

# Introduction to Phylogenetics

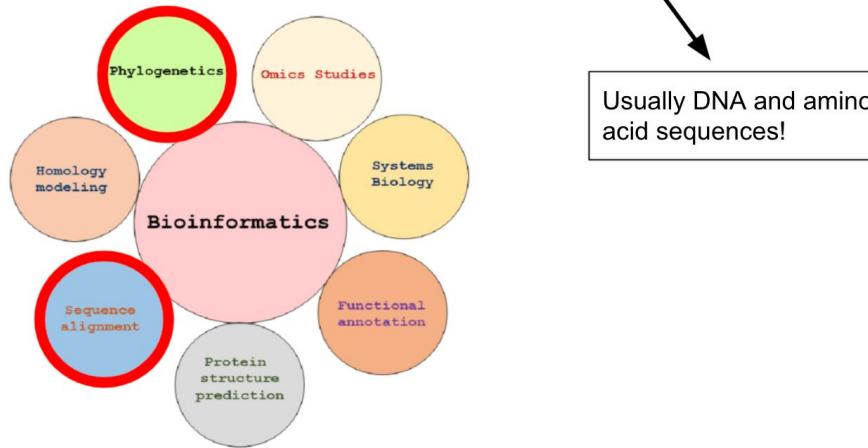
Emilie Skoog

Summer 2021

# Recall...

## What is bioinformatics?

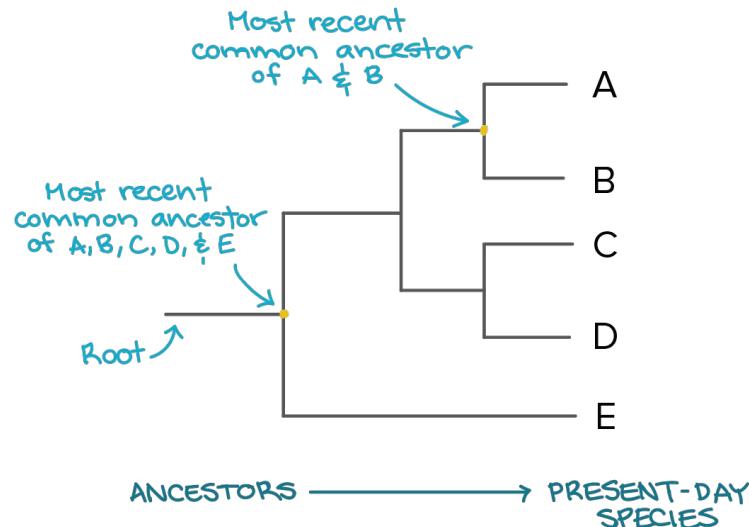
- Combines biology and computer science to **interpret biological data**



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.

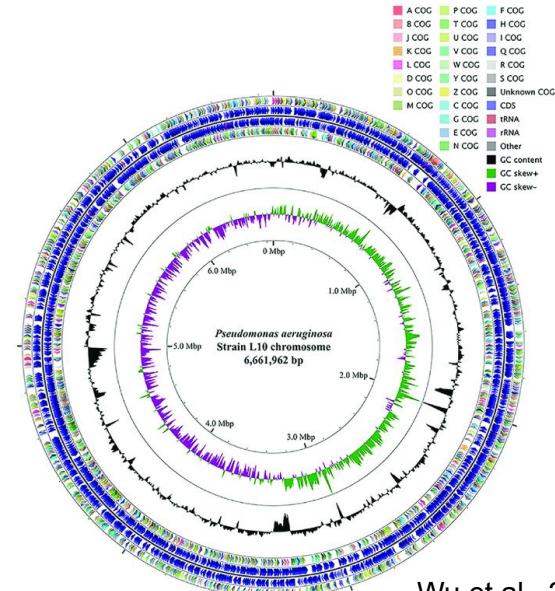


## Key terms

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

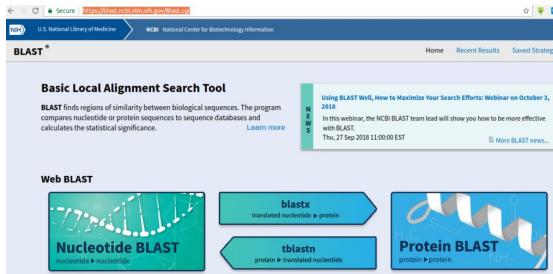
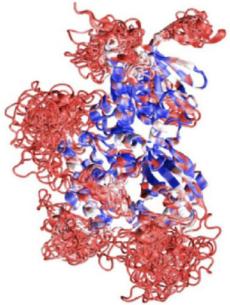
# Phylogenetics vs. Phylogenomics

- **Phylogenetics** compares and analyzes the sequences of single genes, or a small number of **genes**.
- **Phylogenomics** draws information by comparing entire genomes, or at least large portions of **genomes**.



