Introduction to Phylogenetics

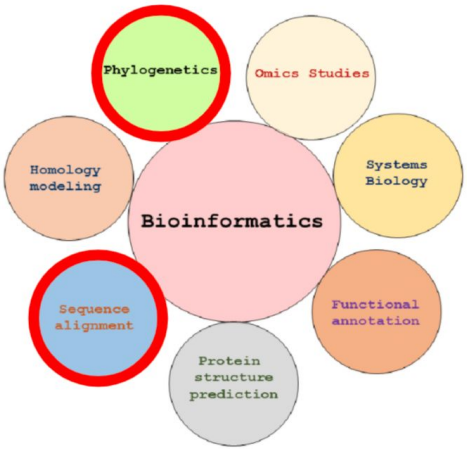
Emilie Skoog

Summer 2021

Recall...

What is bioinformatics?

- Combines biology and computer science to **interpret biological data**



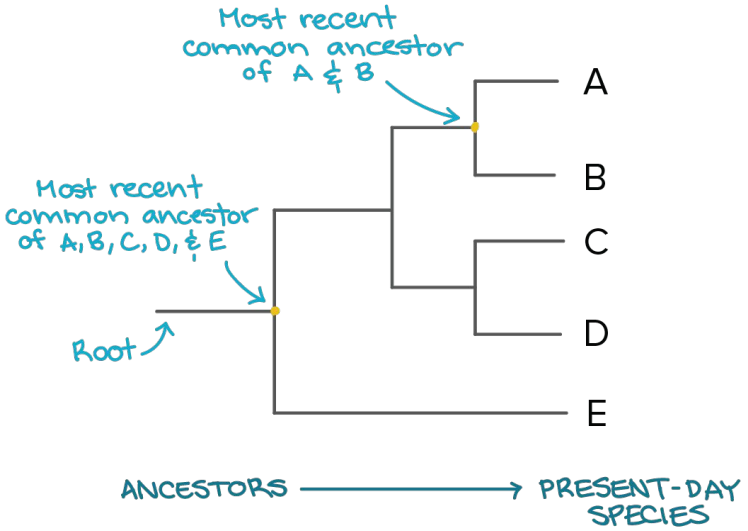
Usually DNA and amino acid sequences!

Phylogenetics is an important component of many bioinformatic studies.

What is phylogenetics?

- **Phylogenetics** is the study of evolutionary relationships among biological entities

- A **phylogenetic tree** is often made to visualize these genetic comparisons or infer evolutionary relationships.



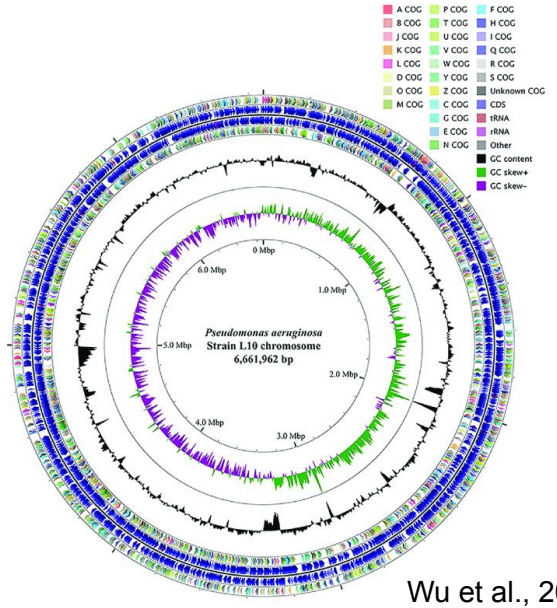
Phylogenetics vs. Phylogenomics

Key terms

genome: the whole genetic makeup of an organism (as opposed to a single gene)

- **Phylogenetics** compares and analyzes the sequences of single genes, or a small number of **genes**.

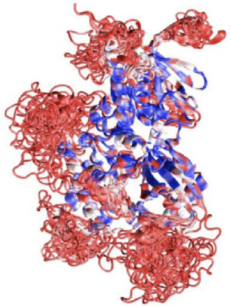
- **Phylogenomics** draws information by comparing enti genomes, or at least large portions of **genomes**.

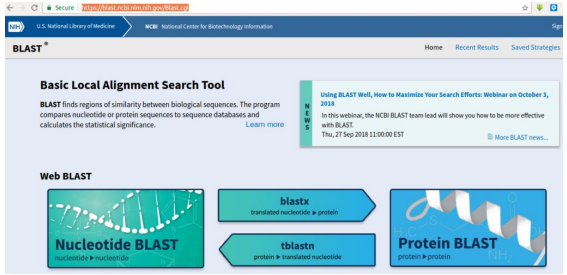


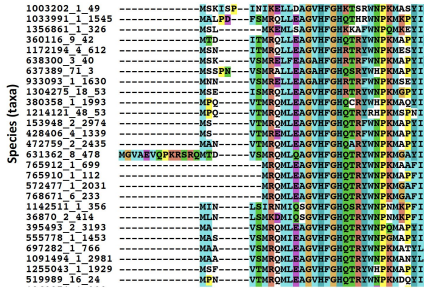
Wu et al., 2018

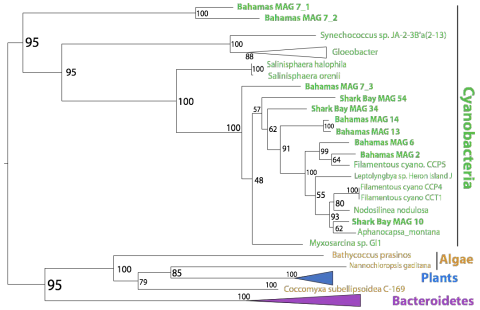
Overview of usual steps in creating a phylogenetic tree

STEPS:









Identify amino acid (protein) or nucleotide (DNA) sequence of interest

Homology search using BLAST

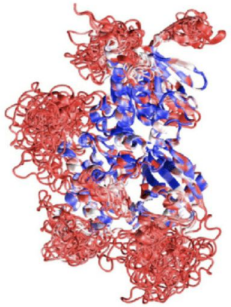
Multiple sequence alignment

Generate phylogenetic tree

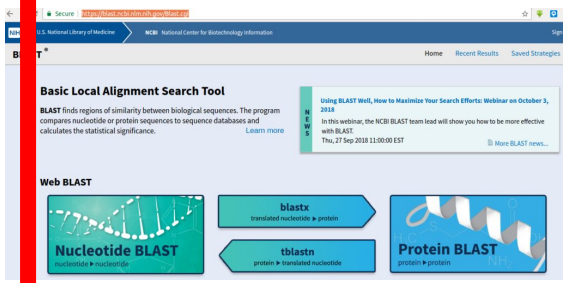
Let’s dive right into it (might make the most sense learning while doing it together and I’ll explain as we go along)

Overview of usual steps in creating a phylogenetic tree

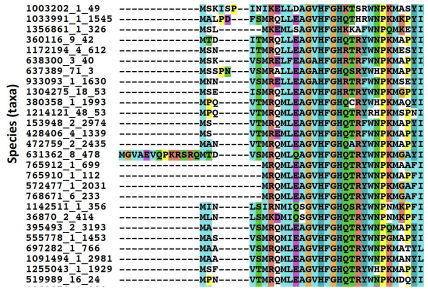
STEPS:



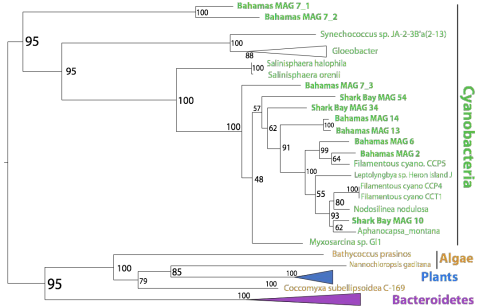
Identify amino acid (protein) or nucleotide (DNA) sequence of interest



Homology search using BLAST



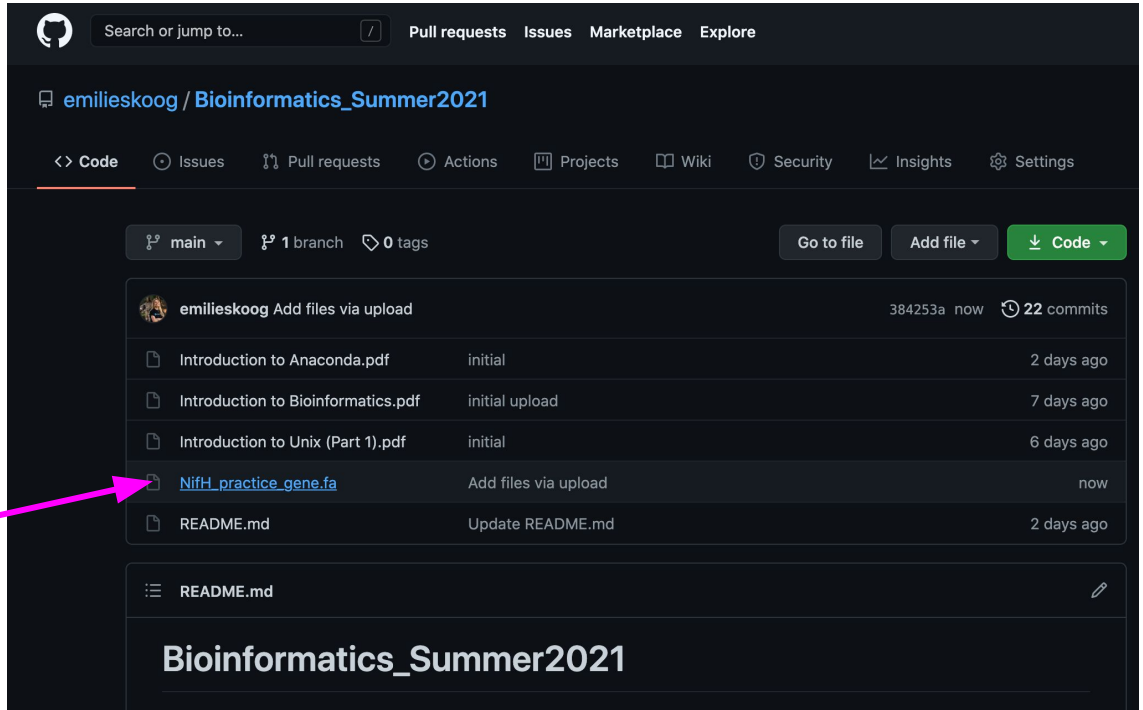
Multiple sequence alignment



Generate phylogenetic tree

First: Acquire a gene of interest (provided)

- Go to our github repository and download protein **query sequence**



Key terms

Query sequence: A DNA or protein sequence submitted to a computerized database for comparison, e.g., a BLAST search. Aka whatever your gene of interest is that you are using for your homology searches

NifH: A nitrogenase gene (role: fixation of atmospheric nitrogen)

You will see a “FASTA file”:

```
>CAA83510.1 NifH [Nostoc sp. PCC 6720]
MTDENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLHAKAKTTVLHLAAERGA
VEDLELHEVMLTGFRGVRCVESGGPEPGVGCAGRGIITAINFLEENGAYQDLDFVSYDVLGDVVCGGFAM
PIREGKAQEIYIVTSGEMMAMYAANNIARGILKYAHSGGVRLGGLICNSRKTDREAELIENLAERLNTQM
IHFVPRDNIVQHAELRRMTVNEYAPDSNQGQEYRALAKKIINNDKLTIPTPIEMDELEALLIEYGILDDD
SKHAEIIGKPAEATK
```

You will see a “FASTA file”:

```
>CAA83510.1 NifH [Nostoc sp. PCC 6720]
MTDENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLHAKAKTTVLHLAAERGA
VEDLELHEVMLTGFRGVRCVESGGPEPGVGCAGRGIITAINFLEENGAYQDLDFVSYDVLGDVVCGGFAM
PIREGKAQEIYIVTSGEMMAMYAANNIARGILKYAHSGGVRLGGLICNSRKTDREAELIENLAERLNTQM
IHFVPRDNIVQHAELRRMTVNEYAPDSNQGQEYRALAKKIINNDKLTIPTPIEMDELEALLIEYGILDDD
SKHAEIIGKPAEATK
```

- In bioinformatics, the FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes.
- The format also allows for sequence names and comments to precede the sequences.

Types of FASTA file extensions:

- Generic FASTA: .fasta, .fas, .fa, .seq, .fsa
- FASTA nucleic acids: .fna
- FASTA nucleotide coding regions for a genome: .ffn
- FASTA amino acids: .faa

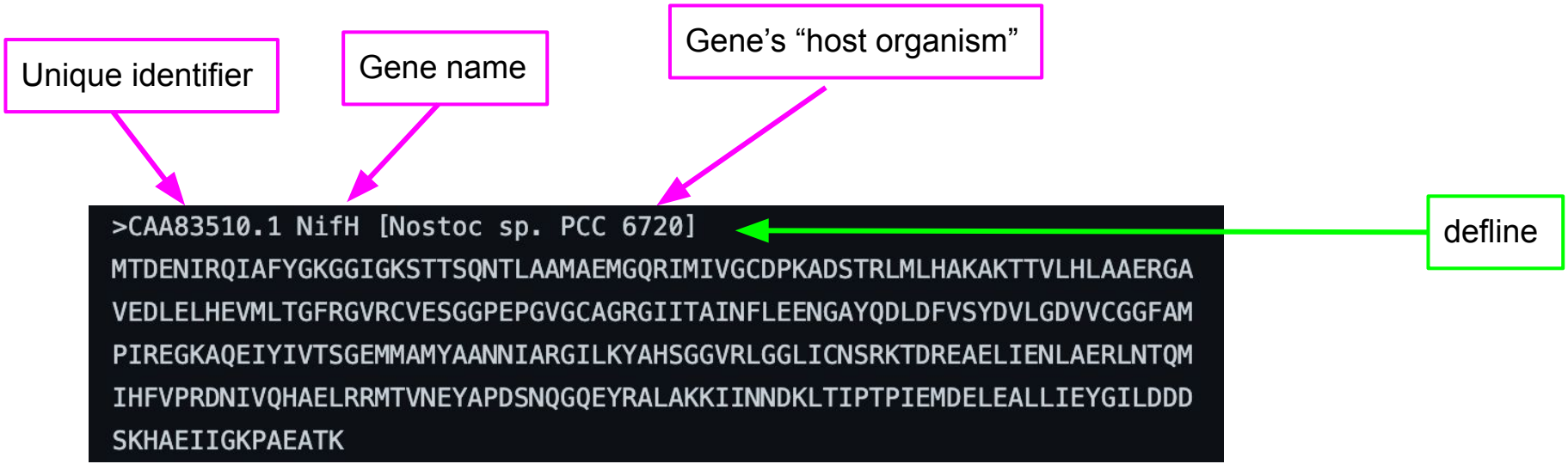
Highlighted are the most common types (at least that I see)

Second: Copy query sequence

- Copy protein query sequence (don't include **defline**)

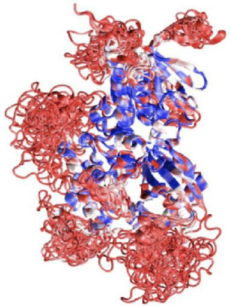
Key terms

Defline: The description line or header/identifier line, which begins with '>', gives a name and/or a unique identifier for the sequence, and may also contain additional information.

