Phylogeny and Molecular Biology

The Evolutionary Trace

- Inspiration for the applications project
  - Lichtarge O., et al. J. Mol. Biol., 257(2): 342-58, 1996.
- If we make a basic assumption:
  - Our evolutionary tree is representative of a family of homologous proteins.
    - We call these proteins “homologs”
- We can find a number of markers that are powerful indicators of molecular function
  - Conserved amino acids are likely to involved in function
  - Amino acids that vary with subfamily variations likely control differences in function
  - Conserved and partly conserved amino acids cluster spatially around active sites

The Evolutionary Trace finds active sites

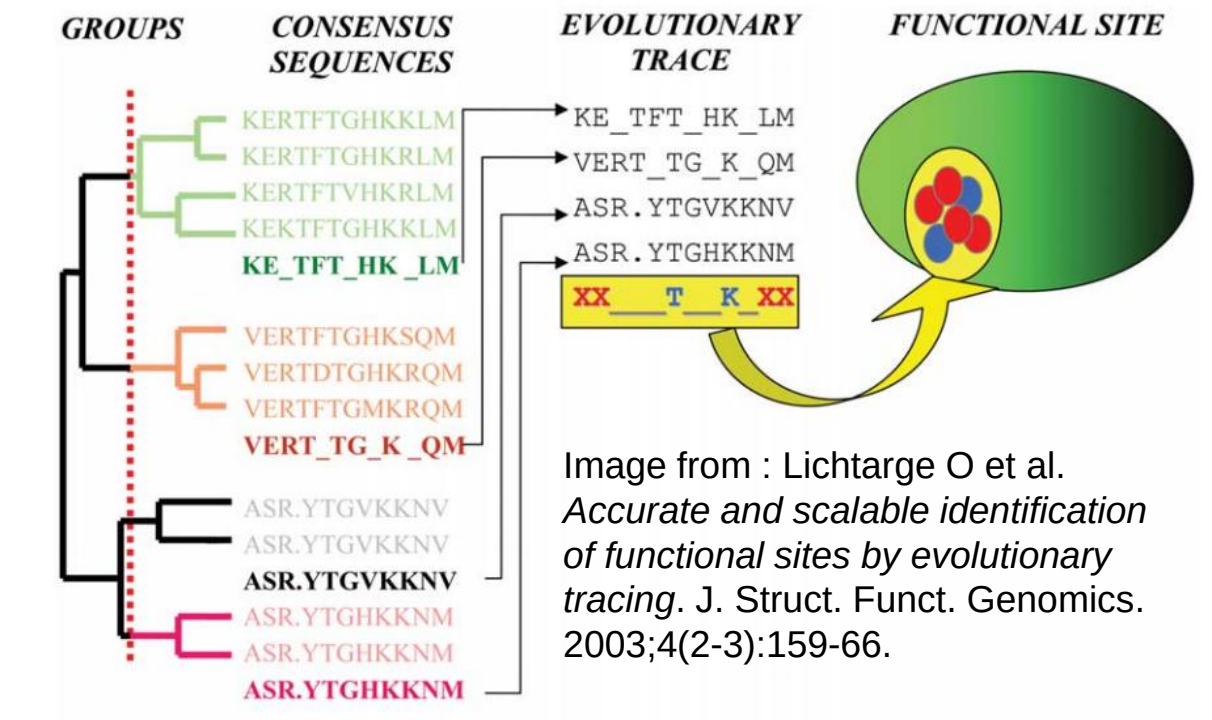


Image from : Lichtarge O et al. *Accurate and scalable identification of functional sites by evolutionary tracing*. J. Struct. Funct. Genomics. 2003;4(2-3):159-66.

Step 1) Build an evolutionary tree

Step 2) choose a level of evolutionary significance

Step 3) Select consensus sequences

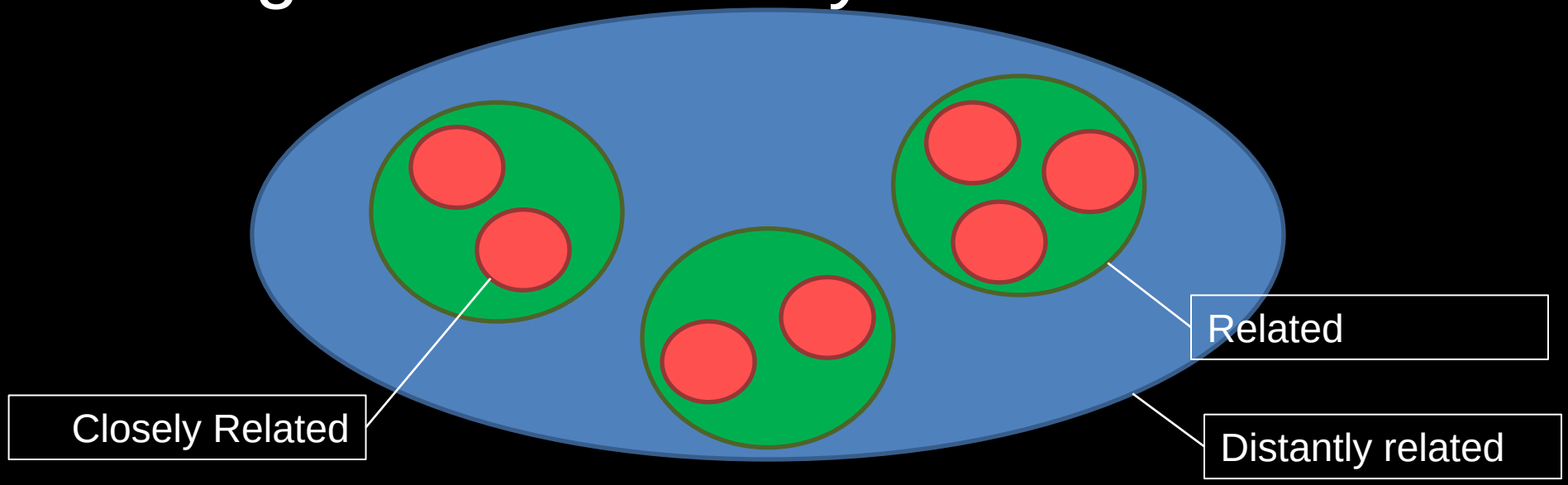
Step 4) Get consensus sequence alignment

Step 5) Map to structure

The theory behind the Evolutionary Trace

- Begin with an evolutionary tree that represents real evolutionary variations
- This tree will separate subfamilies of the multiple sequence alignment that are evolutionarily distinct
- Consensus sequences that indicate subfamily conservation can be aligned to find which sequence positions are conserved globally
- Globally conserved and subfamily conserved sequence positions are important for function.

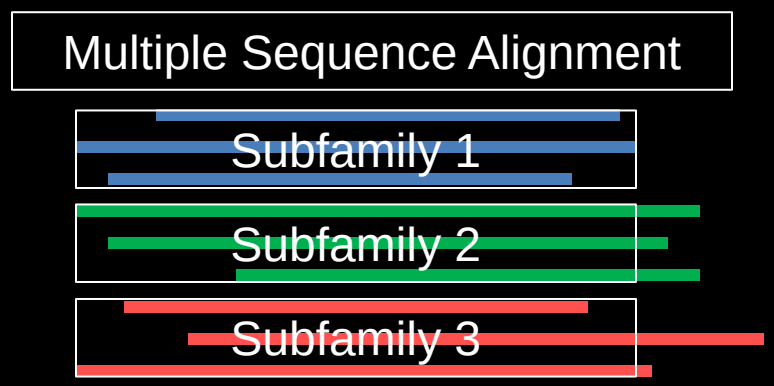
Building an Evolutionary Tree

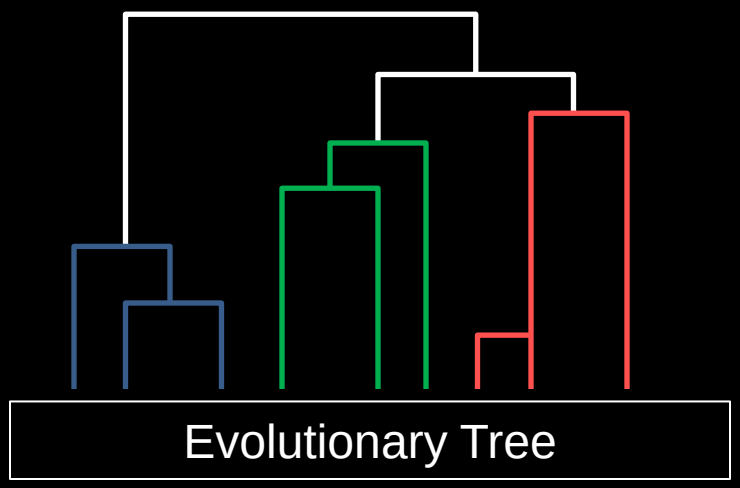


- Every protein is a modern-day descendant of a wide range of ancestors
- Each protein has many cousins and siblings that descend from similar recent ancestors, and the same distant ancestors
- The tough part is to figure out who belongs there

You need an alignment to build a tree

- A *good* multiple sequence alignment is necessary before you build a tree
- This is the hardest and most confusing part of the entire operation
  - In fact, the tree is just a way to visualize the multiple sequence alignment!





