…
PASTA INFO: Masking alignment sites with less than 108 sites before running the
tree step
PASTA INFO: Configuration written to "/Users/timeisen/Dropbox (Personal)/Kuriyan
Lab/Sequences/AncestralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reco
nstruction_4/ConstructNtSeqs/FinalCheck/pastajob_temp_pasta_config.txt".

PASTA INFO: Directory for temporary files created at
/Users/timeisen/.pasta/pastajob/tempxbcc14ep
```

```
PASTA INFO: Name translation information saved to /Users/timeisen/Dropbox (Perso
nal)/KuriyanLab/Sequences/AncestralSequenceReconstruction/SH2_Domain_Reconstruct
ion/sh2_reconstruction_4/ConstructNtSeqs/FinalCheck/pastajob_temp_name_translati
on.txt as safe name, original name, blank line format.
PASTA INFO: Creating a starting tree for the PASTA algorithm…
PASTA INFO: Input sequences assumed to be aligned (based on sequence lengths).
PASTA INFO: Performing initial tree search to get starting tree…
PASTA INFO: Starting PASTA algorithm on initial tree…
PASTA INFO: Max subproblem set to 165
PASTA INFO: Step 0. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 0. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted and score improved.
PASTA INFO: current score: -9937.525, best score: -9937.525
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Step 1. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 1. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted and score improved.
PASTA INFO: current score: -8902.192, best score: -8902.192
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Step 2. Realigning with decomposition strategy set to mincluster
PASTA INFO: Step 2. Alignment obtained. Tree inference beginning…
PASTA INFO: realignment accepted despite the score not improving.
PASTA INFO: current score: -8887.572, best score: -8887.572
PASTA INFO: TreeShrink option has been turned off!
PASTA INFO: Writing resulting alignment to
FinalCheck/pastajob.marker001.SeqsWithKR_Mutant_Controls_Reduce_PASTA_INPUT.aln
PASTA INFO: Writing resulting tree to FinalCheck/pastajob.tre
PASTA INFO: Writing resulting likelihood score to FinalCheck/pastajob.score.txt
PASTA INFO: The resulting alignment (with the names in a "safe" form) was first
written as the file "/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Anc
estralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/Cons
tructNtSeqs/FinalCheck/pastajob_temp_iteration_2_seq_alignment.txt"
PASTA INFO: The resulting tree (with the names in a "safe" form) was first
written as the file "/Users/timeisen/Dropbox (Personal)/KuriyanLab/Sequences/Anc
estralSequenceReconstruction/SH2_Domain_Reconstruction/sh2_reconstruction_4/Cons
tructNtSeqs/FinalCheck/pastajob_temp_iteration_2_tree.tre"
PASTA INFO: Total time spent: 37.57833003997803s
```

```
[ ]:
```