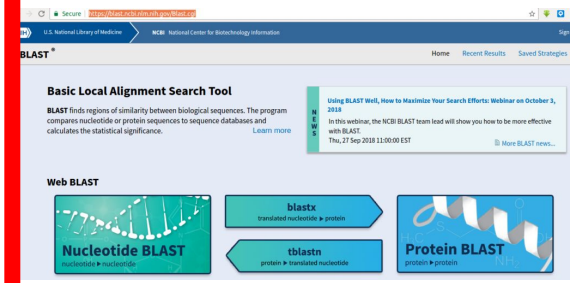


Overview of usual steps in creating a phylogenetic tree

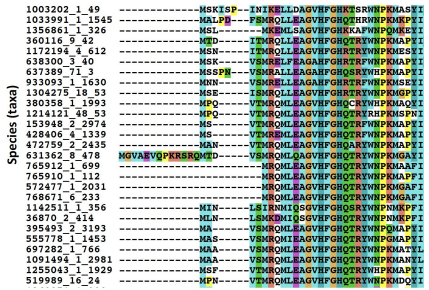
STEPS:



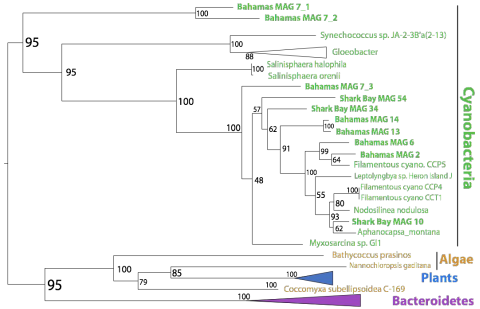
Identify amino acid (protein) or nucleotide (DNA) sequence of interest



Homology search using BLAST

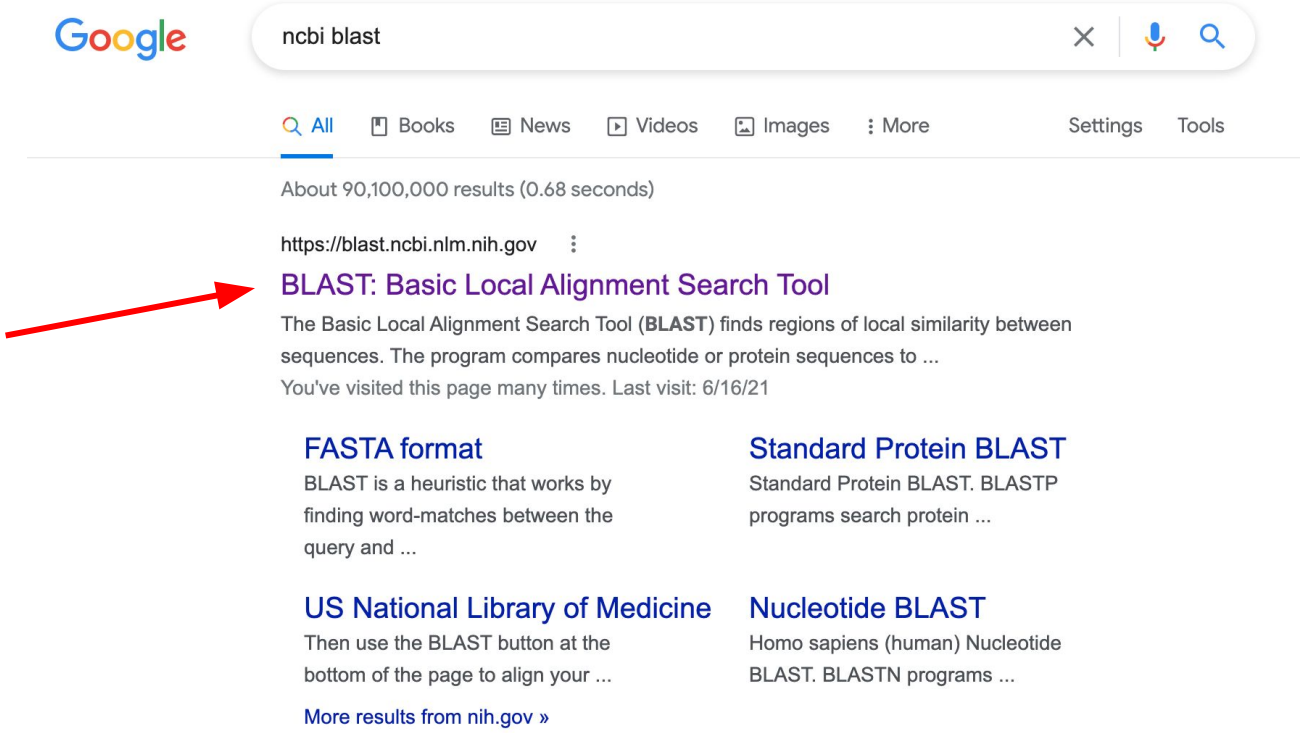


Multiple sequence alignment



Generate phylogenetic tree

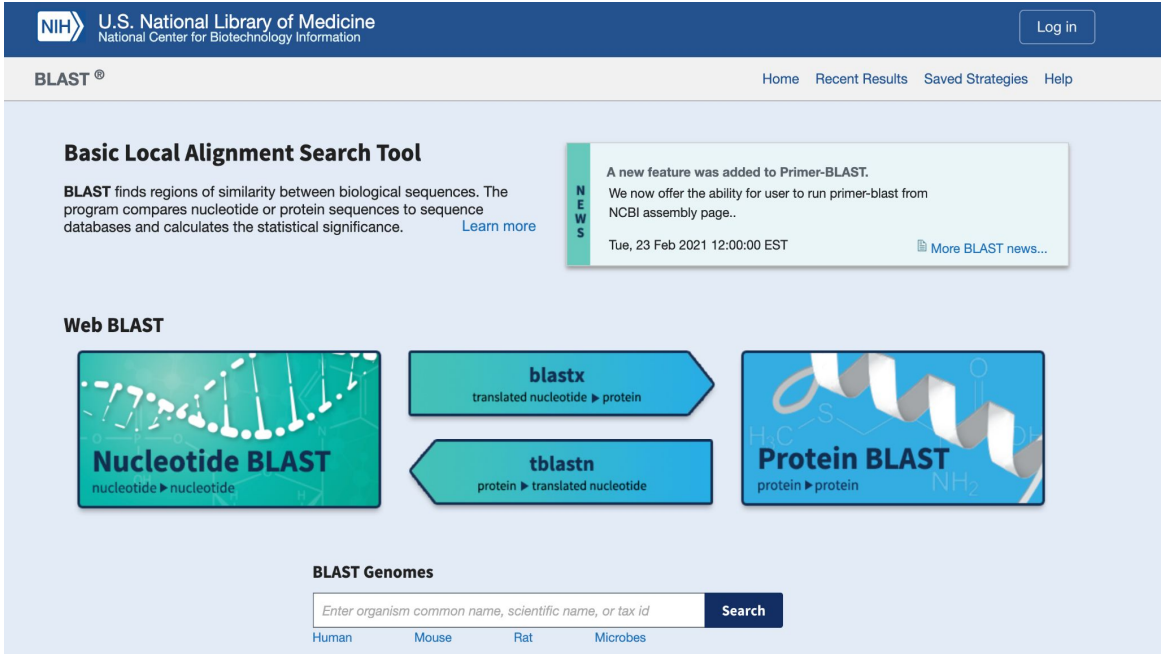
Go to NCBI BLAST



BLAST: Basic Local Alignment Search Tool

- BLAST is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences.

- Pairwise alignment (one type of BLAST) aligns two sequences based on the BLAST algorithm.



BLAST: Basic Local Alignment Search Tool

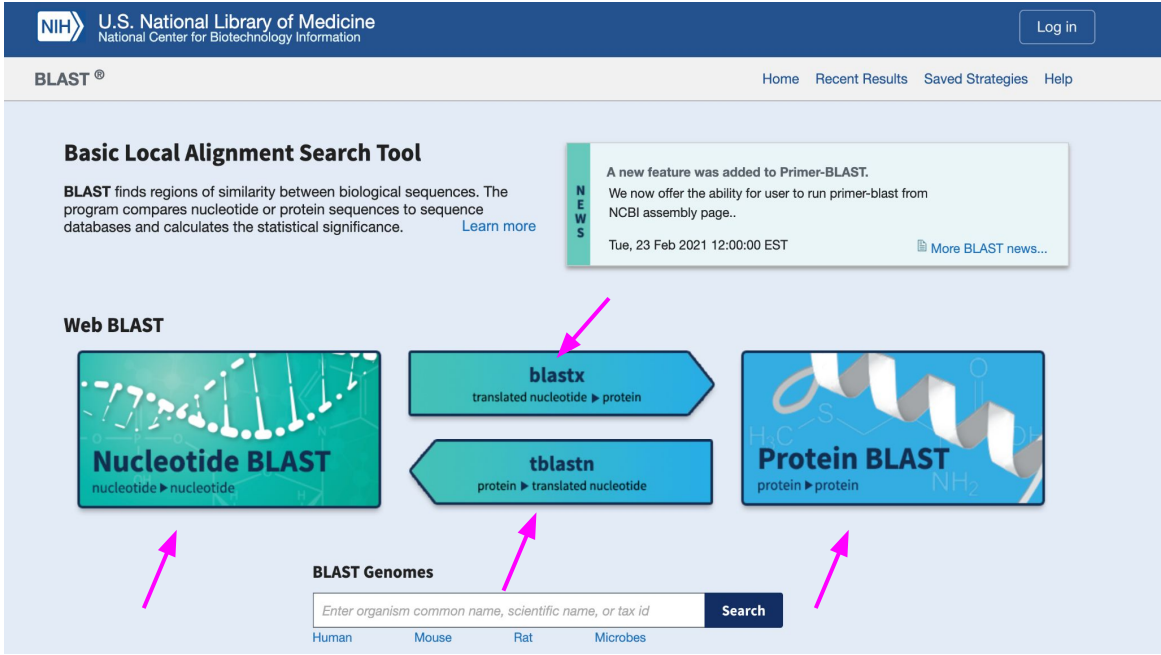
blastn (nucleotide blast): BLAST nucleotide sequence against nucleotide query databases

blastp (protein blast): BLAST protein sequence against protein database

blastx: search protein database using translated nucleotide query

tblastn: search translated nucleotide database using protein query

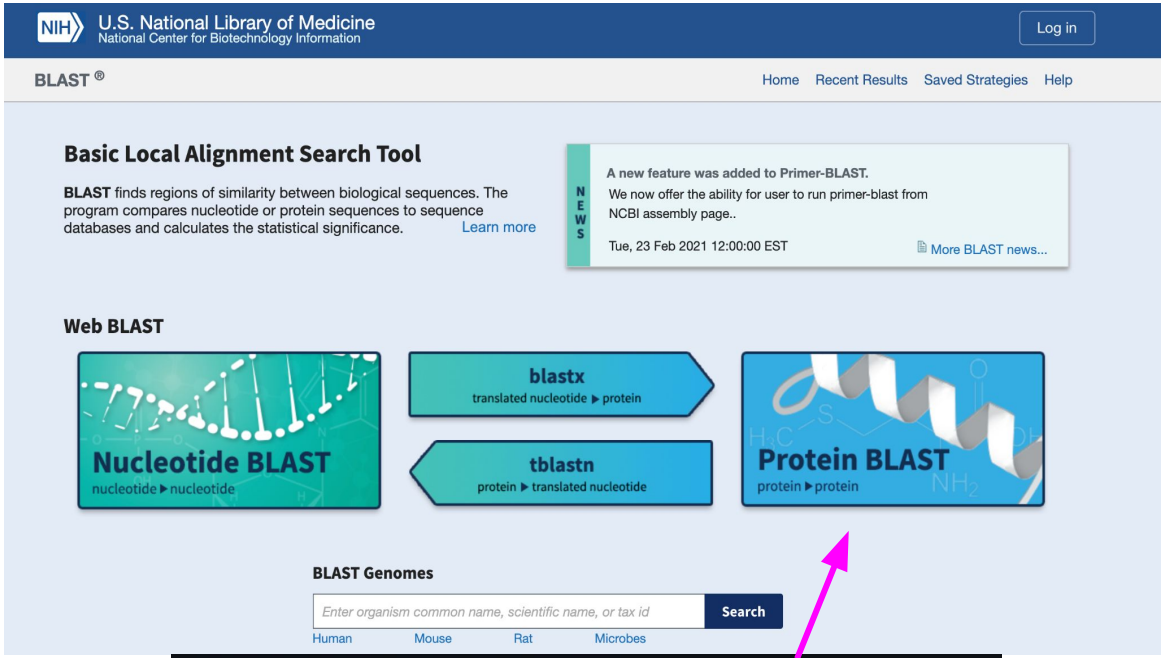
tblastx: search translated nucleotide database using translated nucleotide query



Verbiage taken from Greg Fournier

Search for homologous sequence using NCBI BLAST

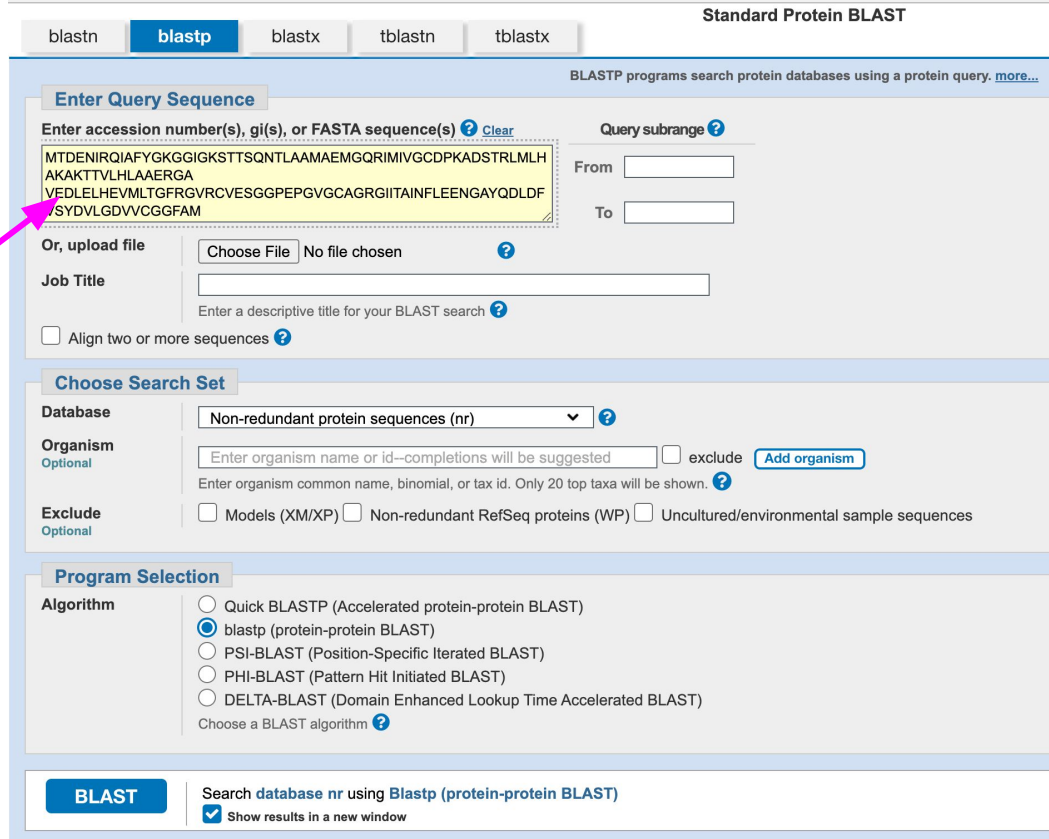
1. Click on ‘protein blast’ (blastp)




```
>CAA83510.1 NifH [Nostoc sp. PCC 6720]
MTDENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLHAKAKTTVLHLAAERGA
VEDLELHEVMLTGFRGVRCVESGGPEPGVGCAGRGIITAINFLEENGAYQDLDFVSYDVLGDVVCGGFAM
PIREGKAQEIYIVTSGEMMAMYAANNIARGILKYAHSGGVRLGGLICNSRKTDREAELIENLAERLNTQM
IHFVPRDNIVQHAELRRMTVNEYAPDSNQGQEYRALAKKIINNDKLTIPTPIEMDELEALLIEYGILDDD
SKHAEIIGKPAEATK
```

Search for homologous sequence using NCBI BLAST

1. Click on ‘protein blast’ (blastp)
2. Paste NifH protein query where it asks you to ‘enter query sequence’



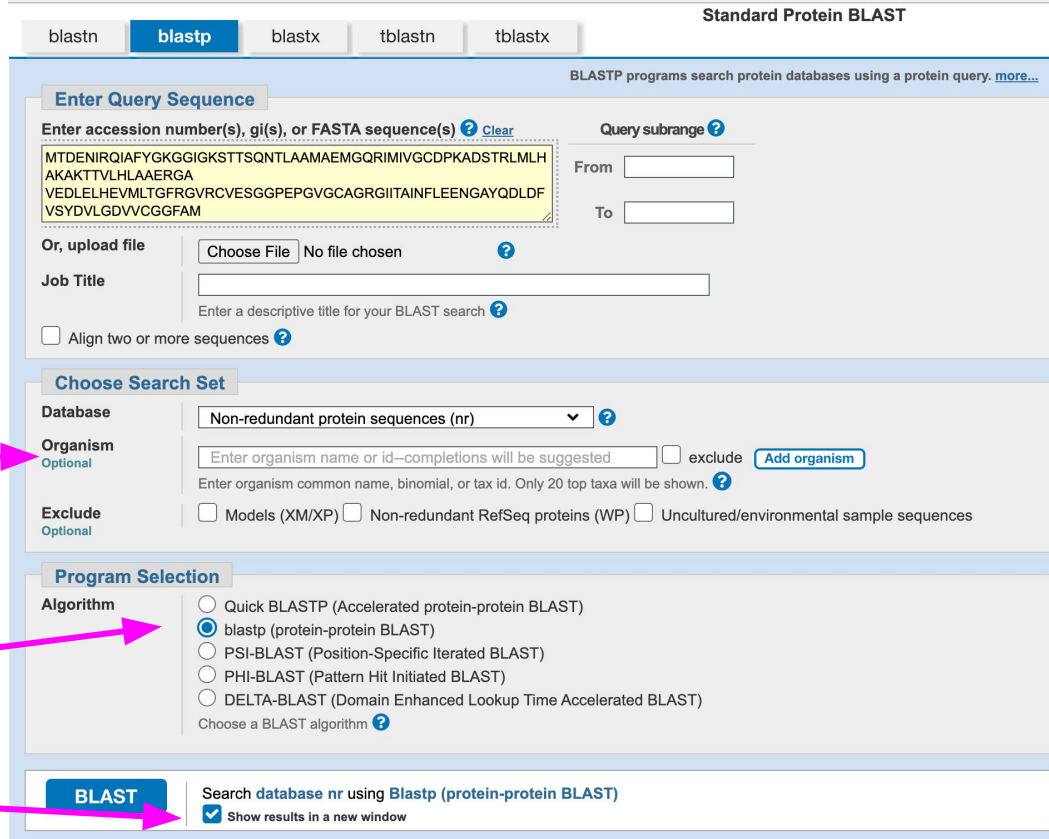
Search for homologous sequence using NCBI BLAST

1. Click on ‘protein blast’ (blastp)
2. Paste NifH protein query where it asks you to ‘enter query sequence’
3. You may change some basic parameters