On “good” multiple sequence alignments

- Many of these guidelines are common sense
- Nonetheless, <u>many</u> people learn these guidelines the hard way.
  - What is the hard way:
  - Making a multiple sequence alignment that is totally pointless, and wasting a lot of time trying to improve it

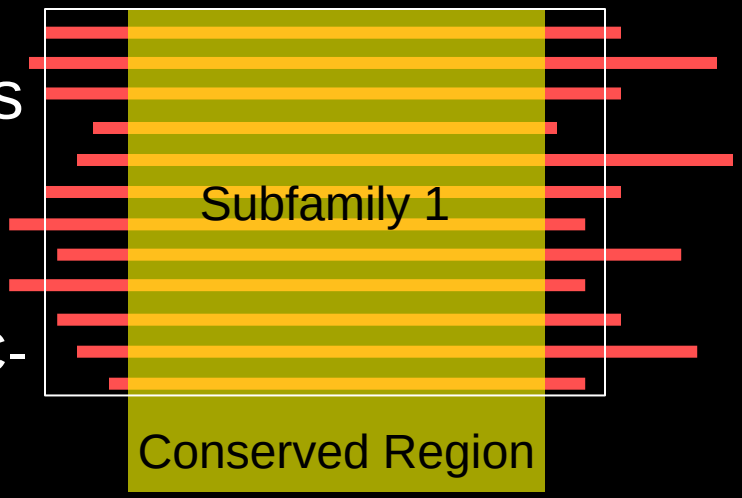
- Today, we will talk about the Evolutionary Trace process, and best practices for getting acceptable multiple sequence alignments.

Good Multiple Sequence Alignments ...

… are <u>Nonredundant</u>

– genBank holds many thousands of protein sequences, many of which are essentially identical

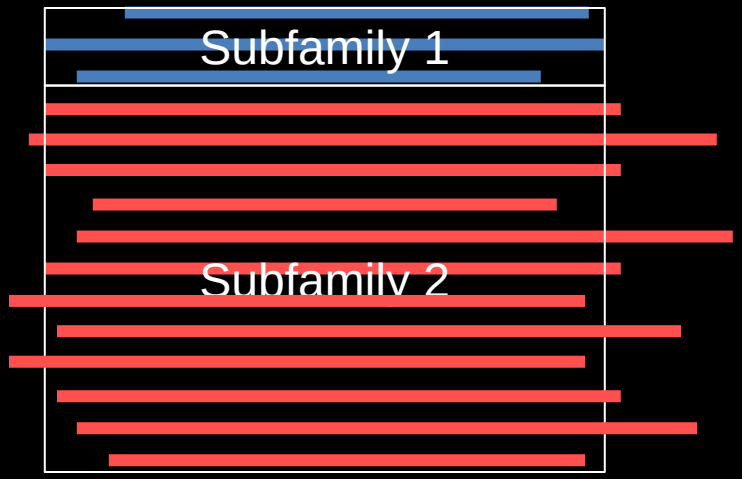
  - perhaps slightly different N- and C-terminii

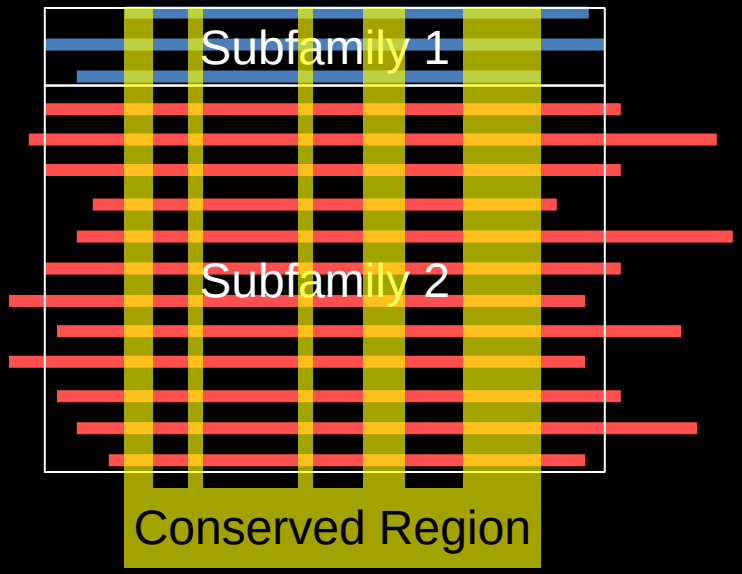


- What happens?
  - MSAs with large numbers of redundant sequences have big conserved regions
  - Redundant subfamilies have misleadingly large numbers of conserved amino acids
  - Those amino acids are not necessarily important, they just come up because too many identical sequences were used

Even small amounts of diversity help

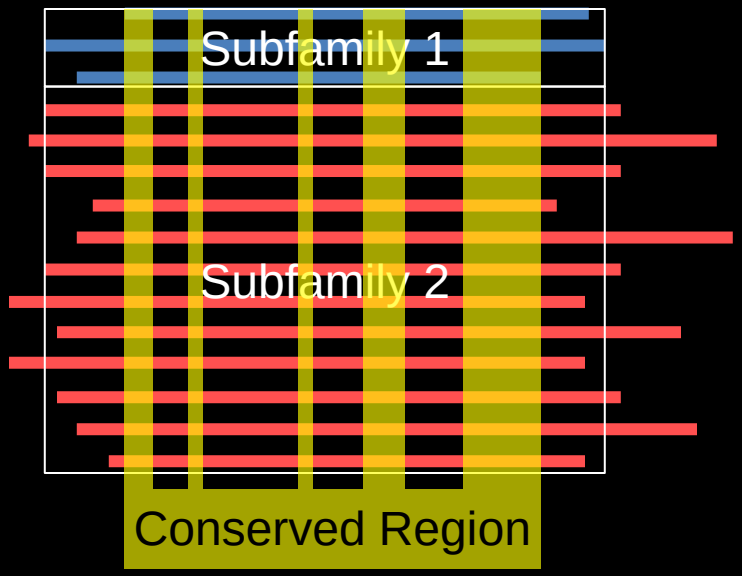
- Both Subfamily1 and Subfamily2 may be fairly redundant
- However, the probability that they are identical at the same residues is low
  - Unless the region if functionally important, in which case we want to detect those amino acids





Small degrees of diversity aren't everything

- Only 3-10 amino acids will be totally conserved among a large family of Proteins
- An additional subfamily will eliminate many conserved positions
- Even if 80% of misleadingly conserved positions are eliminated, the vast majority are still unnecessarily conserved.

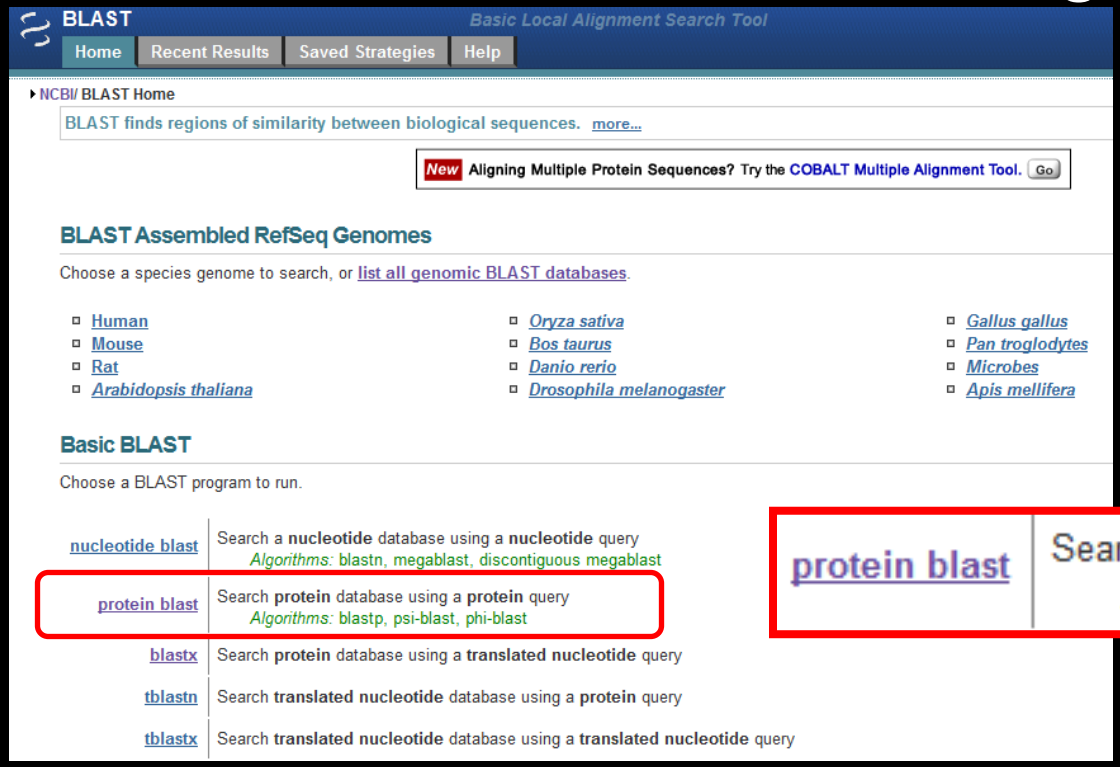


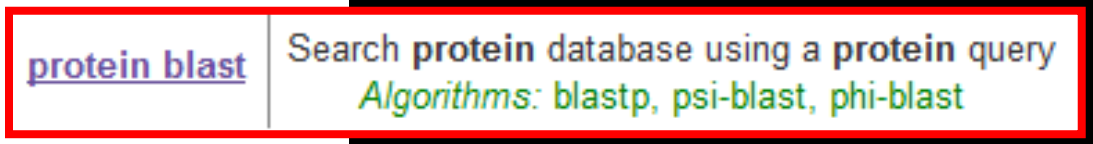
How you can improve diversity

- PSI-Blast searches
  - Psi-blast uses position specific substitution matrices to score its alignments
  - When we talked about sequence alignments, the substitution matrices applied to all parts of the sequence
  - Psi-blast uses a different, dynamically generated substitution matrix at each position of the alignment

- As a result, PSI-Blast identifies homologs at greater evolutionary distances

PSI-blast: www.ncbi.nlm.nih.gov/blast/Blast.cgi?

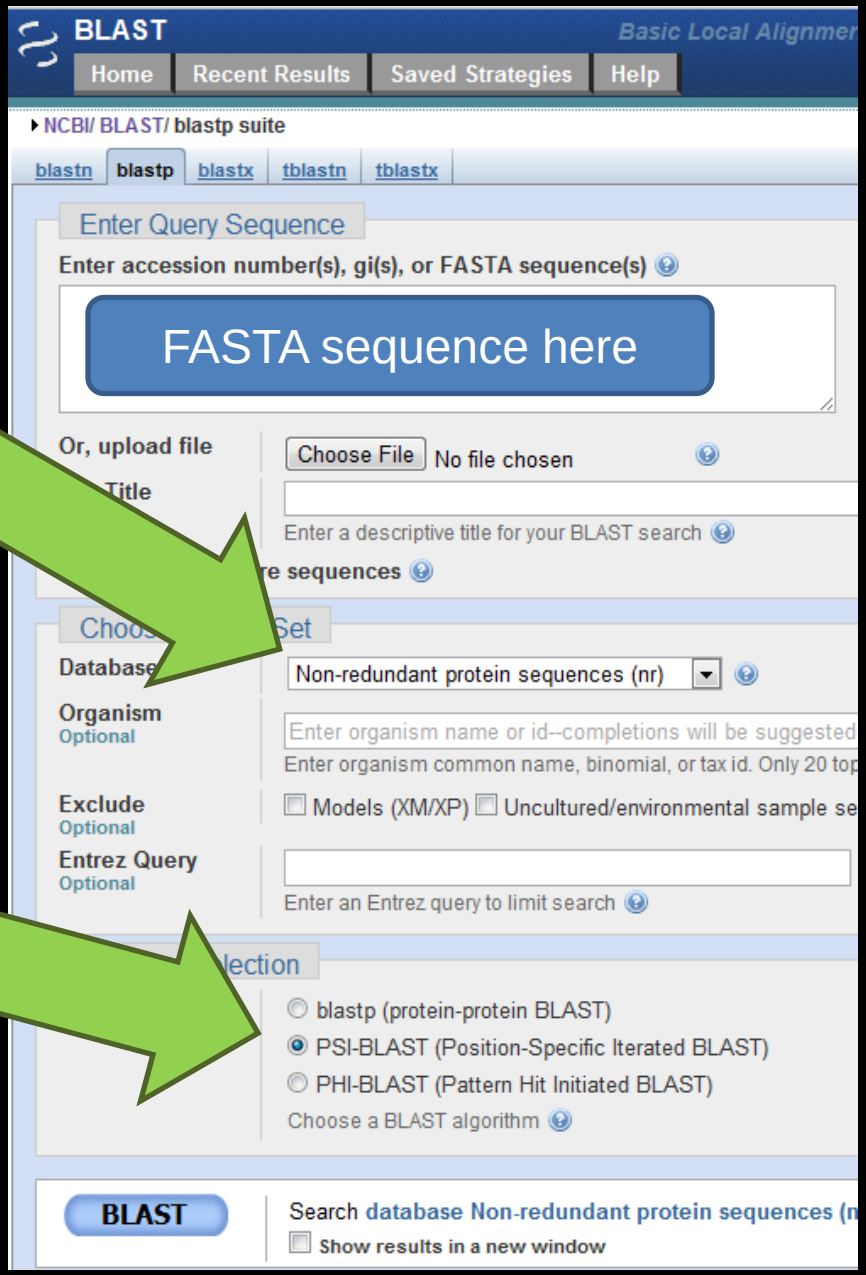




- PSI-Blast works identically to MegaBLAST, except that it searches a database of protein sequences using the unusual weight matrices

Searching with PSI-Blast

- Make sure to select the non-redundant protein sequences database (nr)
- Also select PSI-BLAST in the radio buttons below
- But you know all this.



PSI-blast: Iteratively increasing diversity

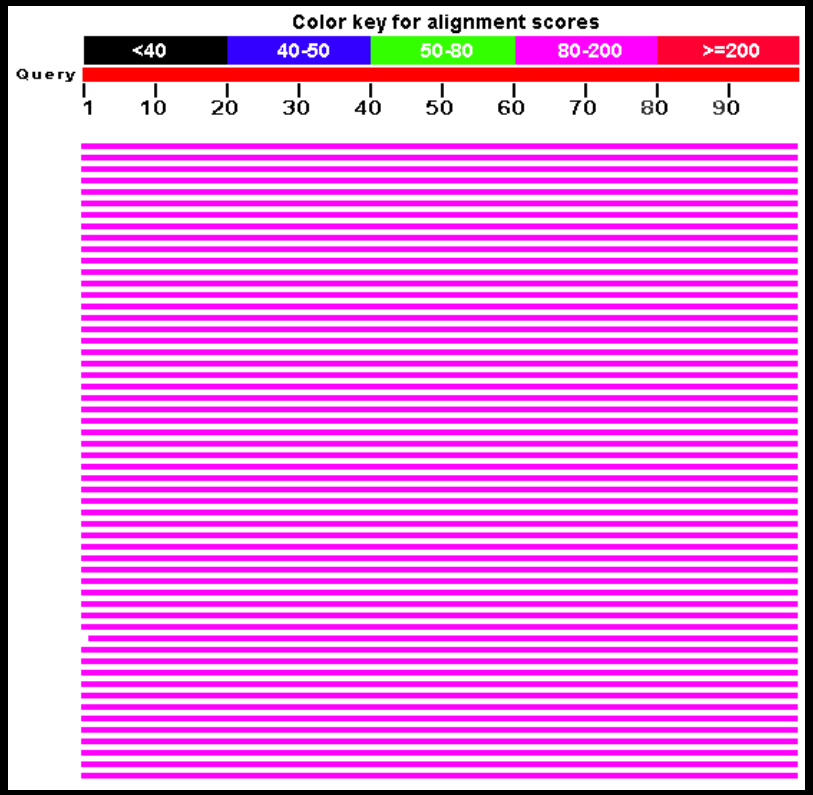
- HIV protease is an extremely well studied virus
- Like many retroviruses, it has a malfuntioning reverse transcriptase protein, which causes it to mutate constantly
- HIV tends to rapidly become resistant to drug treatments
- Sequencing many HIV genomes gives us a sense of how evolution happens in HIV.