Here we can actually select for specific taxa we want to target or exclude. Ex: we only want hits to virus sequences

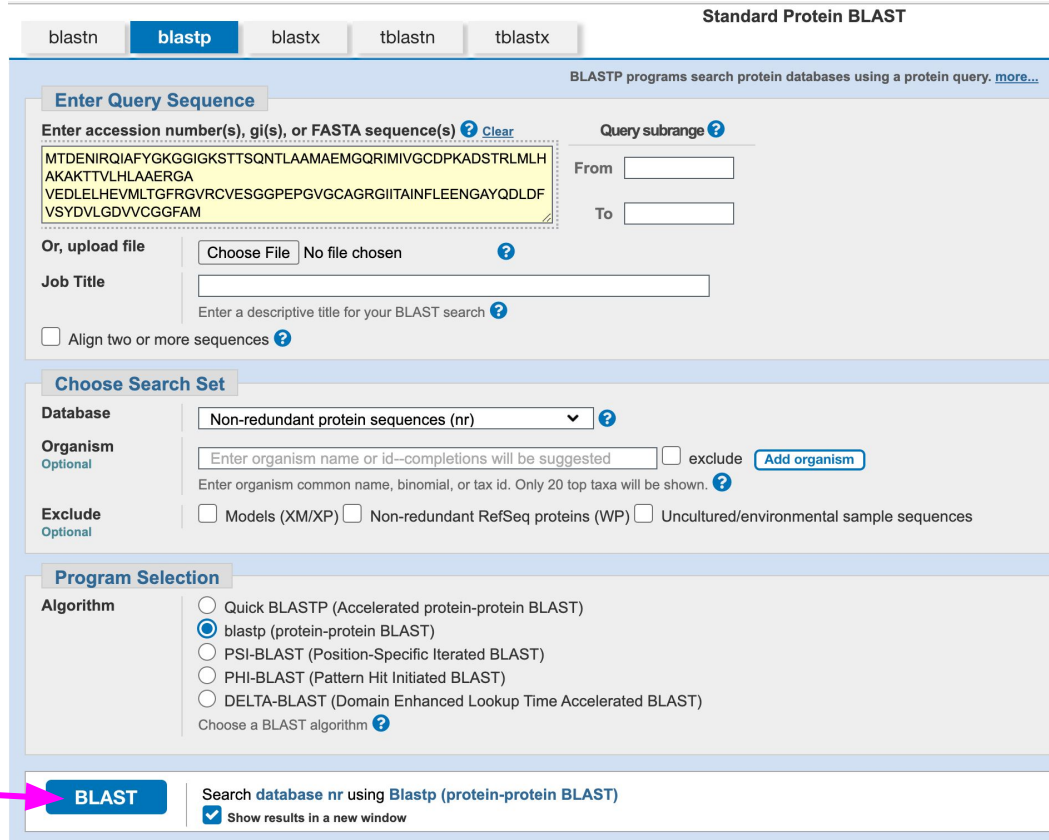
Notice here we have blastp selected

I always select to show results in a new window



Search for homologous sequence using NCBI BLAST

1. Click on ‘protein blast’ (blastp)
2. Paste NifH protein query where it asks you to ‘enter query sequence’
3. You may change some basic parameters
4. ‘Blast’ it! (not technically a verb)



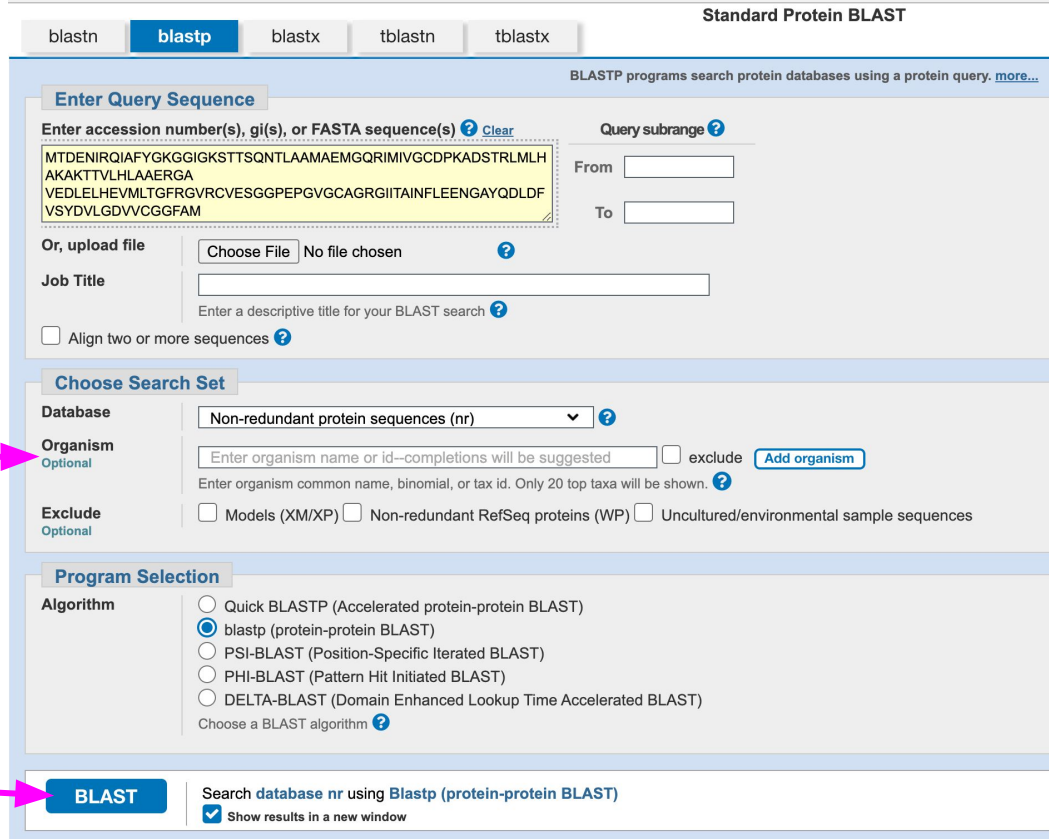
Click ‘BLAST’

Search for homologous sequence using NCBI BLAST

1. Click on ‘protein blast’ (blastp)
2. Paste NifH protein query where it asks you to ‘enter query sequence’
3. You may change some basic parameters

Try changing this parameter to exclude Nostoc or only include viral hits or whatever you are interested in

And hit BLAST again



Our results:

< Edit Search | Save Search | Search Summary | How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	**Protein Sequence**
RID	CNDBG98Y013 *Search expires on 06-18 10:37 am* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_68648
Description	None
Molecule type	amino acid
Query Length	295
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to **E value** to **Query Coverage** to

Filter Reset

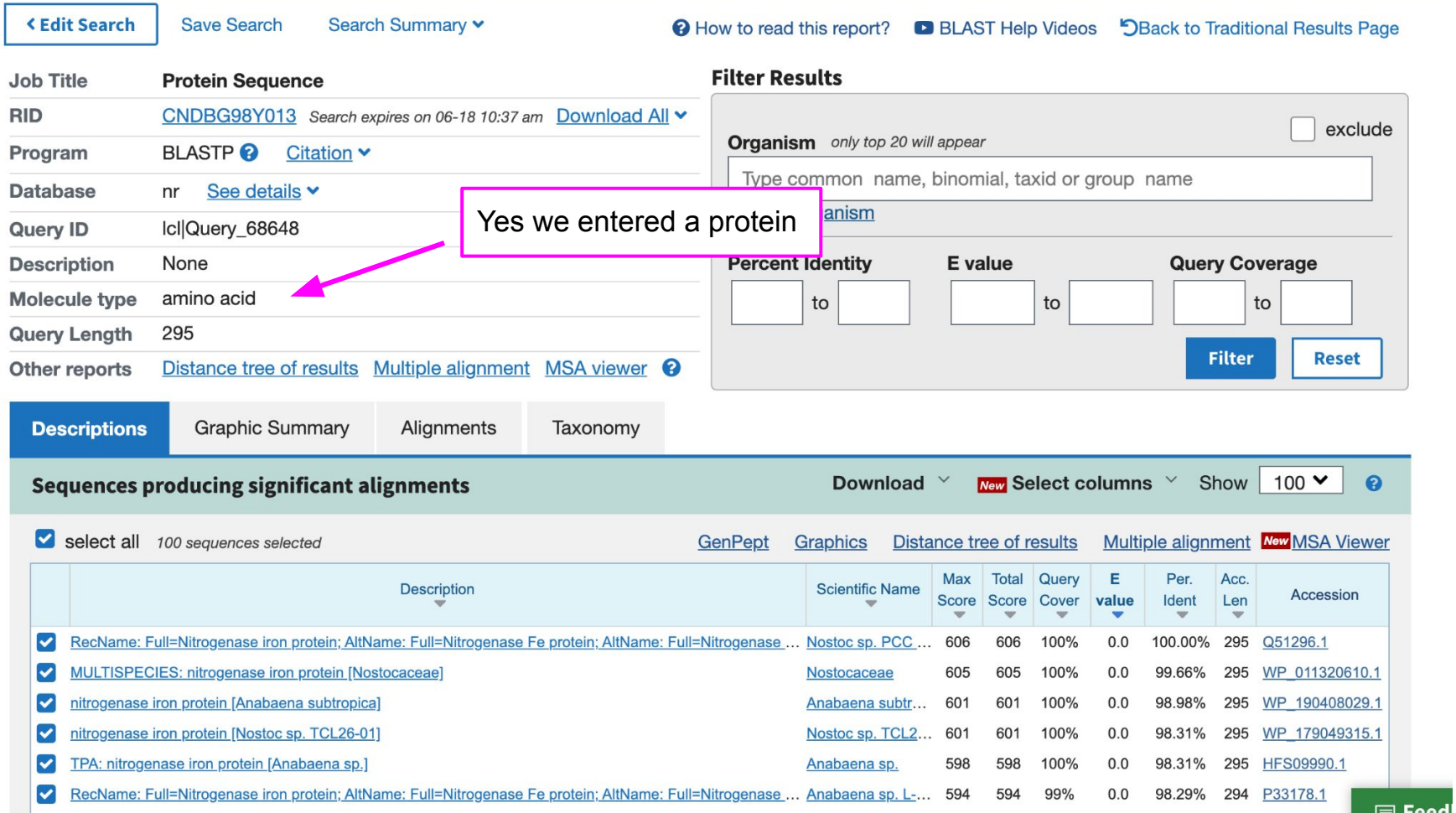
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download New Select columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase …	Nostoc sp. PCC …	606	606	100%	0.0	100.00%	295	Q51296.1
MULTISPECIES: nitrogenase iron protein [Nostocaceae]	Nostocaceae	605	605	100%	0.0	99.66%	295	WP_011320610.1
nitrogenase iron protein [Anabaena subtropica]	Anabaena subtr…	601	601	100%	0.0	98.98%	295	WP_190408029.1
nitrogenase iron protein [Nostoc sp. TCL26-01]	Nostoc sp. TCL2…	601	601	100%	0.0	98.31%	295	WP_179049315.1
TPA: nitrogenase iron protein [Anabaena sp.]	Anabaena sp.	598	598	100%	0.0	98.31%	295	HFS09990.1
RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase …	Anabaena sp. L-…	594	594	99%	0.0	98.29%	294	P33178.1

Our results:



Edit Search | Save Search | Search Summary | How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	Protein Sequence
RID	CNDBG98Y013 *Search expires on 06-18 10:37 am* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_68648
Description	None
Molecule type	amino acid
Query Length	295
Other reports	Distance tree of results Multiple alignment MSA viewer

Yes we entered a protein

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

anism

Percent Identity to | E value to | Query Coverage to

Filter Reset

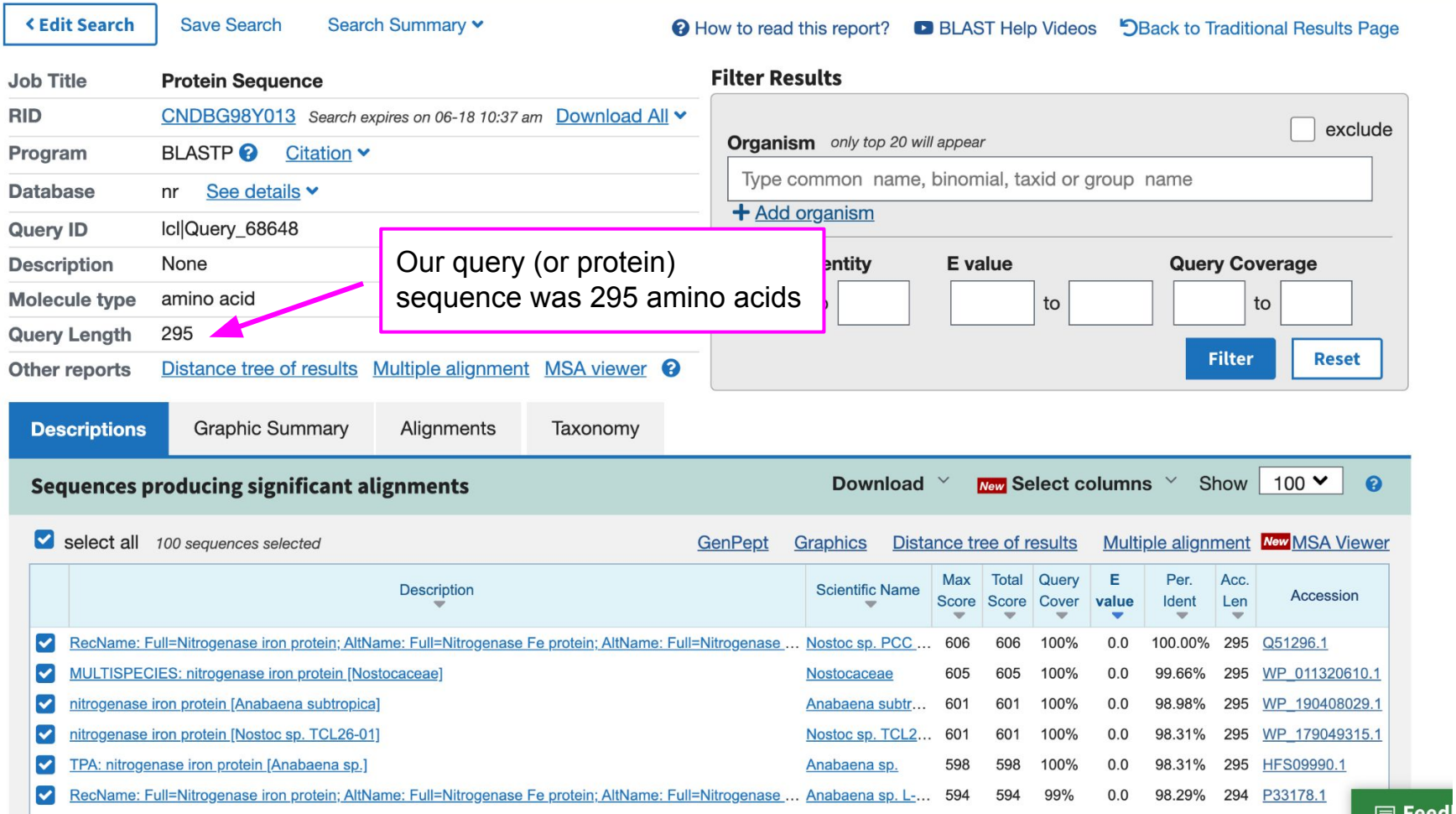
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download | New Select columns | Show 100