The sequence I will use to build my tree:

```
>HIV-Protease
PQITLWKRPLVTIKIGGQLKEALLDT
GADDTVIEEMSLPGRWKPKMIGGIGG
FIKVRQYDQIIIEIAGHKAIGTVLVG
PTPVNIIGRNLLTQIGATLNF
```

Lets see what we get with PSI-Blast

- Running protease on PSI-Blast got us 500 identical matches
- <u>Massive lack of diversity in this alignment</u>
- How do we fix it?
- Lets have a look at which organisms we are matching



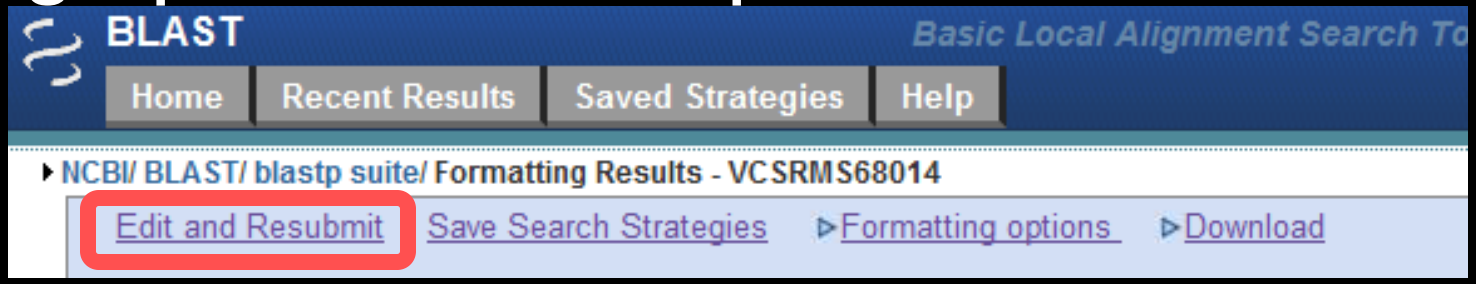
Evolutionary Tree (very flat)

Inspecting the list of PSI-Blast matches

NEW	ADQ92502.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	9e-48
NEW	ADQ92527.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	9e-48
NEW	ADQ92521.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	9e-48
NEW	ADQ92496.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	9e-48
NEW	ADQ92525.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	9e-48
NEW	ADQ92539.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	1e-47
NEW	ADQ92450.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	1e-47
NEW	3CYW_A	Chain A, Effect Of Flap Mutations On Structure Of Hiv-1 Protease And Inl	192	192	100%	1e-47
NEW	1DAZ_C	Chain C, Structural And Kinetic Analysis Of Drug Resistant Mutants Of Hi	192	192	100%	1e-47
NEW	ADQ92500.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	1e-47
NEW	ADQ92458.1	pol protein [Human immunodeficiency virus 1]	192	192	100%	1e-47

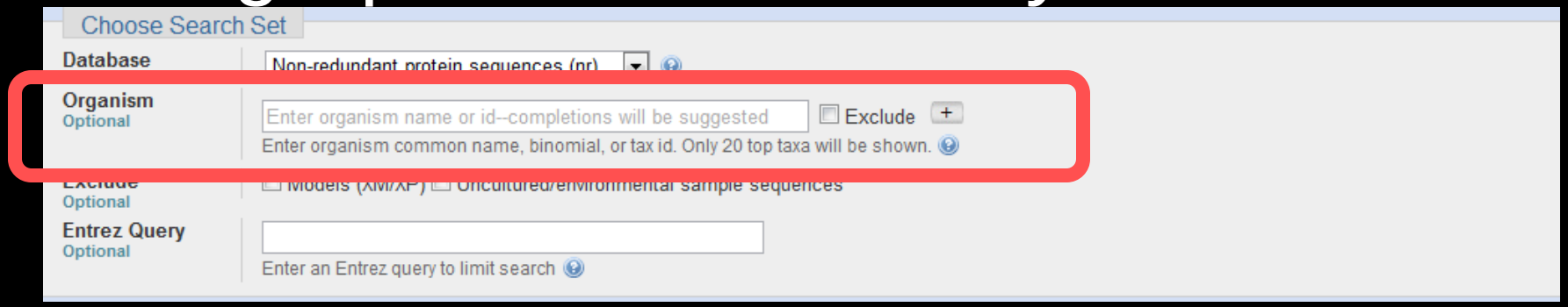
- A huge number of the matches come from [Human immunodeficiency virus 1]
- This is obvious – we submitted a protein from HIV-1. Naturally, we are going to match the same protein.
- While this is an extreme case of lack of diversity, a similar problem may happen in your applications projects: overrepresentation.

Fixing species overrepresentation in MSAs

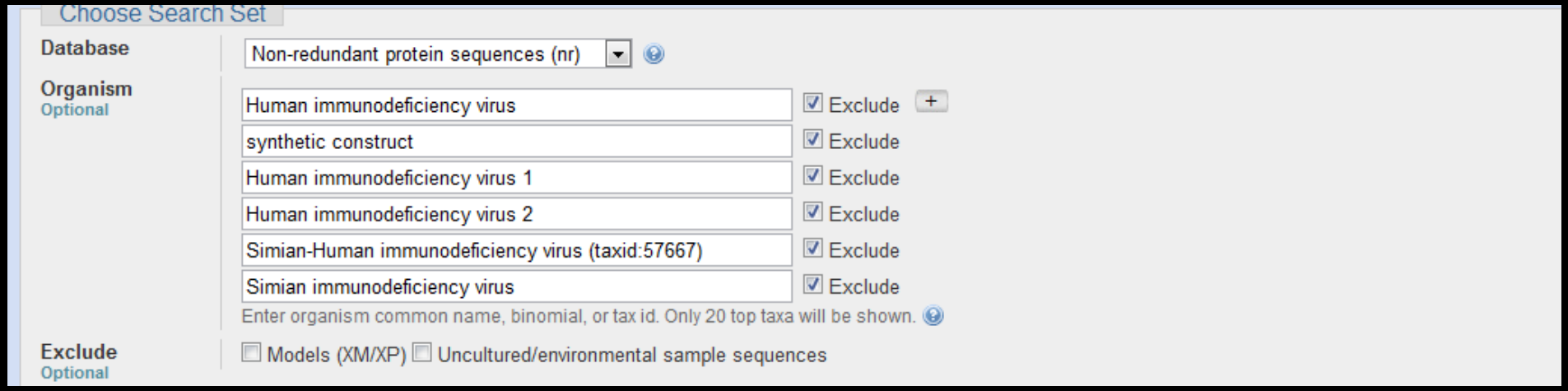


- The “Edit and Resubmit” button is a great way to improve your searches without having to start everything over.
- You can use this in several iterations to keep improving the diversity of your alignment:
  1) Submit
  2) Observe
  3) Resubmit..

Limiting Species redundancy



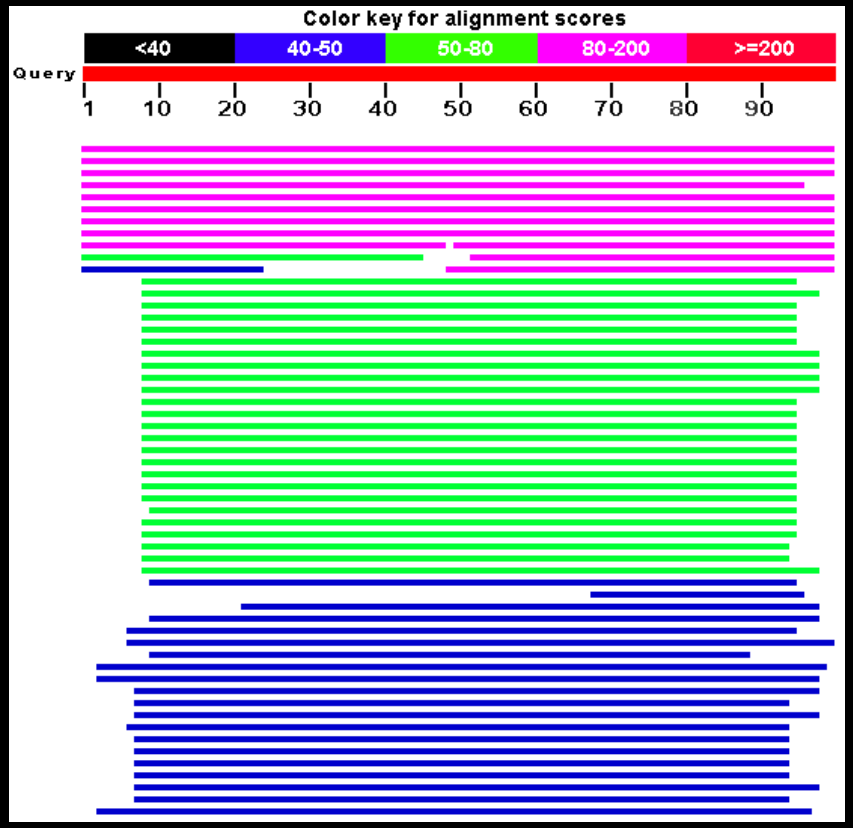
- Here you can limit the search to several species that you select, or more importantly, you can exclude several species you have seen already.
- I'll exclude HIV and some others:



A new filtered alignment

- By excluding HIV1, HIV2 and several other viruses that have nearly identical proteases, we have improved diversity considerably

- Note that this is not perfect: there are some oddly named HIV-1 things here
  - I'd keep fixing this up