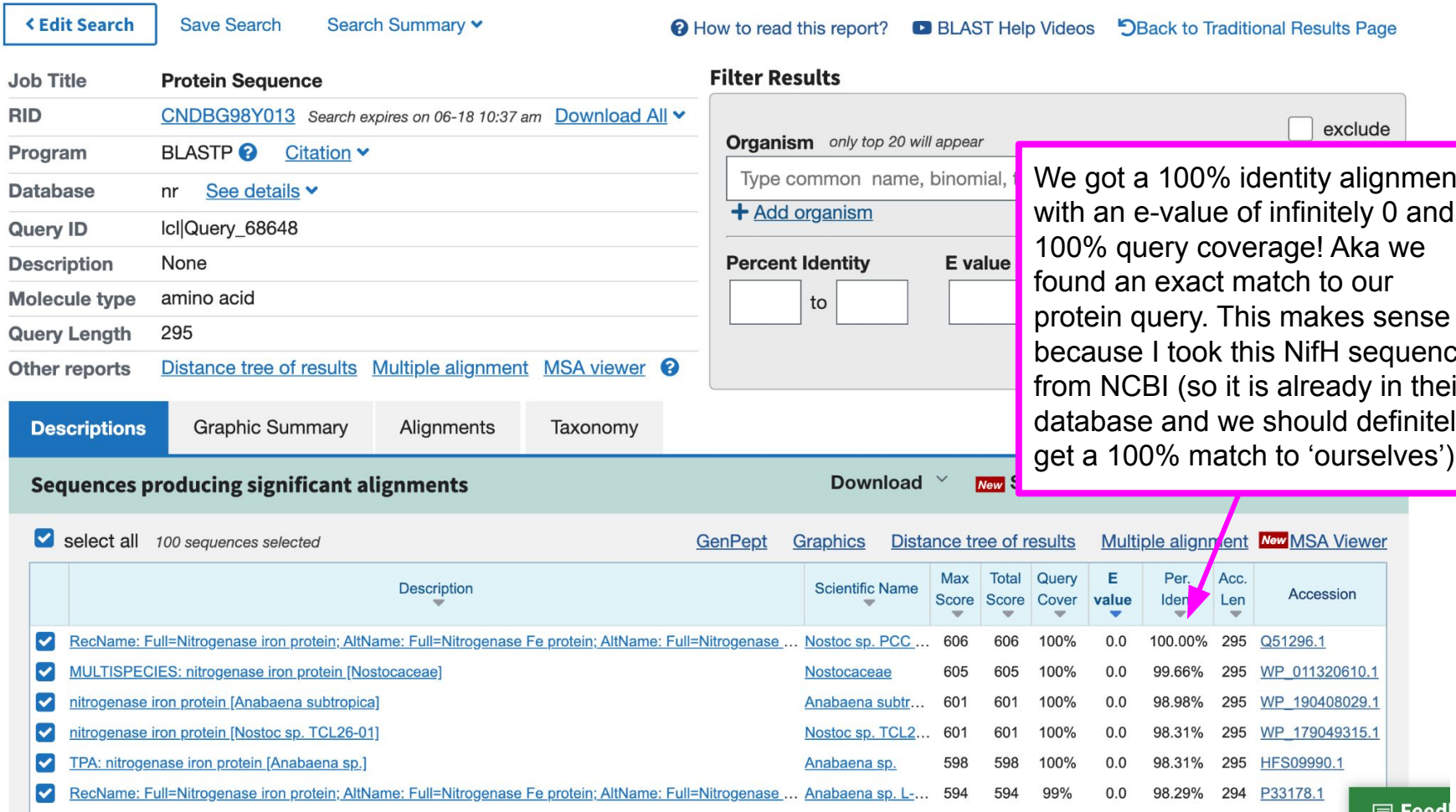
select all *100 sequences selected* GenPept | Graphics | Distance tree of results | Multiple alignment | New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase …	Nostoc sp. PCC …	606	606	100%	0.0	100.00%	295	Q51296.1
MULTISPECIES: nitrogenase iron protein [Nostocaceae]	Nostocaceae	605	605	100%	0.0	99.66%	295	WP_011320610.1
nitrogenase iron protein [Anabaena subtropica]	Anabaena subtr…	601	601	100%	0.0	98.98%	295	WP_190408029.1
nitrogenase iron protein [Nostoc sp. TCL26-01]	Nostoc sp. TCL2…	601	601	100%	0.0	98.31%	295	WP_179049315.1
TPA: nitrogenase iron protein [Anabaena sp.]	Anabaena sp.	598	598	100%	0.0	98.31%	295	HFS09990.1
RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase …	Anabaena sp. L-…	594	594	99%	0.0	98.29%	294	P33178.1

Our results:


Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title Protein Sequence
RID CNDBG98Y013 Search expires on 06-18 10:37 am Download All
Program BLASTP Citation
Database nr See details
Query ID lcl|Query_68648
Description None
Molecule type amino acid
Query Length 295
Other reports Distance tree of results Multiple alignment MSA viewer
Our query (or protein) sequence was 295 amino acids
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
select all 100 sequences selected
GenPept
Graphics
Distance tree of results
Multiple alignment
MSA Viewer
Description
Scientific Name
Max Score
Total Score
Query Cover
E value
Per. Ident
Acc. Len
Accession
RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase … Nostoc sp. PCC … 606 606 100% 0.0 100.00% 295 Q51296.1
MULTISPECIES: nitrogenase iron protein [Nostocaceae] Nostocaceae 605 605 100% 0.0 99.66% 295 WP_011320610.1
nitrogenase iron protein [Anabaena subtropica] Anabaena subtr… 601 601 100% 0.0 98.98% 295 WP_190408029.1
nitrogenase iron protein [Nostoc sp. TCL26-01] Nostoc sp. TCL2… 601 601 100% 0.0 98.31% 295 WP_179049315.1
TPA: nitrogenase iron protein [Anabaena sp.] Anabaena sp. 598 598 100% 0.0 98.31% 295 HFS09990.1
RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase … Anabaena sp. L-… 594 594 99% 0.0 98.29% 294 P33178.1

Our results:



We got a 100% identity alignment with an e-value of infinitely 0 and 100% query coverage! Aka we found an exact match to our protein query. This makes sense because I took this NifH sequence from NCBI (so it is already in their database and we should definitely get a 100% match to 'ourselves').

E-value, percent identity, and query cover

E-value: The E value (expected value) is a number that describes how many times you would expect a match by chance in a database of that size. The lower the E value is, the more significant the match.

Percent Identity: The percent identity is a number that describes how similar the query sequence is to the target sequence (how many characters in each sequence are identical). The higher the percent identity is, the more significant the match.

Query Cover: The query cover is a number that describes how much of the query sequence is covered by the target sequence. If the target sequence in the database spans the whole query sequence, then the query cover is 100%. This tells us how long the sequences are, relative to each other.

E-value, percent identity, and query cover

E-value: The E value (expected value) is a number that describes how many times you would expect a match by chance in a database of that size. The lower the E value is, the more significant the match.

Percent Identity: The percent identity is a number that describes how similar the query sequence is to the target sequence (how many characters in each sequence are identical). The higher the percent identity is, the more significant the match.

Query Cover: The query cover is a number that describes how much of the query sequence is covered by the target sequence. If the target sequence in the database spans the whole query sequence, then the query cover is 100%. This tells us how long the sequences are, relative to each other.

Percent identity: 100%

Query cover: 50%

E-value, percent identity, and query cover

E-value: The E value (expected value) is a number that describes how many times you would expect a match by chance in a database of that size. The lower the E value is, the more significant the match.

Percent Identity: The percent identity is a number that describes how similar the query sequence is to the target sequence (how many characters in each sequence are identical). The higher the percent identity is, the more significant the match.

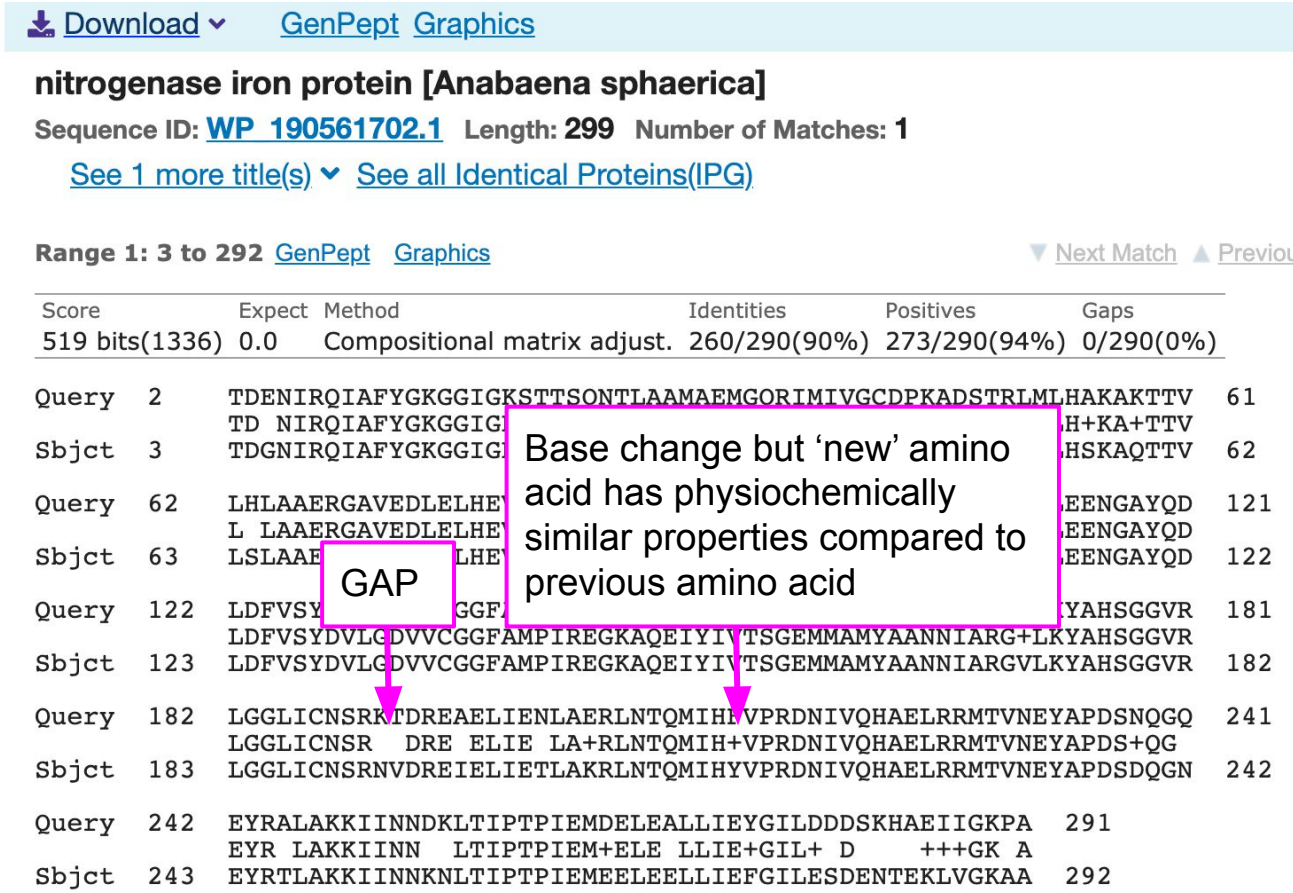
Query Cover: The query cover is a number that describes how much of the query sequence is covered by the target sequence. If the target sequence in the database spans the whole query sequence, then the query cover is 100%. This tells us how long the sequences are, relative to each other.

Very important to consider more than just percent identity!

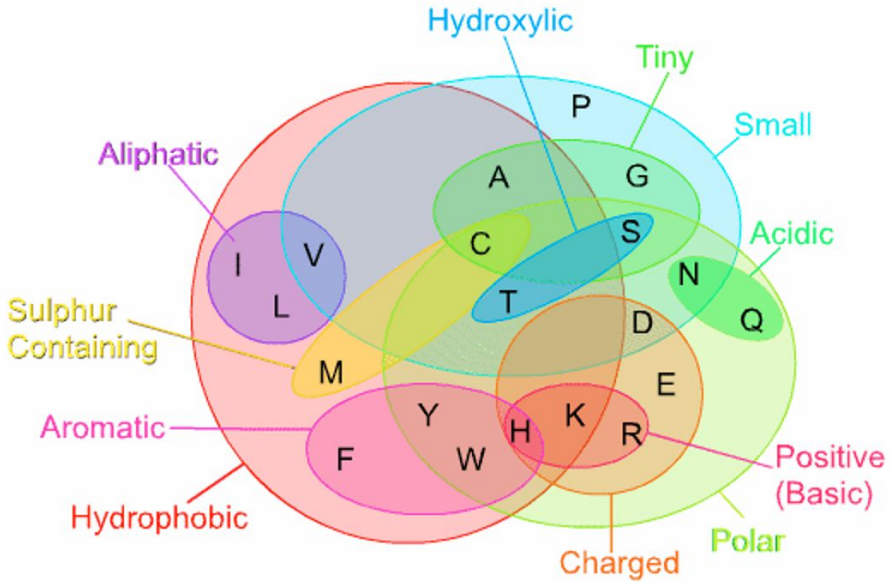
Percent identity: 100%

Query cover: 50%

Reminder: This is a pairwise alignment (aligns and compares 2 sequences - your query sequence and the hit)



‘Physiochemically similar’



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

Reminder: This is a pairwise alignment (aligns and compares 2 sequences - your query sequence and the hit)

Download GenPept Graphics

nitrogenase iron protein [Anabaena sphaerica]

Sequence ID: WP_190561702.1 **Length:** 299 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 3 to 292 GenPept Graphics Next Match Previou

Score	Expect	Method	Identities	Positives	Gaps
519 bits(1336)	0.0	Compositional matrix adjust.	260/290(90%)	273/290(94%)	0/290(0%)

```
Query  2    TDENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLHAKAKTTV  61
            TD NIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLH+KA+TTV
Sbjct  3    TDGNIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLHSKAQTTV  62

Query  62   LHLAAERGAVEDLELHEVMLTGFRGVRCVESGGPEPGVGCAGRGIITAINFLEENGAYQD  121
            L LAAERGAVEDLELHEVMLTGFRGV+CVESGGPEPGVGCAGRGIITAINFLEENGAYQD
Sbjct  63   LSLAAERGAVEDLELHEVMLTGFRGVKCVESGGPEPGVGCAGRGIITAINFLEENGAYQD  122
```

Same concept as the colorful alignments we saw in previous lessons. Ex:

```
472759_2_2435  ----------MAN-----VTMRQMLEAGVHFGHQARYWNPKMAPYIFGKRNNIHIINLEETLPLYNEATNFLGRI
631362_8_478   MGVAEVQPKRSRQMTD-----VSMRQMLQAGVHFGHQTRYWNPKMGAYIFGHRNKIHIINLERTLPLFQDAMNYLGKI
```

We notice that our viral hit results have lower values and worse alignments

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☐	TPA: MAG TPA: ParA [Caudovirales sp.]	Caudovirales sp.	47.4	47.4	80%	9e-04	20.32%	242	DAF22519.1
☐	AAA family ATPase [Enterococcus phage EF62phi]	Enterococcus phage EF62phi	42.7	42.7	25%	0.025	28.75%	267	YP_006218698.1
☐	TPA: MAG TPA: ParA [Myoviridae sp.]	Myoviridae sp.	42.4	42.4	24%	0.032	33.33%	256	DAE74853.1

Note regions of conservation and their importance

Physiochemically similar, so potentially similar protein function as a result of similar protein folding

TPA: MAG TPA: ParA [Caudovirales sp.]

Sequence ID: DAF22519.1 **Length:** 242 **Number of Matches:** 1

Range 1: 10 to 239 GenPept Graphics ▼ Next Match ▲ Previous

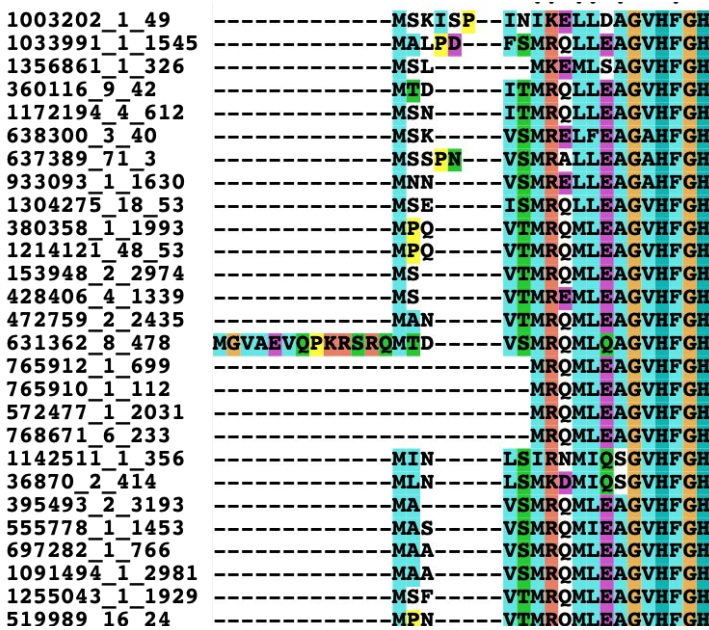
Score	Expect	Method	Identities	Positives	Gaps
47.4 bits(111)	9e-04	Compositional matrix adjust.	51/251(20%)	114/251(45%)	34/251(13%)

```
Query  14   KGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLHAKAKTTVLHLAAERGAVED  73
            KGG+GK+TT+QN  A +     ++++++  D + + T L     K K  +L++
Sbjct  10   KGGVGKTTTAQNLTAGLRLQNKKVLLLDLDAQCNLTLLQQATKYKYNILNVLK-------  62

Query  74   LELHEVMLTGFRGVRCVESGGPEPGVGCAGRGIITAINFL--EENGAYQDLDFVSYDV--  129
               +EV +        VE            G       +N L  E       D D++  D
Sbjct  63   ---NEVDINT-----AVEKDFIAGSKFLVGENTEIELNKLKNELQKLKTDYDYIIIDTPP  114

Query  130  -LGDVVCGGFAMPIREGKAQEIYIVTSGEMMAMYAANNIARGI--LKYAHSGGVRLGGLI  186
             L ++             + +I I  + +++ +    ++ + +  ++   +  + + G++
Sbjct  115  ALSNITINALTA------STDIIITITADLLPIQGLVDLYKTVQSIQKTSNKNLNIKGIL  168

Query  187  CN--SRKTDREAELIENL---AERLNTQMIHFVPRDNI-VQHAELRRMTVNEYAPDSNQG  240
                 +++T     ++ +L   AE+LNT+++   RD+I ++ ++ +   + +YA  S  G
Sbjct  169  VTRFNKRTILGRTMLNSLIDIAEKLNTKVLNTKIRDSISIKESQAKMTDIFKYARYSTAG  228

Query  241  QEYRALAKKII  251
             ++YRAL K+I+
Sbjct  229  RDYRALVKEIL  239
```

Pairwise alignment vs multiple sequence alignment

- BLAST is pairwise alignment (limited to comparing 2 sequences)
- A multiple sequence alignment compares well… multiple sequences (<u>necessary</u> for building phylogenetic trees)

vs

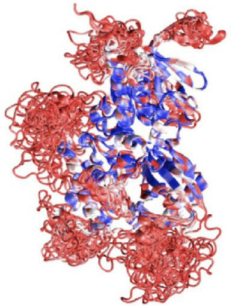


Now you may be asking yourself, why did we bother using a pairwise alignment and ‘blast’ our query sequence? And to this I respond:

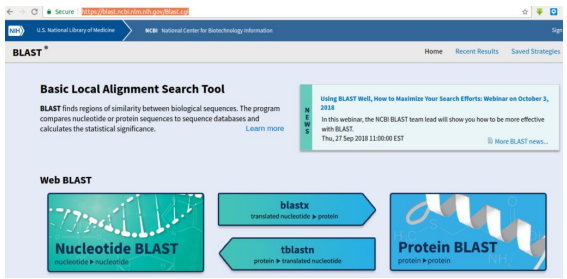
- Now we know what organisms have the closest gene hit and what other organisms may have this gene!
- We can see how similar these genes are to one another.
- If we changed parameters and excluded cyanobacteria (Nostoc was our ‘sequence host’), and got strong hits to viruses for example, we can infer some viral-mediated horizontal gene transfer.
- WE CAN ALSO GATHER OTHER SEQUENCES FOR A MULTIPLE SEQUENCE ALIGNMENT!
    - We will download the fasta files of whichever hits we want and we will create a multiple sequence alignment to make a phylogenetic tree.

Overview of usual steps in creating a phylogenetic tree

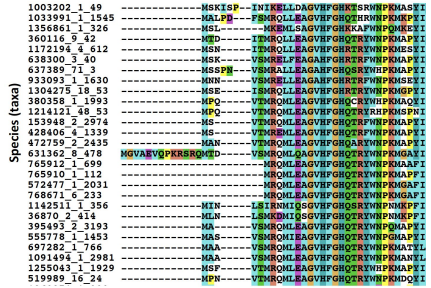
STEPS:



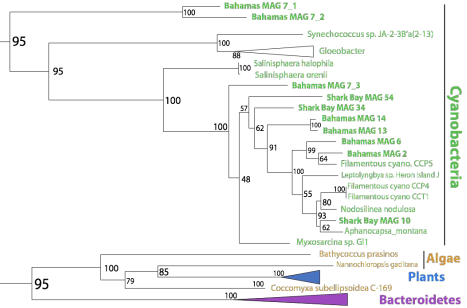
Identify amino acid (protein) or nucleotide (DNA) sequence of interest



Homology search using BLAST



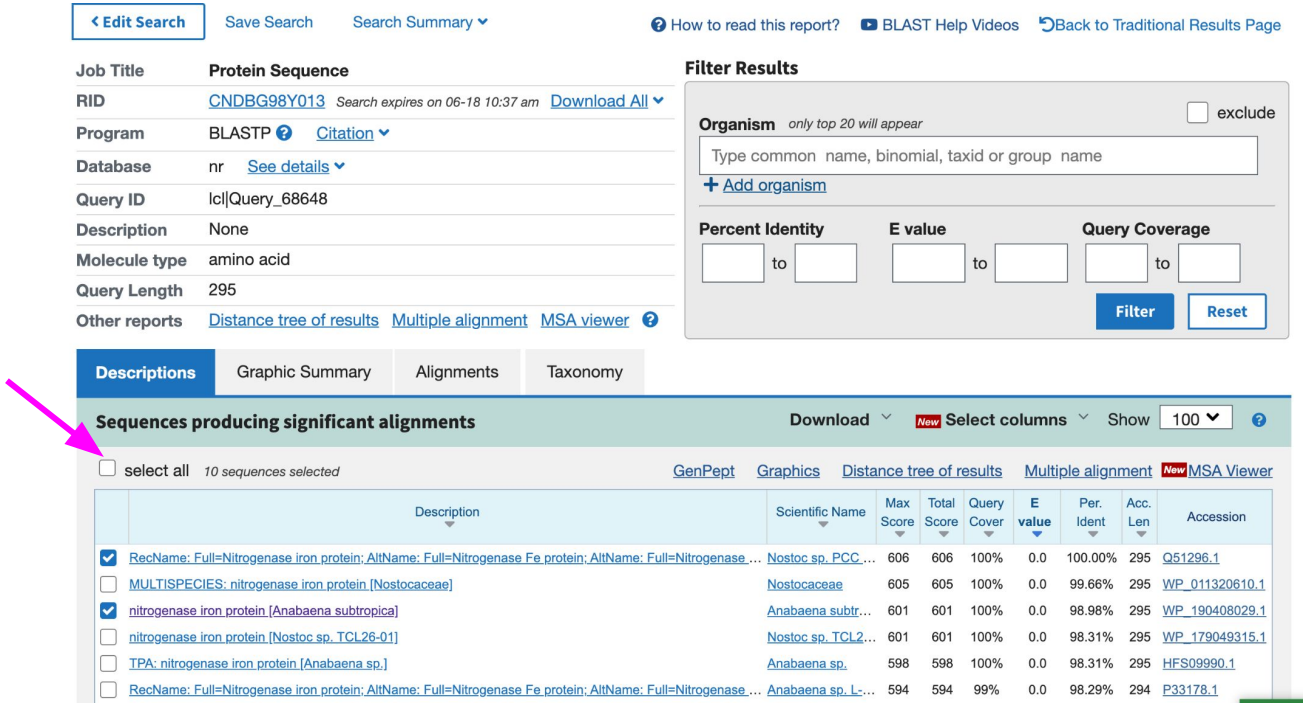
Multiple sequence alignment



Generate phylogenetic tree

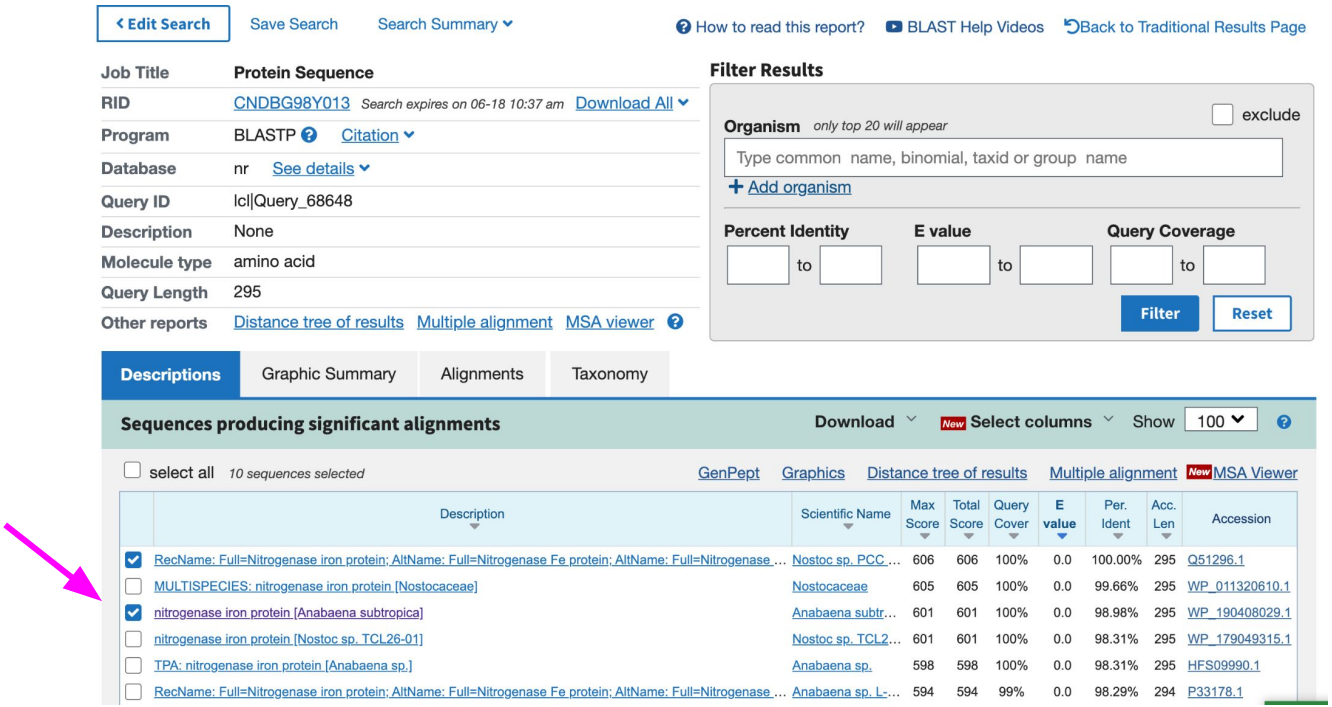
Gather sequences for multiple sequence alignment

1. Uncheck ‘select all’



Gather sequences for multiple sequence alignment

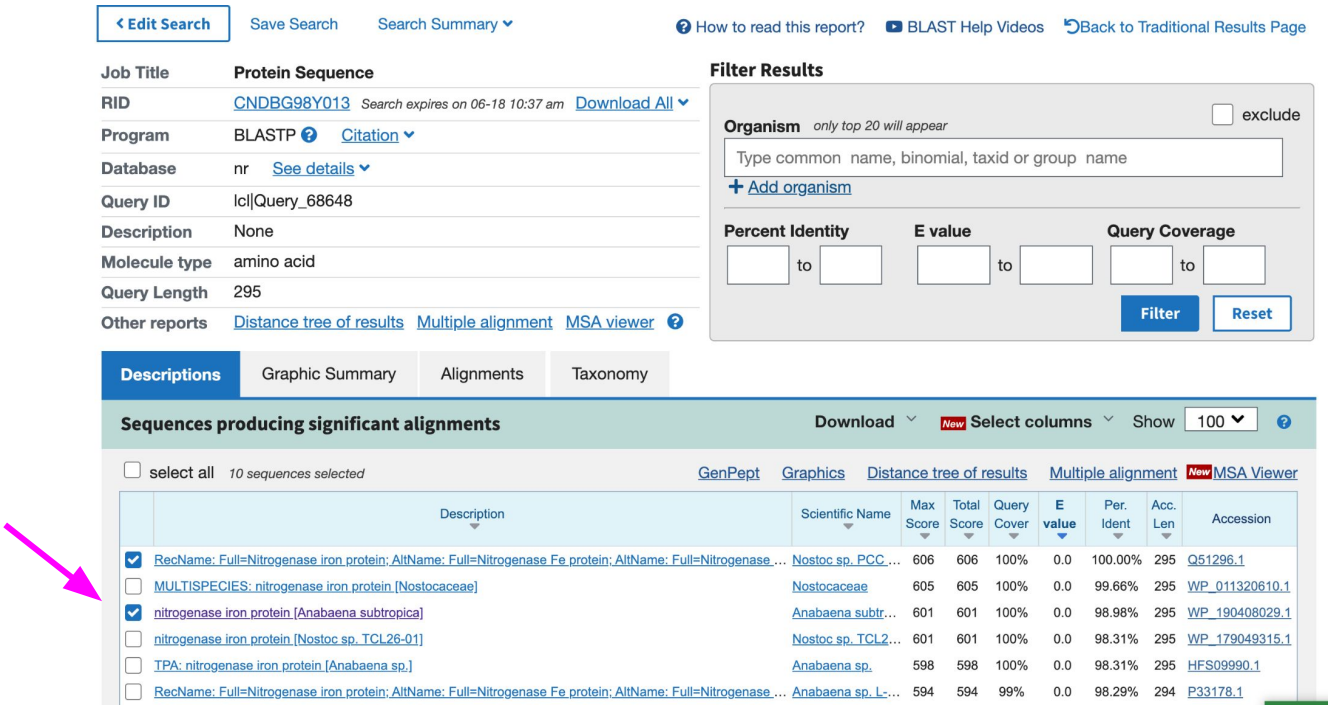
1. Uncheck ‘select all’
2. Check whichever sequences you wish to include in your phylogenetic tree (can scroll down)



Gather sequences for multiple sequence alignment

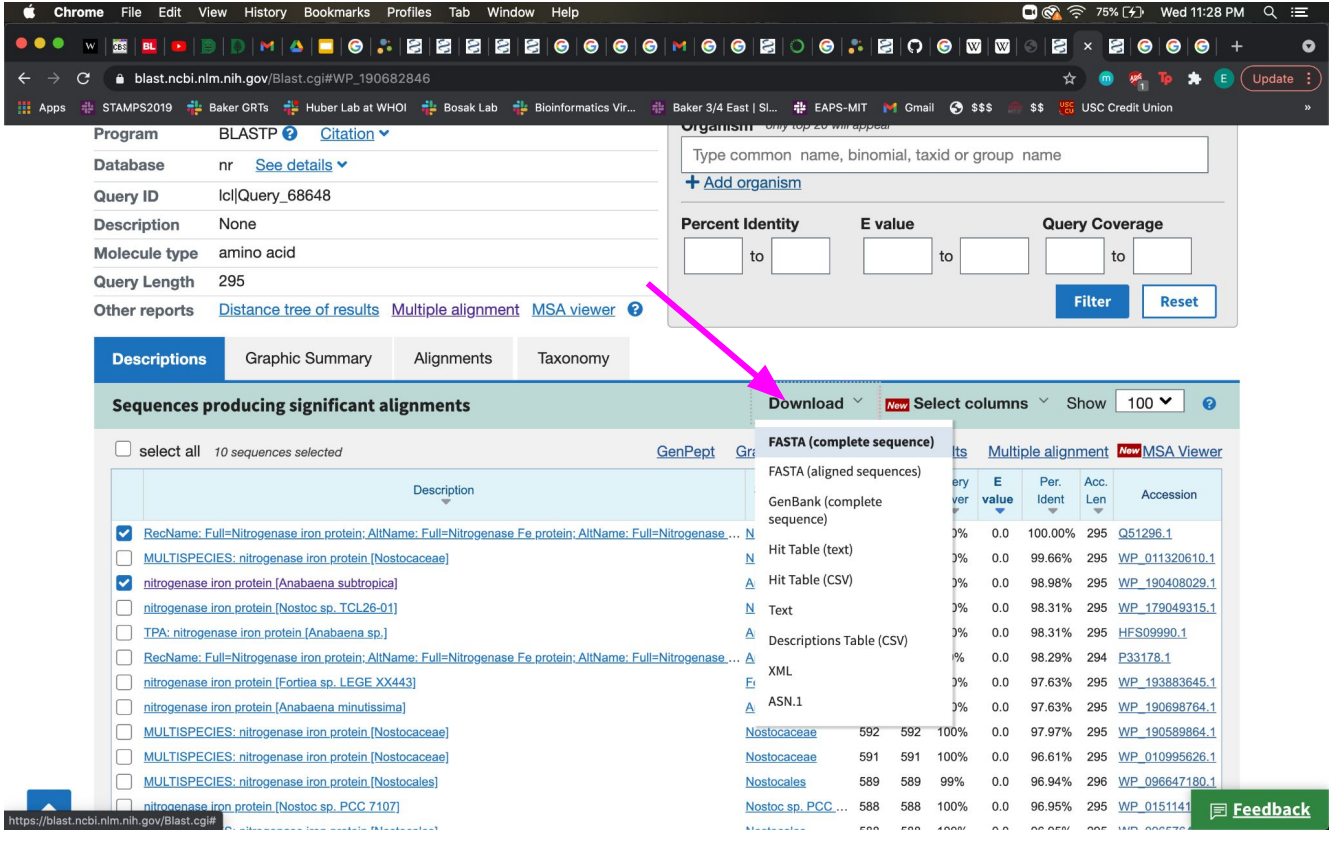
1. Uncheck ‘select all’
2. Check whichever sequences you wish to include in your phylogenetic tree (can scroll down)

And make sure to include those from your cyano-excluded or those with more dissimilar results (lover values)!