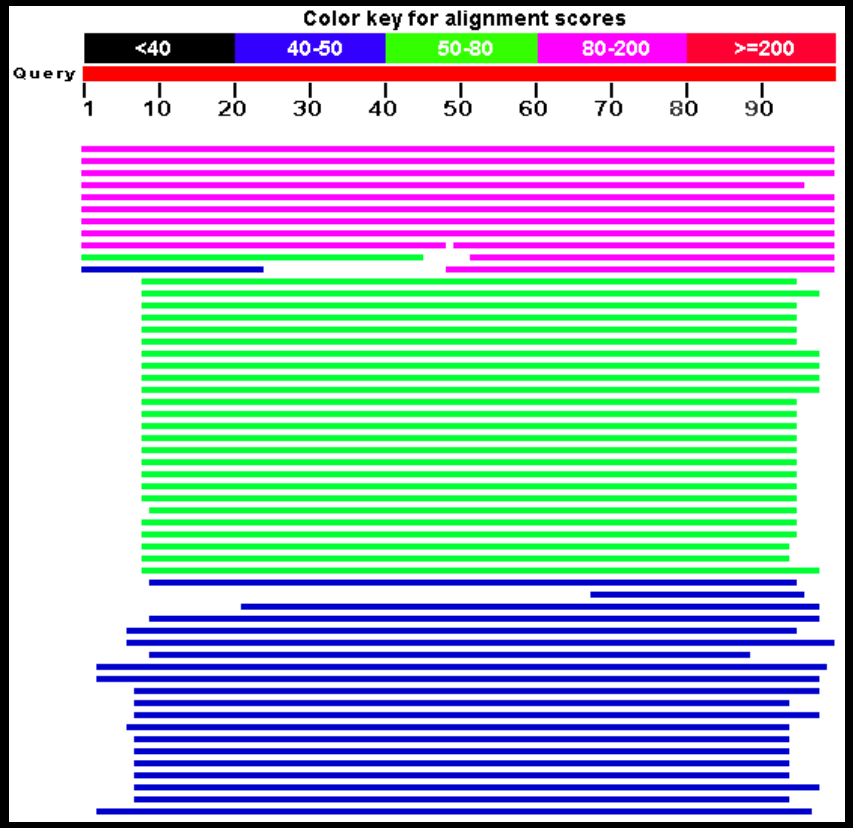


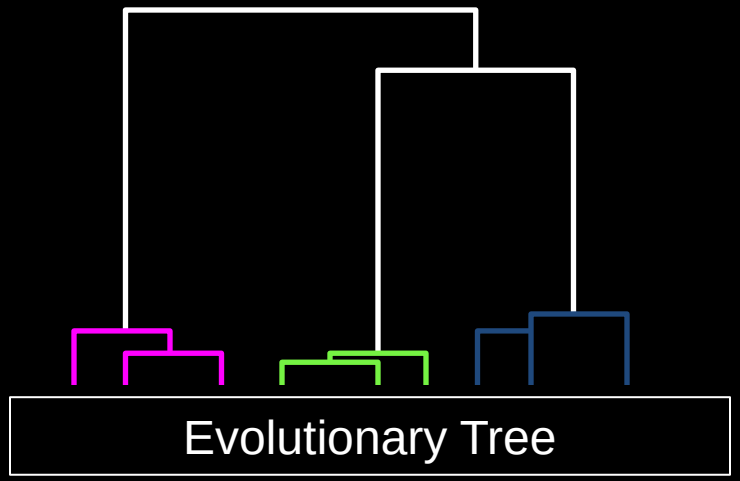
Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Links
NEW AAK08484.2	pol polyprotein [HIV-1 vector pNL4-3]	192	192	100%	1e-47	
NEW BAF34642.1	pol polyprotein [HIV-1 vector pNL-DT5R]	188	188	100%	2e-46	
NEW ACY01941.1	pol protein [HIV whole-genome vector AA1305#18]	188	188	100%	2e-46	
NEW 1Q9P_A	Chain A, Solution Structure Of The Mature Hiv-1 Protease Monomer	187	187	95%	6e-46	S
NEW 3FSM_A	Chain A, Crystal Structure Of A Chemically Synthesized 203 Amino Acid	175	175	100%	2e-42	S
NEW AAV69858.1	Pol polyprotein [SIV vector pCLN8]	110	110	100%	5e-23	
NEW ABY60459.1	Gag-Pol-Nef protein [Expression vector MVA--89.6P-SIVGPN]	110	110	100%	7e-23	
NEW ADL66917.1	protease [Cloning vector pMC1s::WT-HIV2Pr]	108	108	100%	2e-22	
NEW ZP_02875333.1	hypothetical protein cdivTM_33997 [candidate division TM7 single-cell	87.8	87.8	50%	5e-16	
NEW ZP_02871245.1	hypothetical protein cdivTM_13331 [candidate division TM7 single-cell	87.4	87.4	48%	5e-16	

Diversity visible in the phylogenetic tree

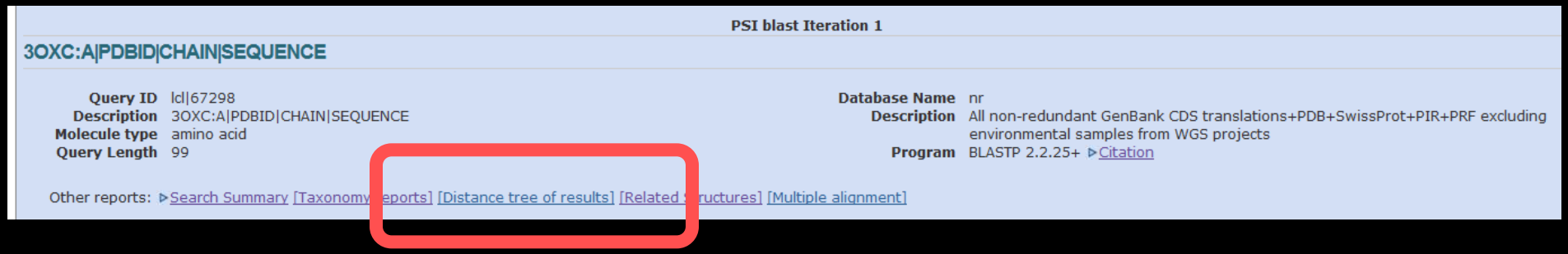
- By excluding HIV1, HIV2 and several other viruses that have nearly identical proteases, we have somewhat improved diversity





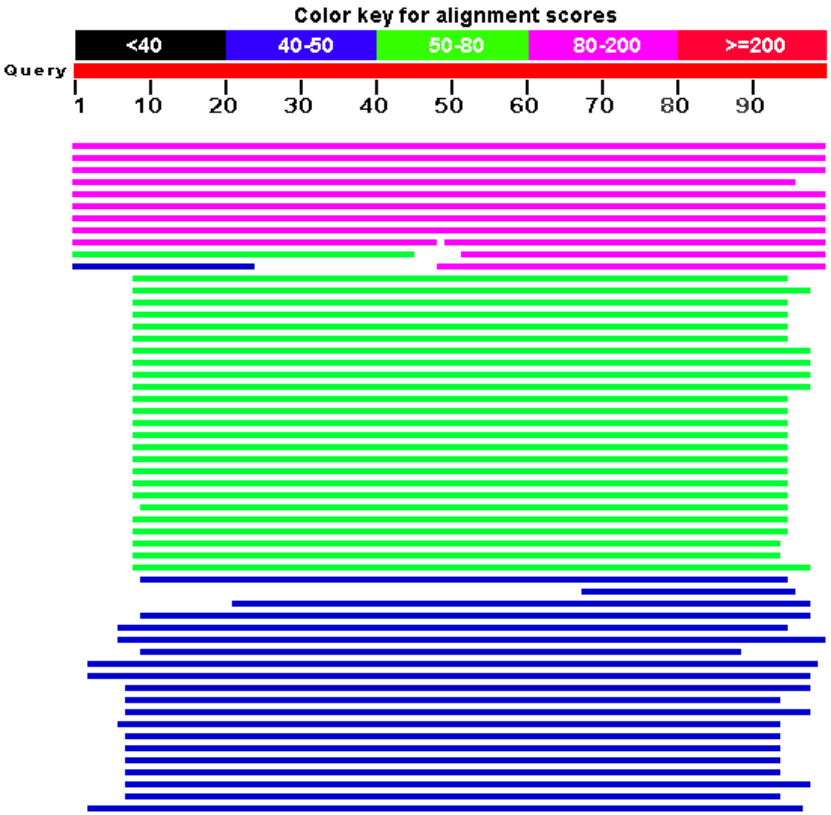
- The tree becomes an effective tool for estimating diversity