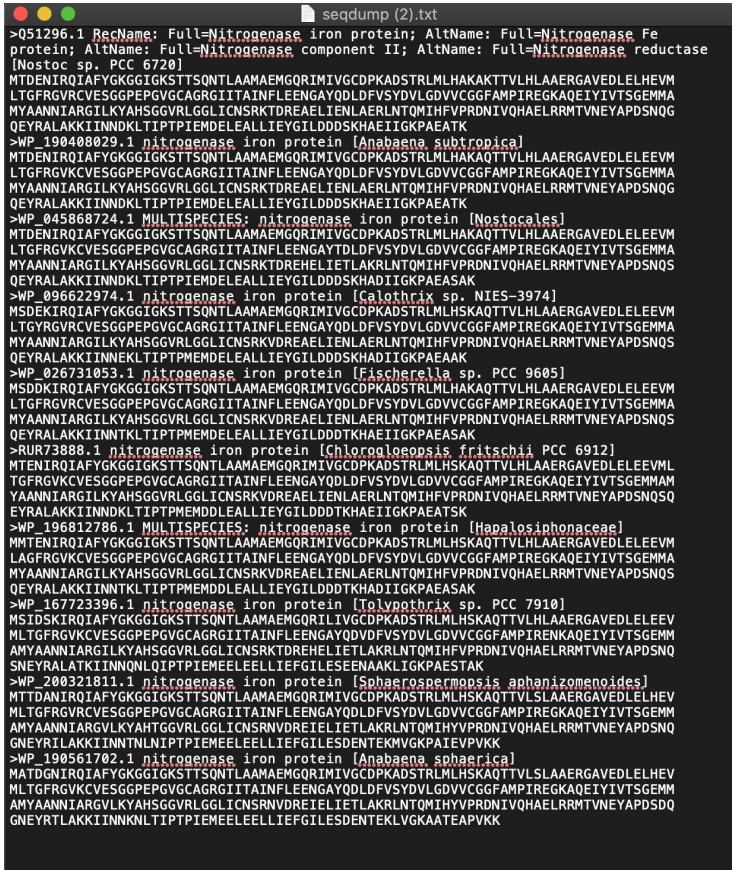
Gather sequences for multiple sequence alignment

1. Uncheck ‘select all’
2. Check whichever sequences you wish to include in your phylogenetic tree (can scroll down)
3. Click ‘Download’ then ‘FASTA (complete sequence)



Gather sequences for multiple sequence alignment

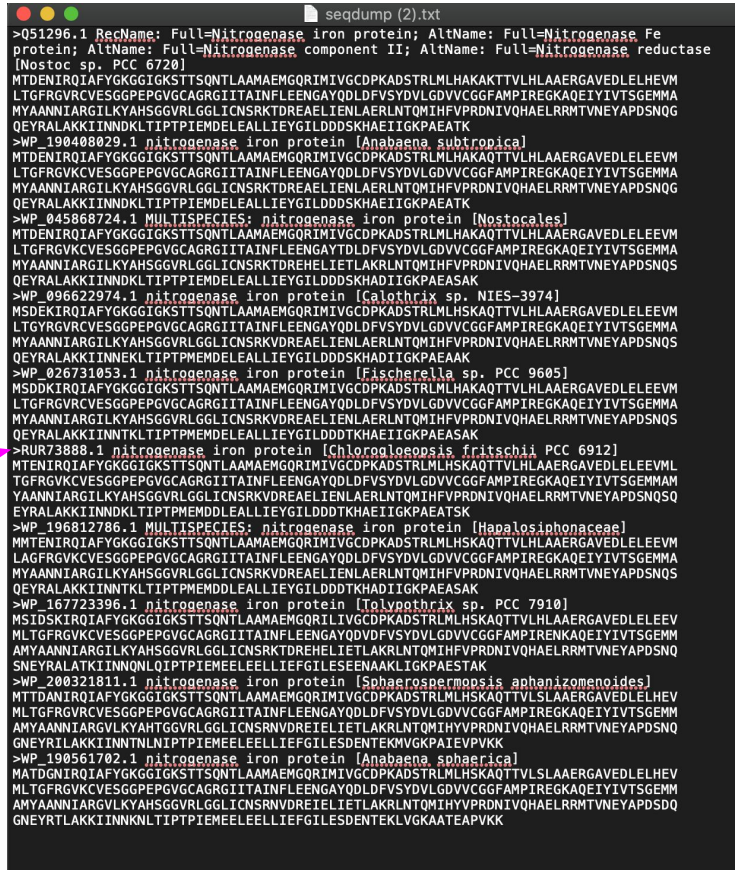
A file called seqdump.txt will appear (likely in your download) and may look like this:



Gather sequences for multiple sequence alignment

A file called seqdump.txt will appear (likely in your download) and may look like this:

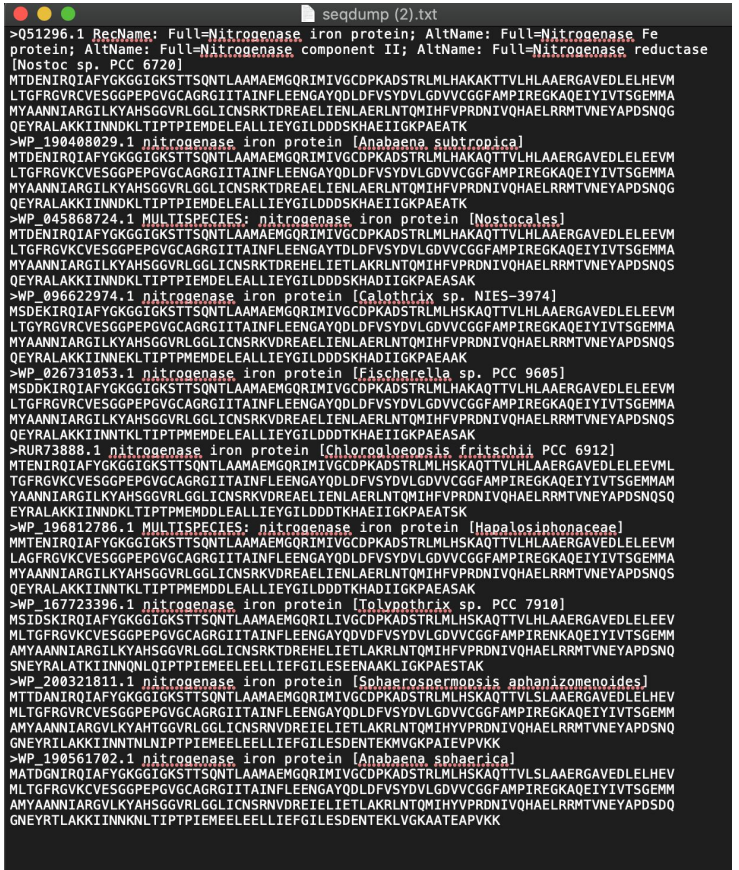
Notice our deflines :)
This is in fasta format!



Gather sequences for multiple sequence alignment

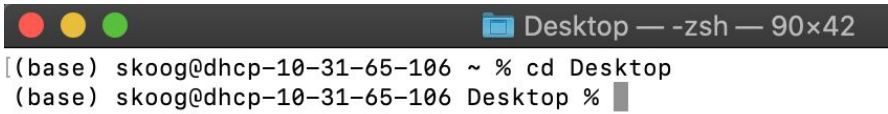
A file called seqdump.txt will appear (likely in your download) and may look like this:

Now move this file to your Desktop and rename it ‘NifH.fa’



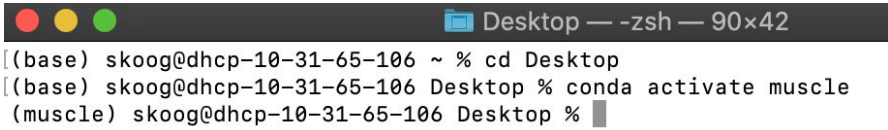
Terminal time!!

1. Open your terminal and ‘cd’ into your Desktop directory



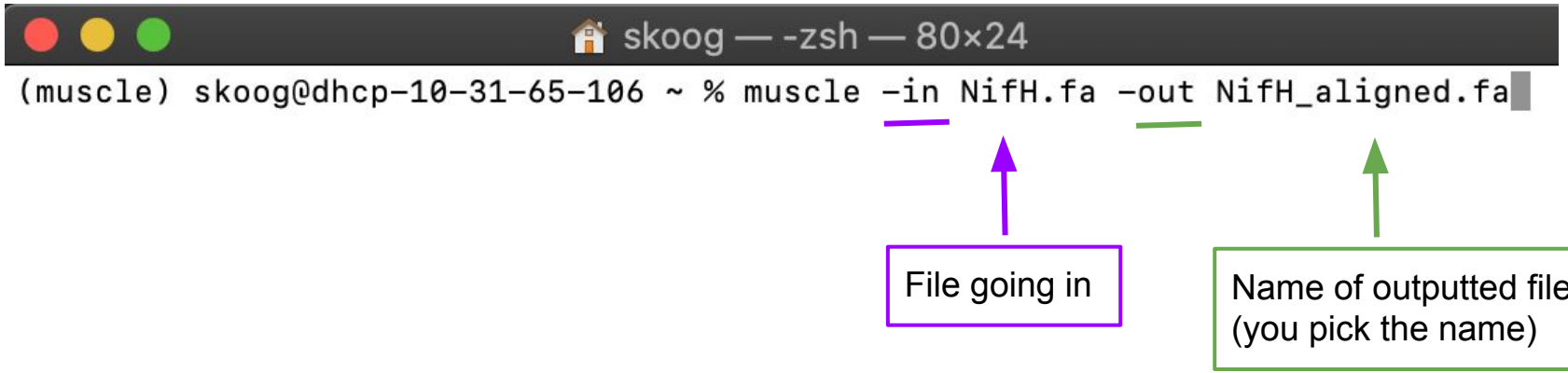
Terminal time!!

1. Open your terminal and ‘cd’ into your Desktop directory
2. Enter your ‘muscle’ environment



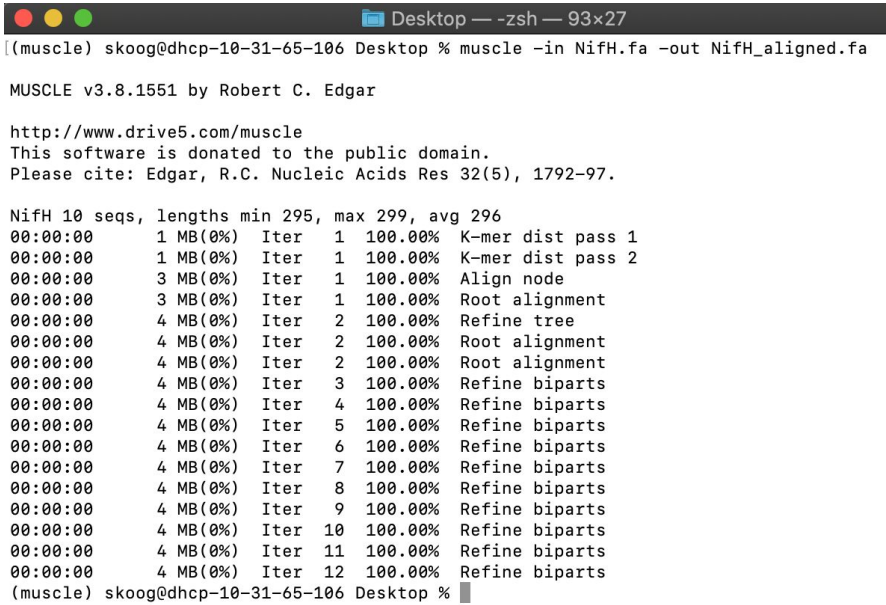
Terminal time!!

1. Open your terminal and ‘cd’ into your Desktop directory
2. Enter your ‘muscle’ environment
3. Now align those sequences using muscle!



Terminal time!!

1. Open your terminal and ‘cd’ into your Desktop directory
2. Enter your ‘muscle’ environment
3. Now align those sequences using muscle!

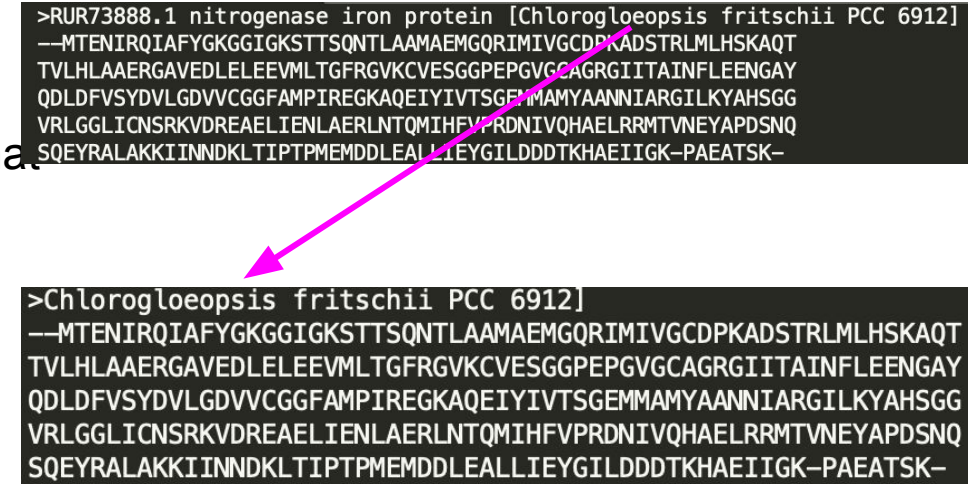


Terminal time!!

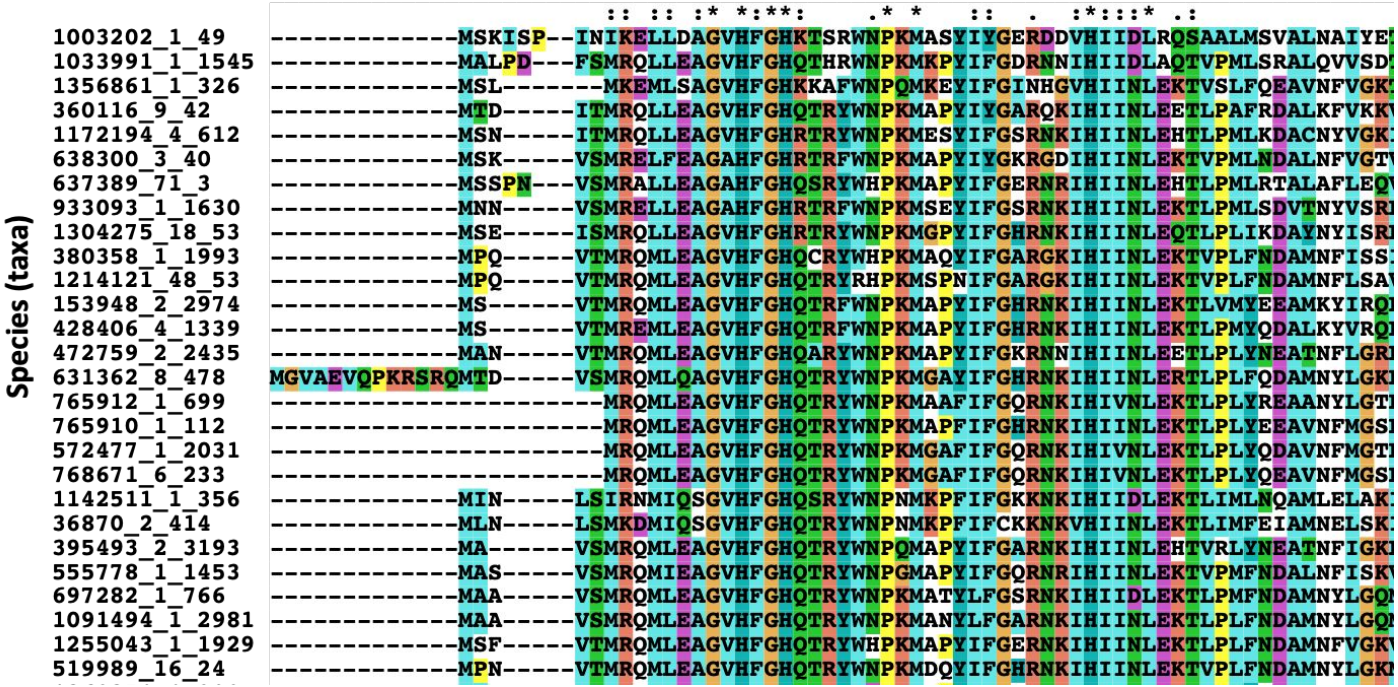
1. Open your terminal and ‘cd’ into your Desktop directory
2. Enter your ‘muscle’ environment
3. Now align those sequences using muscle!
4. Take a look at the alignment! (open the file)

Terminal time!!

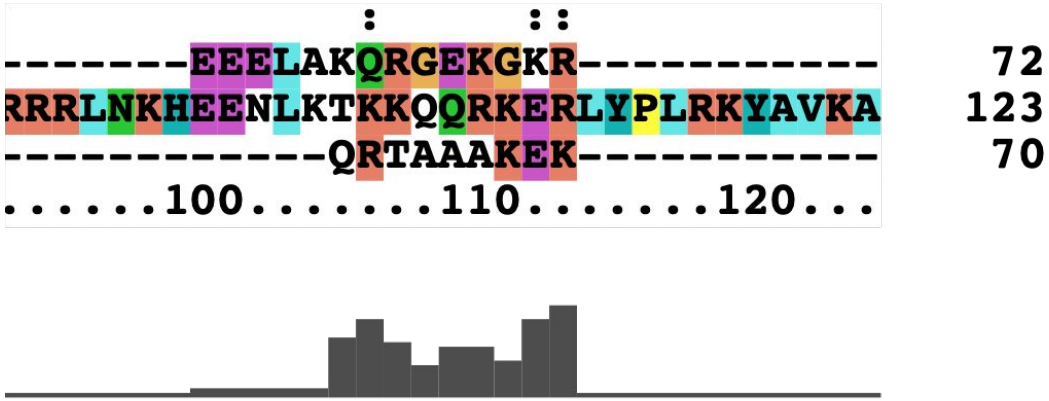
1. Open your terminal and ‘cd’ into your Desktop directory
2. Enter your ‘muscle’ environment
3. Now align those sequences using muscle!
4. Take a look at the alignment! (open the file)
5. Last thing (will come in handy later during tree viewing): remove the first section of the defline (leave the >) so that only the organism name is present



Multiple sequence alignment viewing programs exist



Multiple ways to align sequences (parameters matter)



Alternative alignment?

```
KQ-R-----GEKGKR
KQQR-----KER---
--QRTAAAKEK---
```

?

Slide courtesy of Greg Fournier

How do we determine which sites should be aligned?

•Alignment Scores: Count well-aligned sites in sequences, scoring matches and penalizing mismatches, and penalizing gaps;

•Scoring matches: better alignments have more sites across homologs with the same residues. Sites with similar amino acids are also more likely to be homologous. Regions with these sites are “well-aligned”.