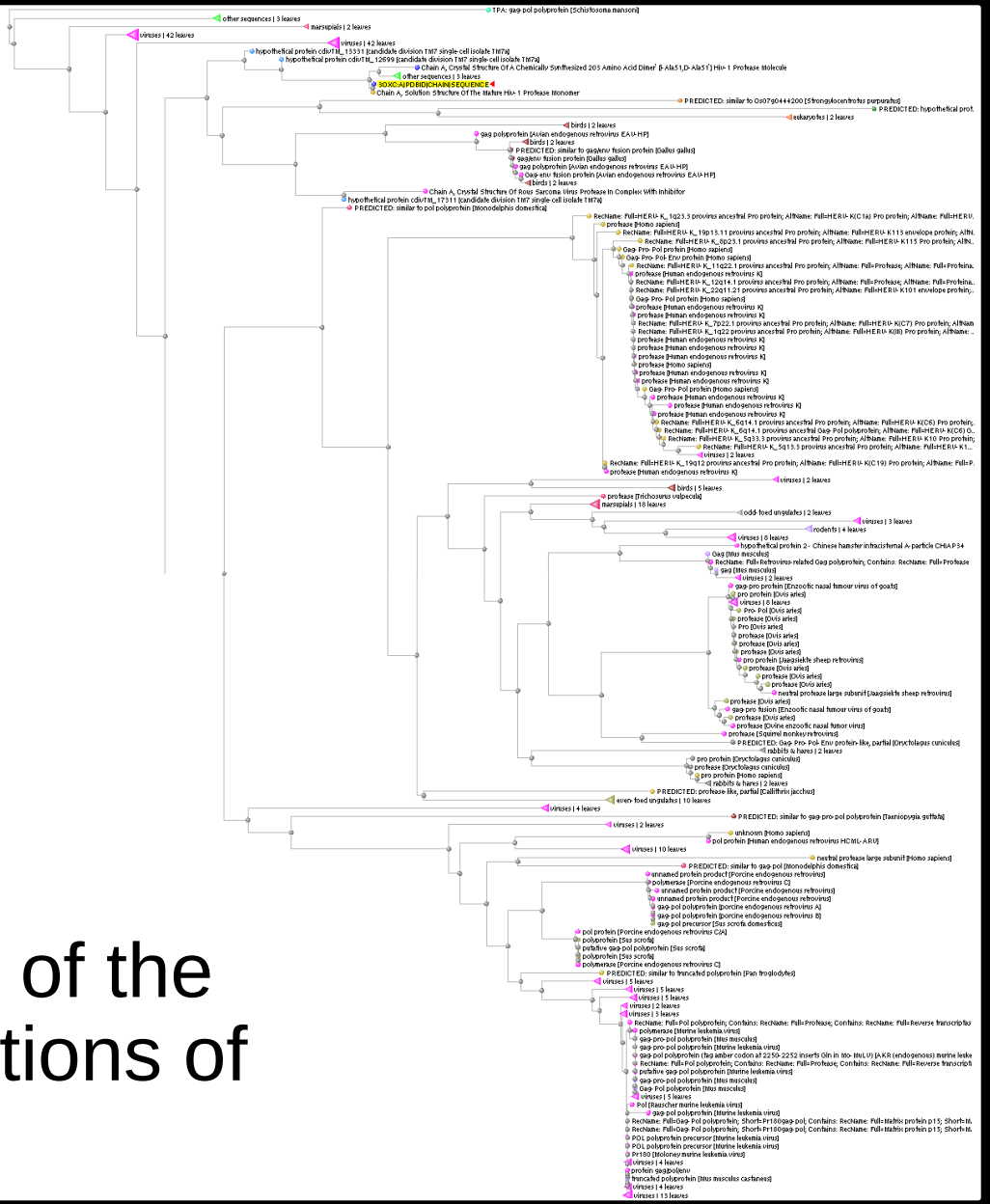
Getting a tree from PSI-Blast



- While you can download the fasta and rescore the alignment, you can also use a distance tree generated at NCBI
  - This will be faster for your earlier results
- The tree itself is generated from pairwise Blast runs, rather than an actual multiple sequence alignment.
  - This makes it slightly less accurate than what you get from ClustalW + Phylip
  - Good enough as you make your set more diverse

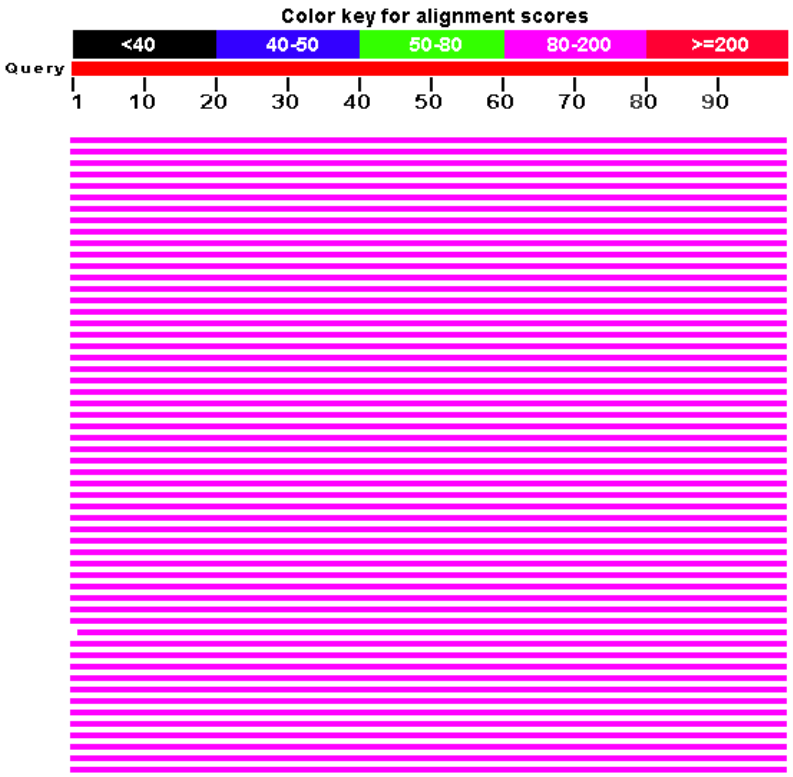
The tree from PSI-Blast

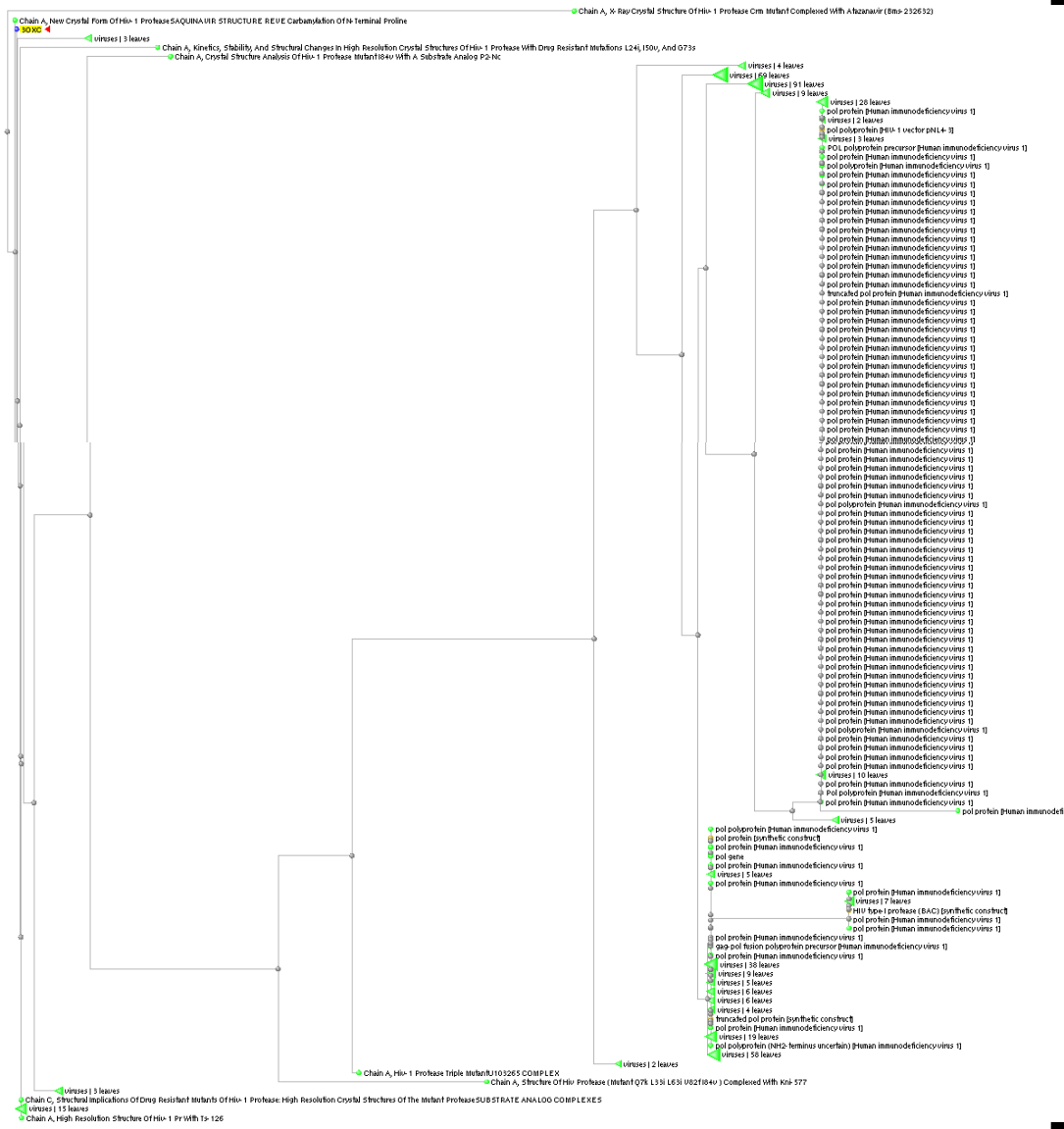




- Notice how the clades of the tree correlate with sections of the alignment

Compared to before..





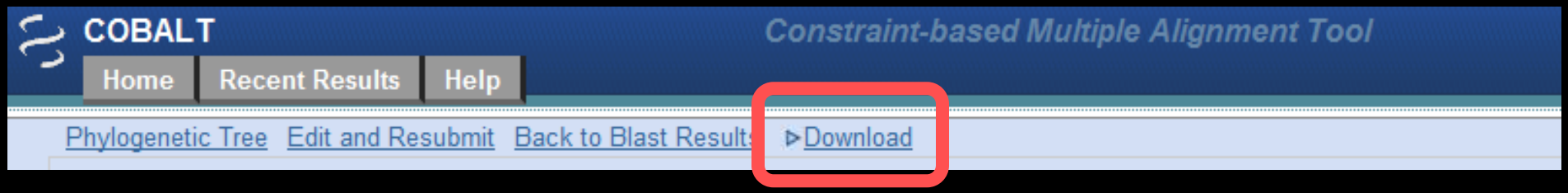
- Notice how the clades of the tree correlate with sections of the alignment

The phylogenetic tree measures diversity

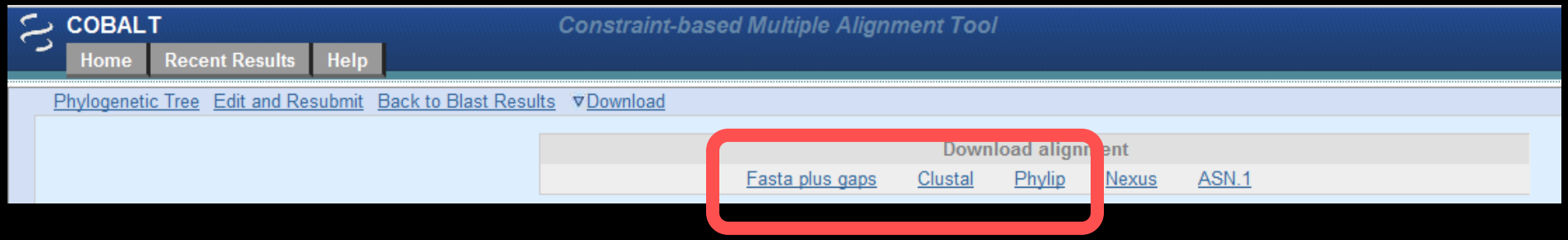
- You can see immediately that the two trees are different
- Get used to noticing more:
  - The topology of the first tree contains a more diverse (though not highly diverse) group of clades
  - The topology of the second tree maps clades that classify identical sequences
- As you build your representative sets, monitor the tree for diversity.
- Periodically save out the fasta files for your records, so you don’t lose anything online

Saving your data

- From the Multiple Sequence Alignment view
  - At the top, you will notice this bar:

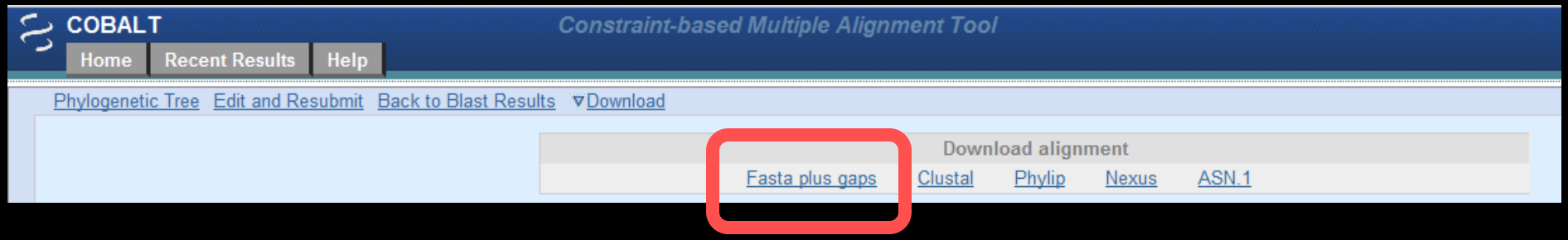


- Select “Download”, to get the following menu:



- Two members of this menu are the most important for saving your progress:

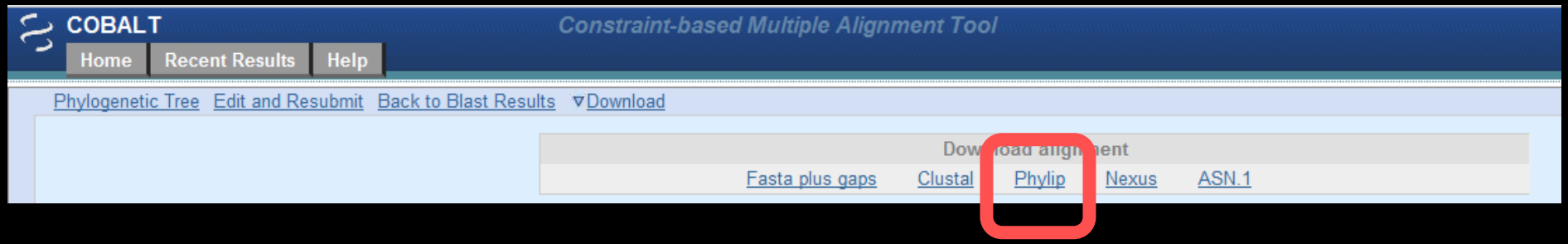
Saving your data



- This is a FASTA file of all the sequences you have in your alignment
- Gaps are added into the sequences to preserve the current alignment
- If you give this fasta file to ClustalW, it will rescore it for you

```
---------------
>gi|296556483 pol polyprotein [HIV-1 vector pNL4-3]
------------------------------------------------------------
------------------------------------------------------------
----------------------------FFREDLAFPQGKAREFSSEQTRANSPT---RR
ELQVWGRDNNSLSEA---GADRQGTVSFSFPQITLWQRPLVTIKIGGQ----------LK
EALL-DTGADDTVLEEMNLPGRW-----KPKM-IGGIGGFIKVRQY-DQILIEICGHKAI
G--T--VLVGP----TPVNIIGRNLLTQIGCTLNFP---ISPIETVPVK---------LK
PGMDGPKVKQWPLTEEKIKALVEIC--------TEMEKEGKI-SKIGPENPYNT------
PVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKQKKSVTVLDVGDAYFSV
PLDKDFRKYTAFTIPSINNETPGIRY---------------------------------Q
YNVLPQGWKGSPAIFQCSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEEL
RQH-LLRWGFTTPDKKHQKEPPFLWMGYELHPDKWTVQPIVLPE--KDSWTVNDIQKLVG
KLNWASQIYAGIKVRQLCKLLRGTKALTEVVPLTEEAELELAENREILKEPVHGVYYDPS
KDLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARMKGAHTNDVKQLTEAVQKIATESIVI
WGKTPKFKLPIQKETW--EAWWTEYWQATWIPEWEFVNTPPLVKLWYQLEKEPIIGAETF
YVDGAANRETKLGKAGYVTDRGRQKVVPLTDTTNQKTELQAIHLALQDSGLEVNIVTDSQ
YALGIIQAQPDKS--ESELVSQIIEQLIKKEKVYLAWVPAHKGIGGNEQVDKLVS-----
-AGI-------------------------------------------------------
---------------------------------RK--------------------------
---------VLFLDGIDKAQEEHEKYHSNWRAMASDFNLPPVVAKEIVASCDKCQLKG--
EAMHGQVDCSPGIWQLDCTHLEGKVILVAVHVASGYIEAEVIPAETGQETAYFLLKLAGR
WPVKTVHTDNGSNFTSTTVKAACWWAGIKQEFGIPYNPQSQGVIESMNKELKKIIGQVRD
QAEHLKTAVQMAVFIHNFKRKGGIGGYSAGERIVDIIATDIQTKELQKQITKIQNFRVYY
R---DSRDPVWKGPAKLLWKGEGAVVIQDNSDIK--VVPRRKAKIIRDYGKQMAGDDCVA
SRQDED--------
```