•Gaps: indels cause gaps within alignments. “Gap penalties” impose a cost both for introducing a gap within an alignment, and also for extending this gap.

•In long sequences, some regions will be well-aligned by chance (false positives). For example, two random nucleotide sequences will be 25% identical by chance alone. For protein sequences, ability to detect distantly related homologs drops off sharply below 20% sequence identity.

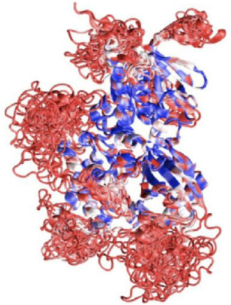
Slide courtesy of Greg Fournier

Review: Basic concepts in creating a phylogenetic tree

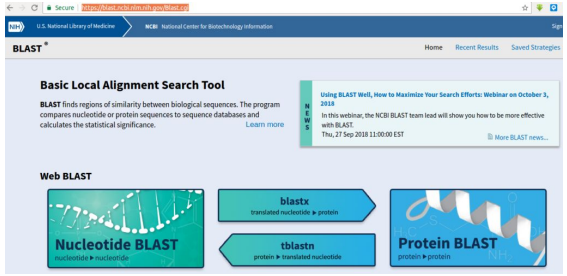
1. **Acquire DNA or protein of interest** and others (perhaps found through NCBI BLAST) that you are interested in comparing.
2. **Align sequences of interest** (can be nucleotide or amino acid sequences) using a program that does its best to align sequences so that they are as aligned or “matched” as possible (inserts gaps and mismatches where needed). We call this a **multiple sequence alignment**.
3. **Make a phylogenetic tree** using a program that does its best to take the aligned sequence file and create a tree with the best representation of the evolutionary relationship (not always the most-likely) based on the alignment

Overview of usual steps in creating a phylogenetic tree

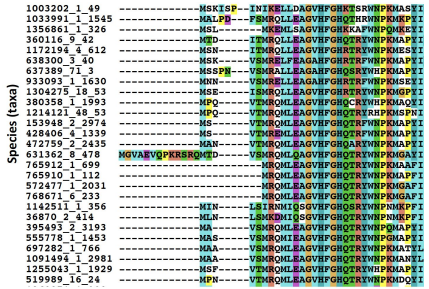
STEPS:



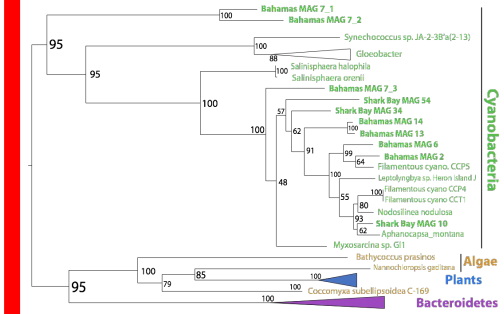
Identify amino acid (protein) or nucleotide (DNA) sequence of interest



Homology search using BLAST



Multiple sequence alignment



Generate phylogenetic tree


There are 3 TRILLION Trees in the World
(http://www.npr.org/sections/goatsandsoda/2015/09/02/436919052/tree-counter-is-astonished-by-how-many-trees-there-are)
How many taxa are needed to generate this many different tree topologies?
Slide courtesy of Greg Fournier


There are 3 TRILLION Trees in the World
(http://www.npr.org/sections/goatsandsoda/2015/09/02/436919052/tree-counter-is-astonished-by-how-many-trees-there-are)
How many taxa are needed to generate this many different tree topologies?
15
Slide courtesy of Greg Fournier

Many types of trees and programs to make trees

Different software for generating trees:

- IQ-TREE
- TREE PUZZLE
- PhyML
- RAxML
- Many, many more

These have different tree methods for building trees. They take in different types of sequences. And they run at different speeds.

Download IQ-TREE

IQ-TREE: phylogenetic tree builder based on maximum likelihood method

Briefly:

Maximum Likelihood: What is the likelihood of the DATA given the tree and the model?

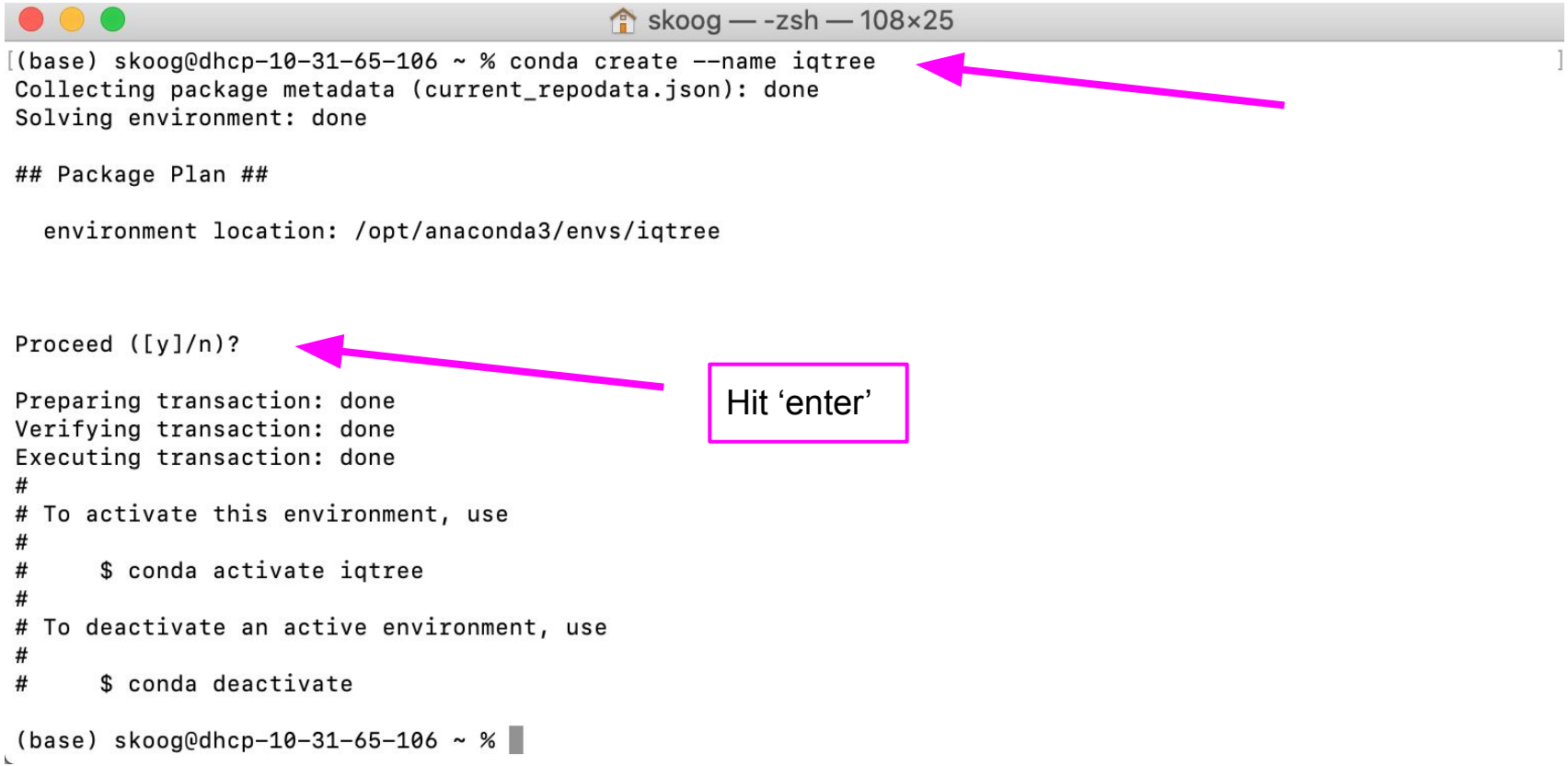
Bayesian Inference: What is the probability of the tree and the model given the data?

Verbiage from Greg Fournier

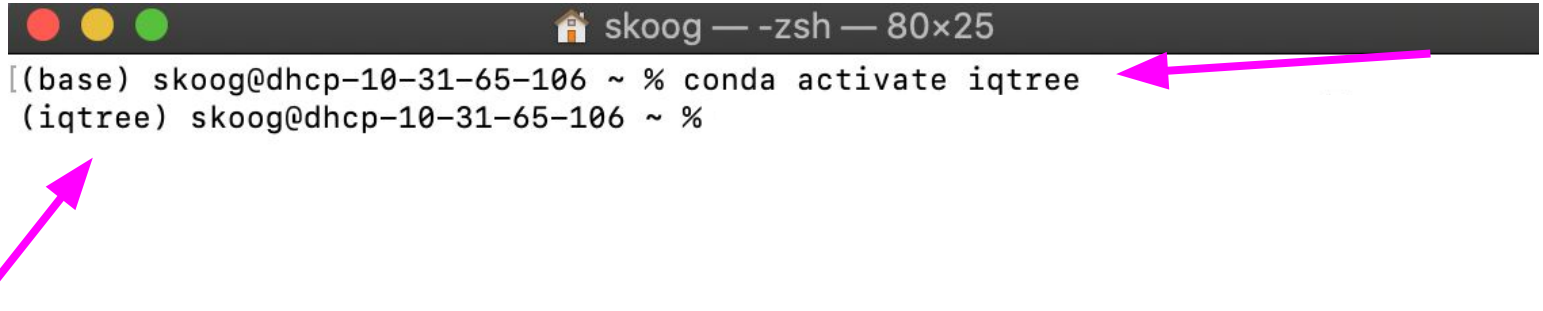
IQ-TREE: Exit MUSCLE environment

Notice change back to ‘base’

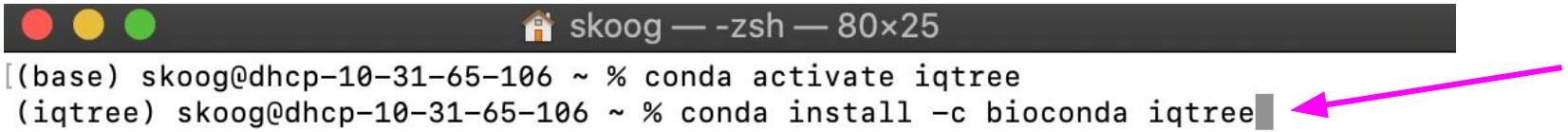
IQ-TREE: Create conda environment



MUSCLE: Activate environment

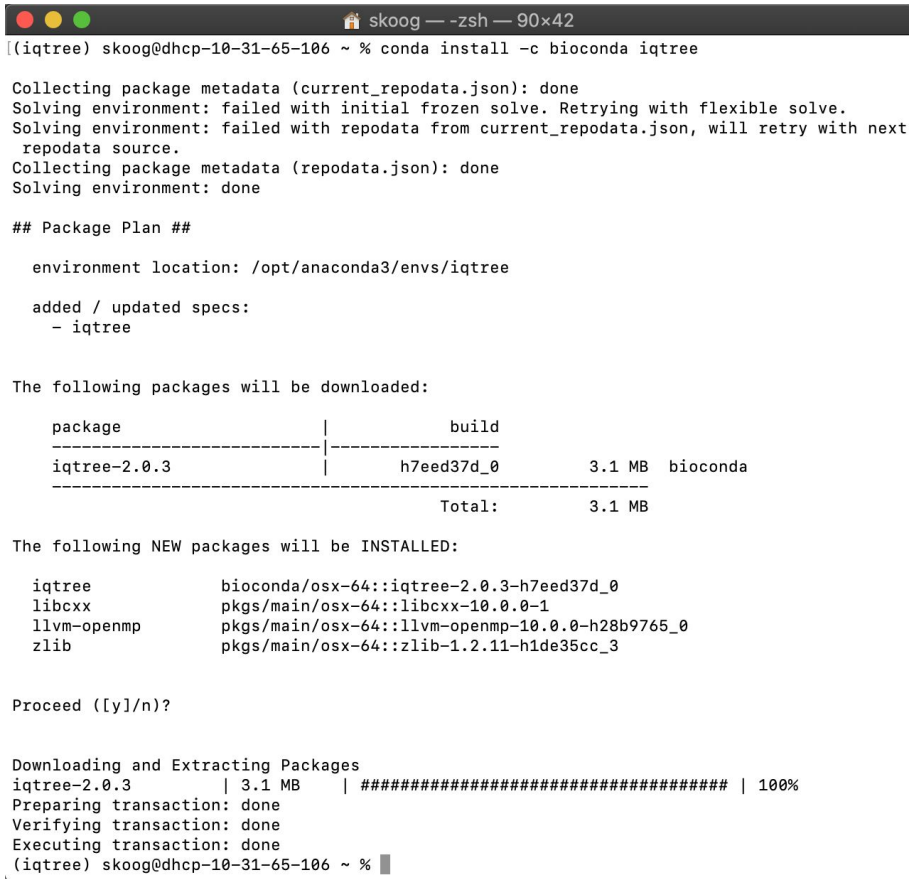


IQ-TREE: Install program inside of environment



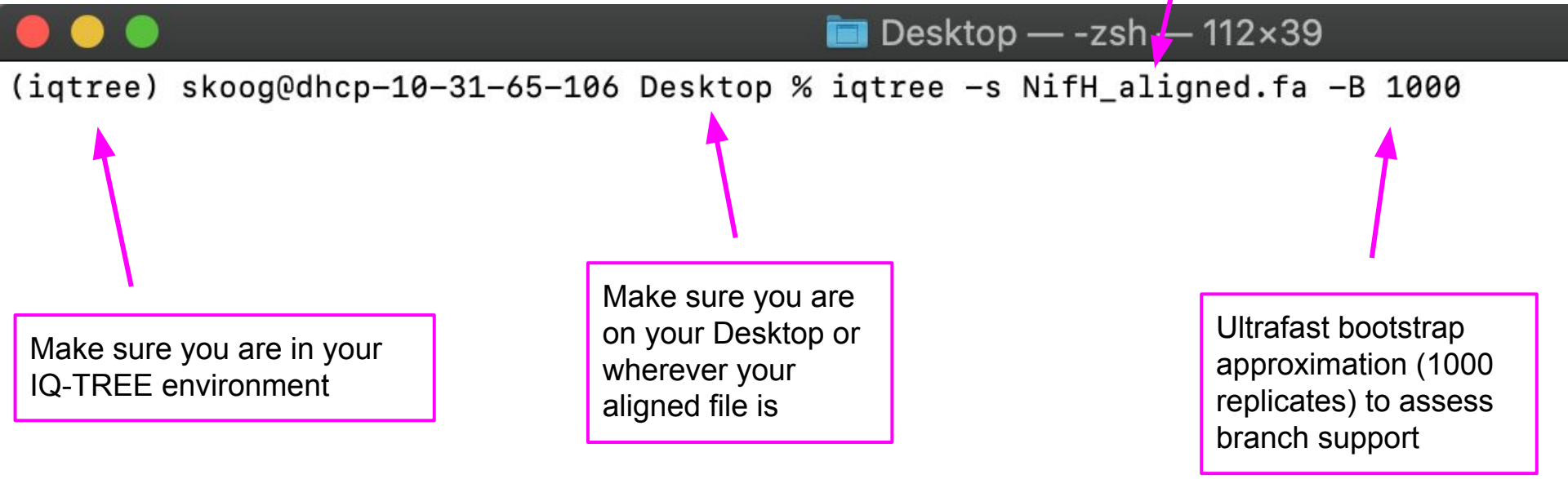
IQ-TREE: Install program inside of environment

Hit ‘enter’



```
skoog — -zsh — 90×42
(iqtree) skoog@dhcp-10-31-65-106 ~ % conda install -c bioconda iqtree

Collecting package metadata (current_repodata.json): done
Solving environment: failed with initial frozen solve. Retrying with flexible solve.
Solving environment: failed with repodata from current_repodata.json, will retry with next
 repodata source.
Collecting package metadata (repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: /opt/anaconda3/envs/iqtree

  added / updated specs:
    - iqtree


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    iqtree-2.0.3               |       h7eed37d_0         3.1 MB  bioconda
    ------------------------------------------------------------
                                           Total:         3.1 MB

The following NEW packages will be INSTALLED:

  iqtree             bioconda/osx-64::iqtree-2.0.3-h7eed37d_0
  libcxx             pkgs/main/osx-64::libcxx-10.0.0-1
  llvm-openmp        pkgs/main/osx-64::llvm-openmp-10.0.0-h28b9765_0
  zlib               pkgs/main/osx-64::zlib-1.2.11-h1de35cc_3


Proceed ([y]/n)?


Downloading and Extracting Packages
iqtree-2.0.3         | 3.1 MB    | ##################################### | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
(iqtree) skoog@dhcp-10-31-65-106 ~ %
```

Run IQ-TREE!