Saving your data



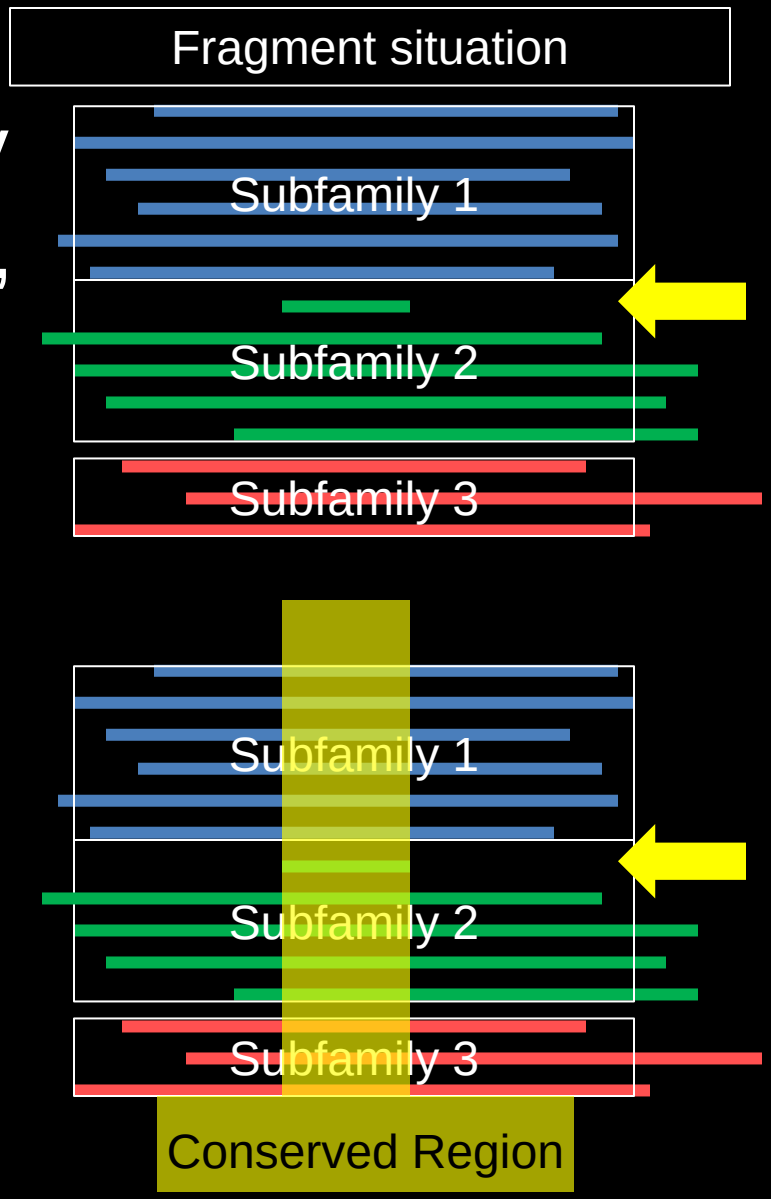
- This outputs a sequence alignment that is compatible with Phylip's protpars program
- This will allow you to compute a high-parismony tree using the NCBI alignment

```
  49   1810
3OXC_A_PD ---------------------------------------------------
pol_polyp ---------------------------------------------------
pol_polyp ---------------------------------------------------
pol_prote ---------------------------------------------------
Chain_A__ ---------------------------------------------------
Chain_A__ ---------------------------------------------------
Pol_polyp ---------------------------------------------------
Gag_Pol_N MRVRNSVLSGKKADELEKIRLRPNGKKKYMLKHVVWAANELDRFGLAESLL
protease_ ---------------------------------------------------
hypotheti ---------------------------------------------------
hypotheti ---------------------------------------------------
hypotheti ---------------------------------------------------
hypotheti ---------------------------------------------------
hypotheti ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_polyp ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_polyp ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_polyp ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_polyp ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_prote ---------------------------------------------------
pol_prote ---------------------------------------------------
RecName__ ---------------------------------------------------
pol_prote ---------------------------------------------------
```

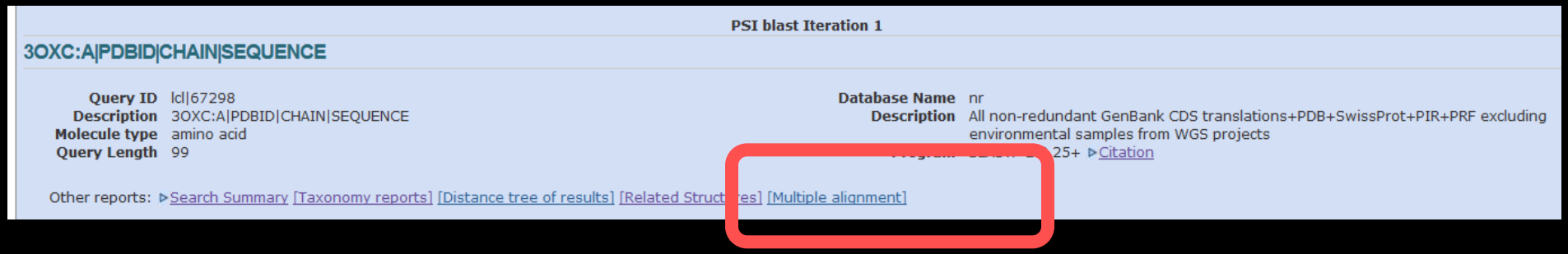
Good Multiple Sequence Alignments …

Avoid fragments

- Fragments might align very well with the other proteins,
  - but they convey little information
- Amino acids that align to gaps cannot be considered fully conserved:
  - Fragments mess up the interpretation of all the other sequences



Fixing the fragments problem



- We can fix the fragments by removing them from the alignment.
- In the results page after a list of matches computed with PSI-Blast, we find this command: [Multiple alignment].
- NCBI computes its multiple sequence alignments with COBALT, which is also a very effective alignment software

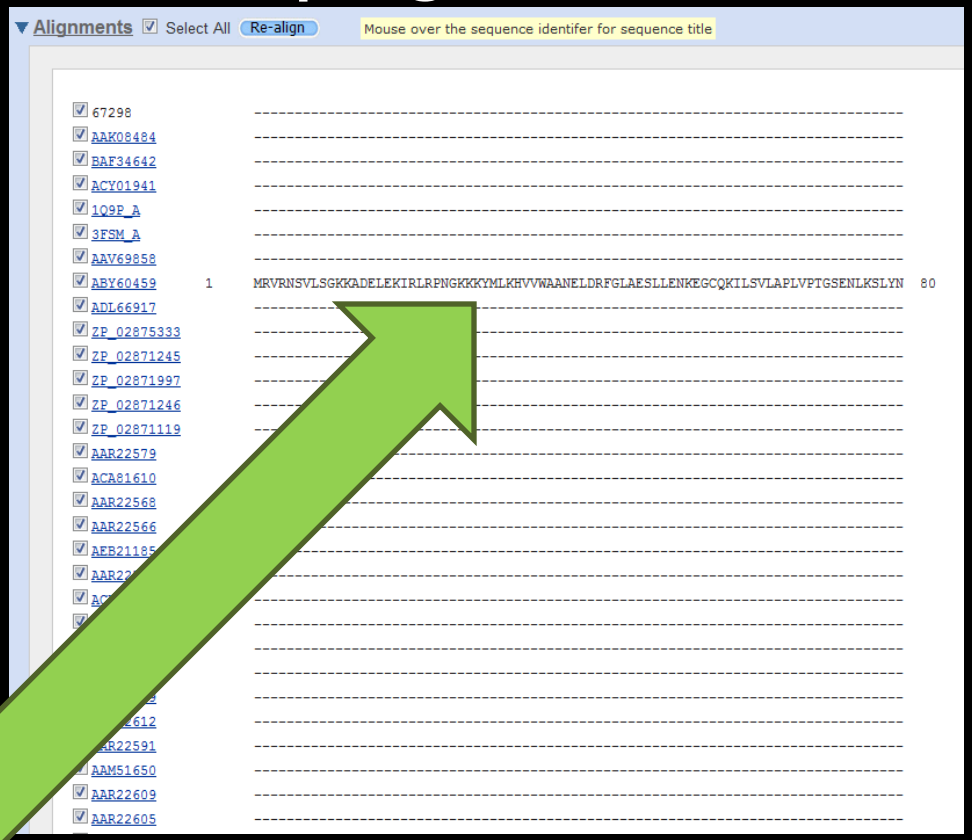
The Multiple Sequence Alignment Page

Accession	Description	Links
lcl|67298	3OXC:A|PDBID|CHAIN|SEQUENCE	
AAK08484.2	pol polyprotein [HIV-1 vector pNL4-3] >gb|ACB38949.1| pol protein [Human immunodeficiency virus 1] >gb|ACM50125.1| pol protein [Human immunodeficiency virus 1] >gb|AA	
BAF34642.1	pol polyprotein [HIV-1 vector pNL-DT5R]	
ACY01941.1	pol protein [HIV whole-genome vector AA1305#18]	
1Q9P_A	Chain A, Solution Structure Of The Mature Hiv-1 Protease Monomer	S
3FSM_A	Chain A, Crystal Structure Of A Chemically Synthesized 203 Amino Acid Dimer' [I-Ala51,D-Ala51'] Hiv-1 Protease Molecule	S
AAV69858.1	Pol polyprotein [SIV vector pCLN8] >gb|AAA91931.1| pol polyprotein [Simian immunodeficiency virus]	
ABY60459.1	Gag-Pol-Nef protein [Expression vector MVA--89.6P-SIVGPN]	

- There are two major sections to the Multiple Sequence Alignment page
  - The top part is simply a list of descriptions and checkboxes.
  - Based on the description, uncheck some of the boxes to the left. This removes that sequence from the multiple sequence alignment
- This section allows you to eliminate any sequences that you would not have wanted based on the wrong information
  - Wrong organism, fragment, etc.

The second half of the MSA page

- Here we get to see the actual Multiple sequence alignment.
- Here you might want to eliminate a protein based on some other reason:
  - The way the protein got aligned is messing up the larger alignment



- Long unaligned tails make the start or end of the sequences align a little arbitrarily

Next time: finding the active site from the alignment

Questions