Aligned file
Desktop — -zsh — 112×39
(iqtree) skoog@dhcp-10-31-65-106 Desktop % iqtree -s NifH_aligned.fa -B 1000
Make sure you are in your IQ-TREE environment
Make sure you are on your Desktop or wherever your aligned file is
Ultrafast bootstrap approximation (1000 replicates) to assess branch support

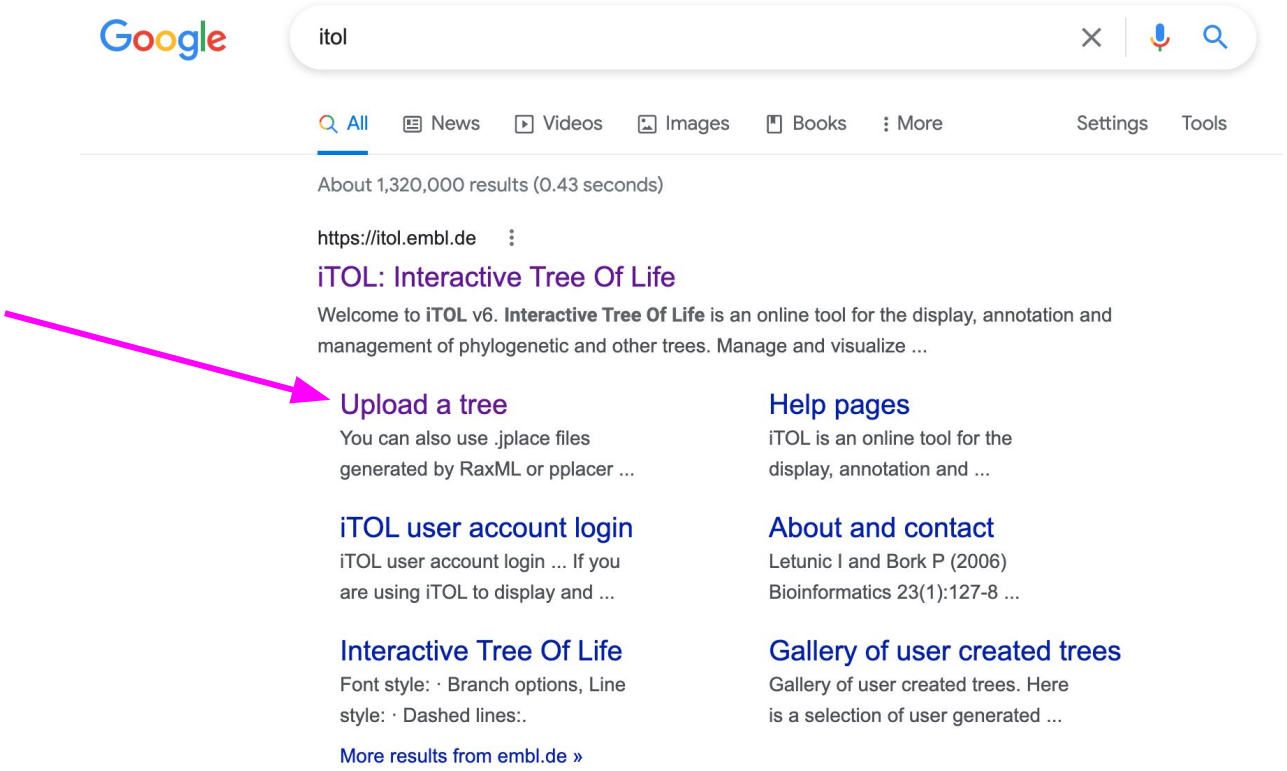
Run IQ-TREE

This should have created multiple files on your desktop.

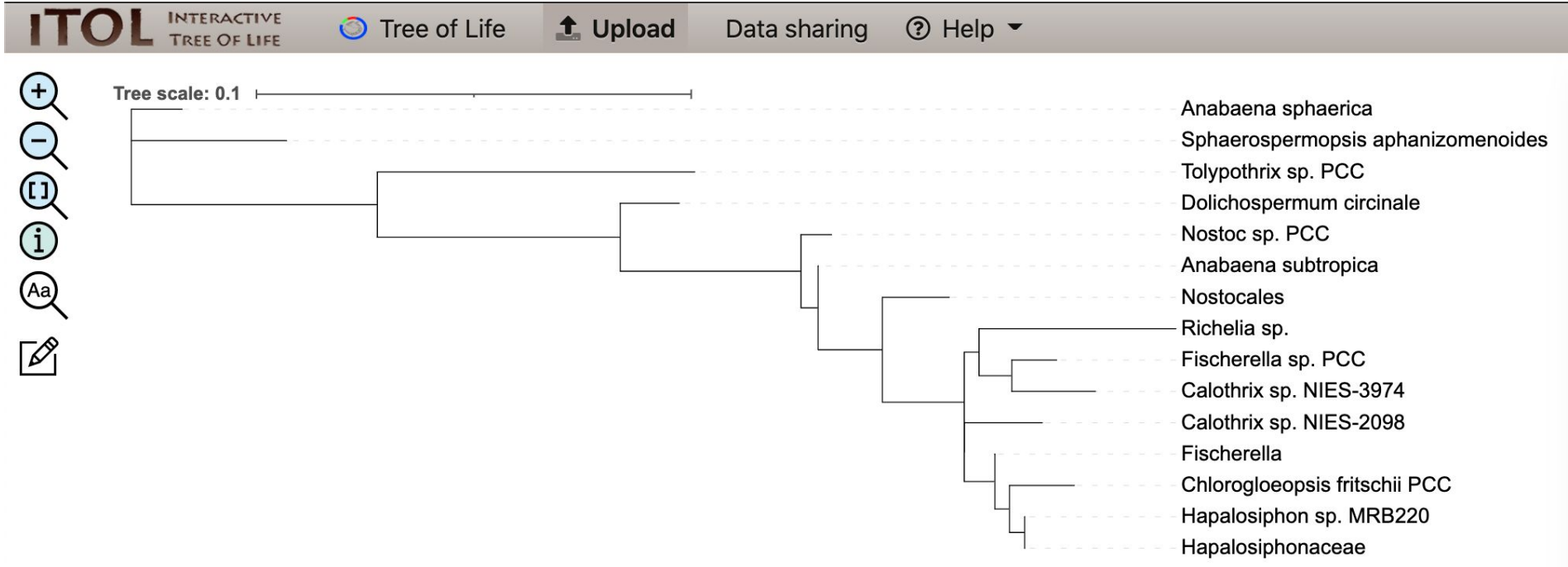
The one we are interested in is the one with the `.treefile` ending (in my case NifH_aligned.fa.treefile)

View tree in iTOL

- Open iTOL and upload your file ending in .treefile



View tree in iTOL



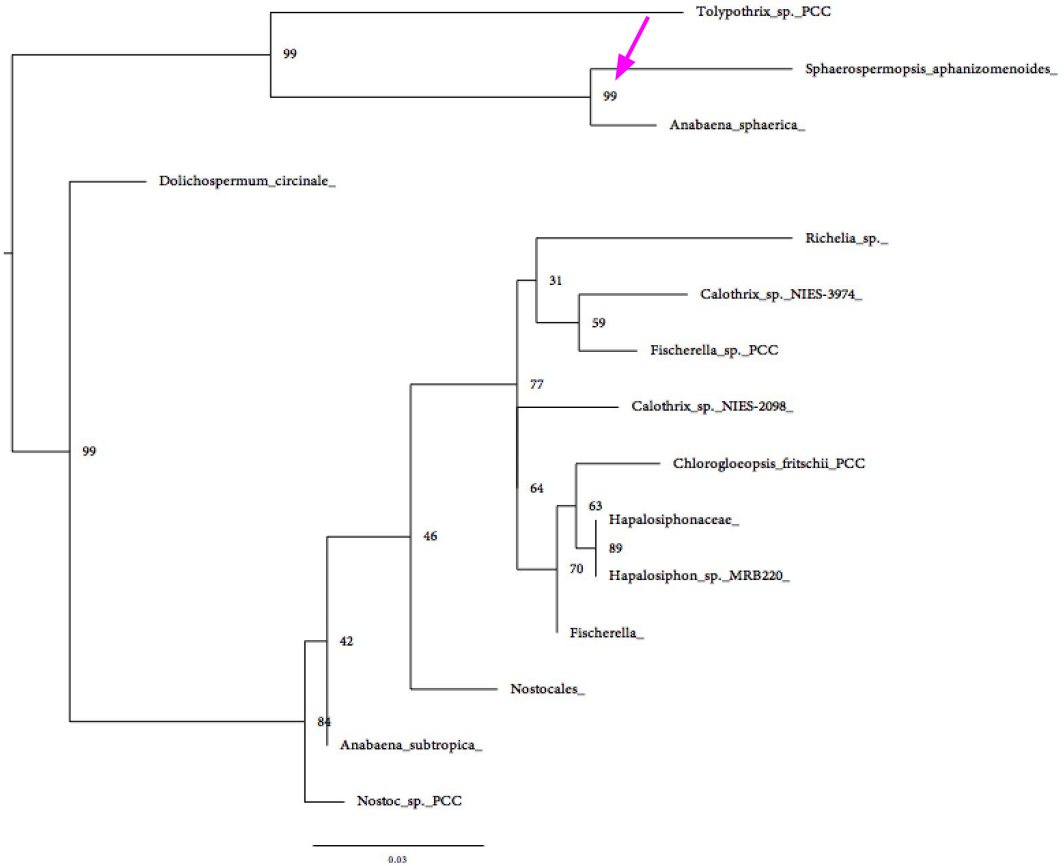
Other tree visualization softwares

There are tree visualization softwares that you can download onto your computer if you want to be serious about tree building! The one I am most familiar with is FigTree.

(FigTree can look very skeletal and ugly without colors, but then I beautify in Affinity Designer (similar to Adobe Illustrator (stay tuned for a crash course on figure making in Affinity Designer (waaaaaay cheaper than Illustrator (but just as awesome (in my opinion))))))

Same tree in FigTree

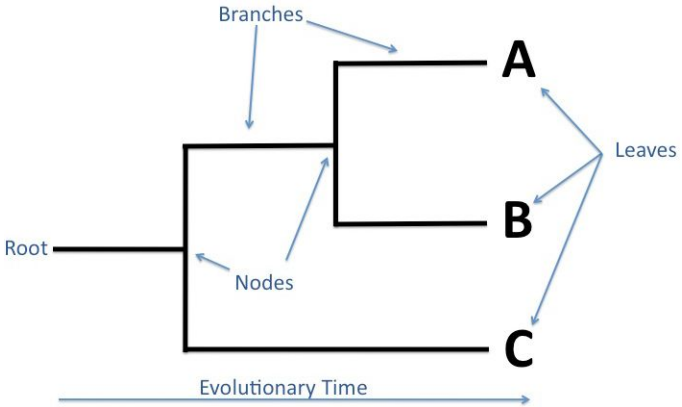
Bootstrap values: indicates how many times out of 100, the same branch was observed when repeating the phylogenetic reconstruction (you can be more confident in branches where node bootstrap values are closer to 100)

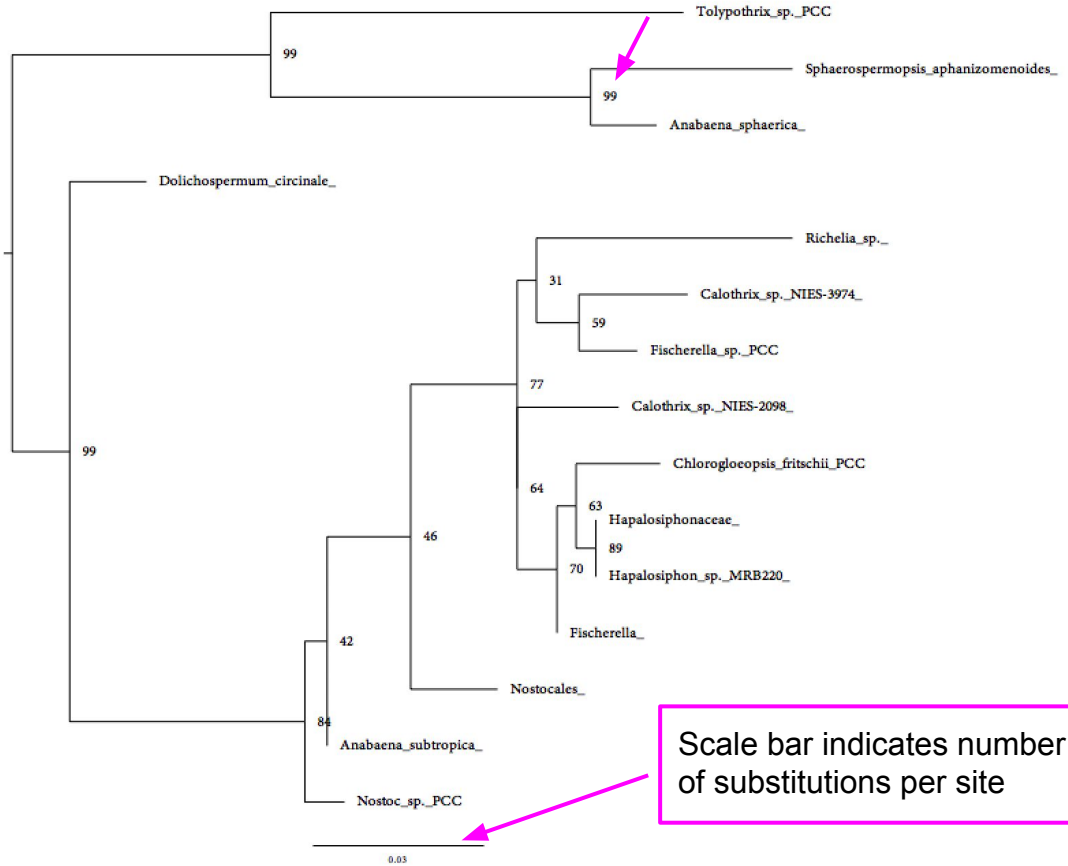


Same tree in FigTree

Bootstrap values: indicates how many times out of 100, the same branch was observed when repeating the phylogenetic reconstruction (you can be more confident in branches where node bootstrap values are closer to 100)

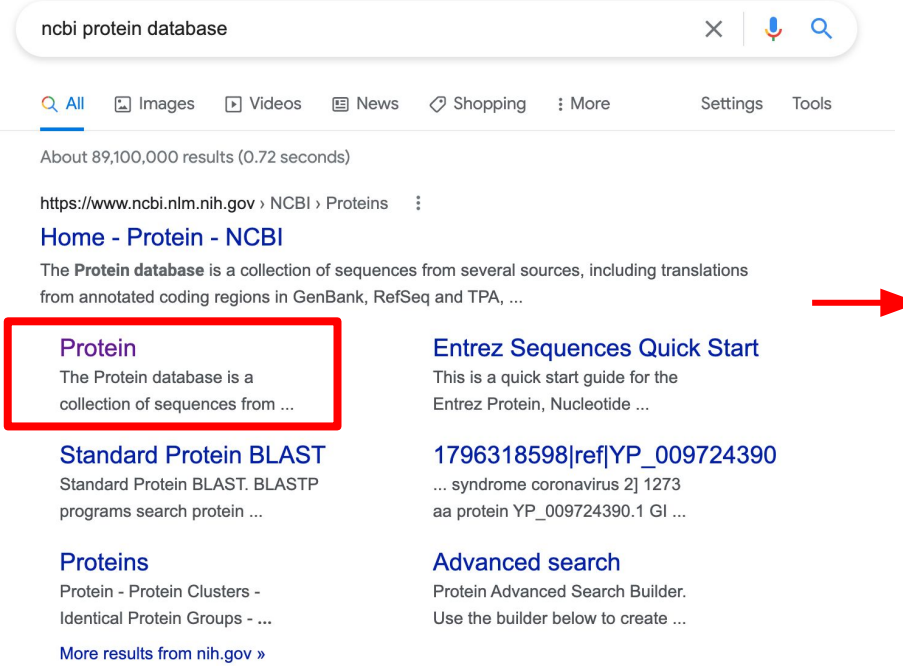
Parts of a phylogenetic tree

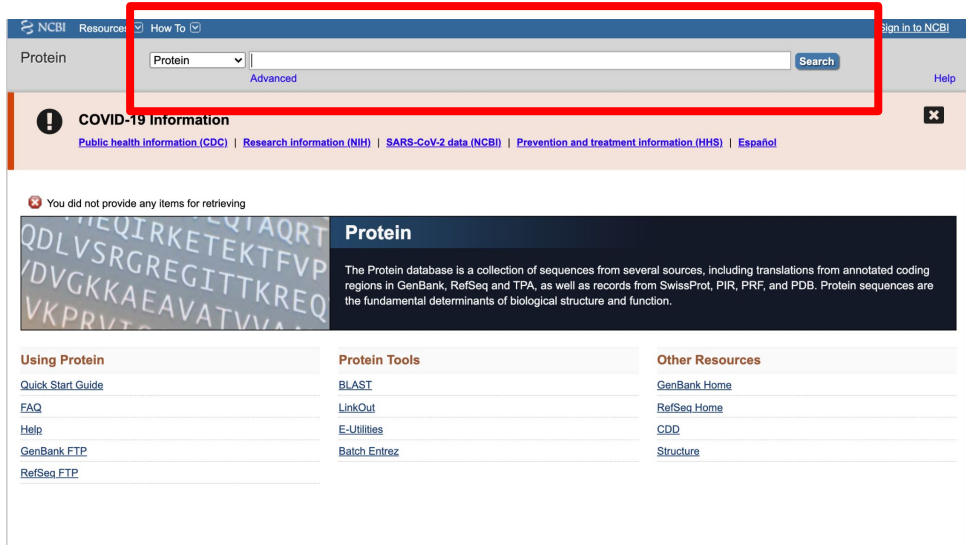




Scale bar indicates number of substitutions per site

Some ways to acquire protein sequence: NCBI





Some ways to acquire protein sequence: JGI IMG

- Useful if you are mining certain organisms for certain genes that you then want to ‘BLAST’
- If you have your own data uploaded here, you can easily get sequences for specific annotated genes of interest