# Overview of usual steps in creating a phylogenetic tree

STEPS:



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

Homology search using BLAST

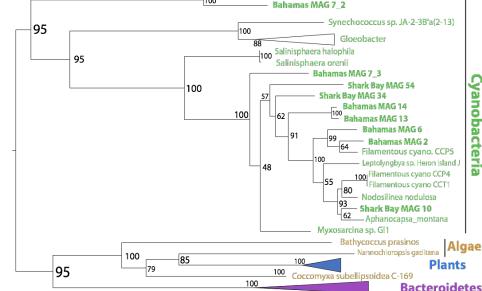


Multiple sequence alignment

Species (taxa)

10033201\_1\_49 HSKICP-----TINKEELDAAGVIFGQTRSPNIPWPKAYI  
1033991\_1\_1945 HALP-----PSMQLLEAGAVIFGQTRHNNPKMKPYI  
1356861\_1\_326 MSL-----MKEMLSAGVIFGKKAFTVNPNMKCYI  
360116\_9\_42 KBD-----ITDROLLAEAGVIFGQSRVNNPKHAPYI  
117311\_1\_612 LBD-----ITDROLLAEAGVIFGQSRVNNPKHAPYI  
638300\_3\_30 MSK-----VSMGELFLAGAAMPFGESRSPVNPHEAPYI  
637389\_71\_3 MSPN-----VSMGELFLAGAAMPFGESRSPVNPHEAPYI  
933092\_1\_1630 HNN-----VSMGELLLAGAAIFGHRERFPRVNPHEAPYI  
10033201\_1\_1946 HPO-----VSMGELLLAGAAIFGHRERFPRVNPHEAPYI  
380359\_1\_1993 HPO-----VSMGELLLAGAVIFGQCRVNPHEAPYI  
1214121\_48\_53 HPO-----VSMGELLLAGAVIFGQTRVNPHEAPYI  
153948\_2\_2974 MS-----VTMQMLLAAGAVIFGQSRVNNPKHAPYI  
404412\_1\_12 MS-----VTMQMLLAAGAVIFGQSRVNNPKHAPYI  
472759\_2\_2435 HAN-----VSMGELLAAGAVIFGQSRVNPHEAPYI  
631362\_8\_478 NOVAEVOPKRKROWMD-----VSMGELLAAGAVIFGQTRVNPHEAPYI  
765912\_1\_699 MRLML-----MRLMLAGAVIFGQSRVNNPKHAPYI  
770001\_1\_12 MRLML-----MRLMLAGAVIFGQSRVNNPKHAPYI  
572477\_1\_2031 MRLML-----MRLMLAGAVIFGQSRVNPHEAPYI  
768671\_6\_233 MRLML-----MRLMLAGAVIFGQTRVNPHEAPYI  
1142511\_1\_356 HIN-----L1RNIMISGQVIFGQSRVNNPKHAPYI  
346871\_1\_12 HIN-----L1RNIMISGQVIFGQSRVNNPKHAPYI  
395493\_2\_3193 MA-----VSMQMLSAGVIFGKKAFTVNPNMKCYI  
555779\_1\_1453 HAS-----VSMQMLSAGVIFGQTRVNPHEAPYI  
697282\_1\_766 HAA-----VSMQMLLAAGAVIFGQSRVNNPKHAPYI  
1001494\_1\_1981 HAA-----VSMQMLLAAGAVIFGQSRVNPHEAPYI  
1250000\_1\_1929 MSP-----VSMGELFLAGAAMPFGESRSPVNPHEAPYI  
519989\_16\_24 HPA-----VSMGELFLAGAAMPFGESRSPVNPHEAPYI

Multiple sequence alignment



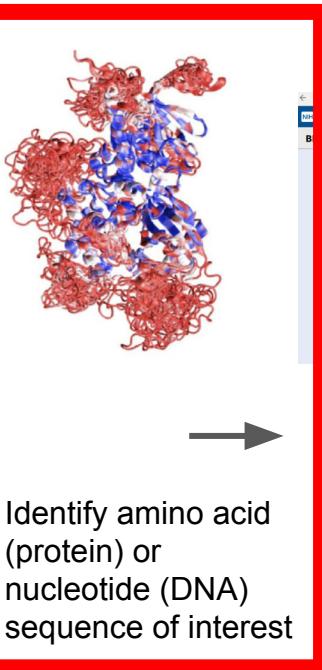
Generate phylogenetic tree

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



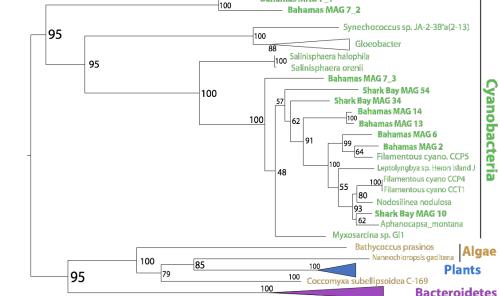
# Overview of usual steps in creating a phylogenetic tree

## STEPS:



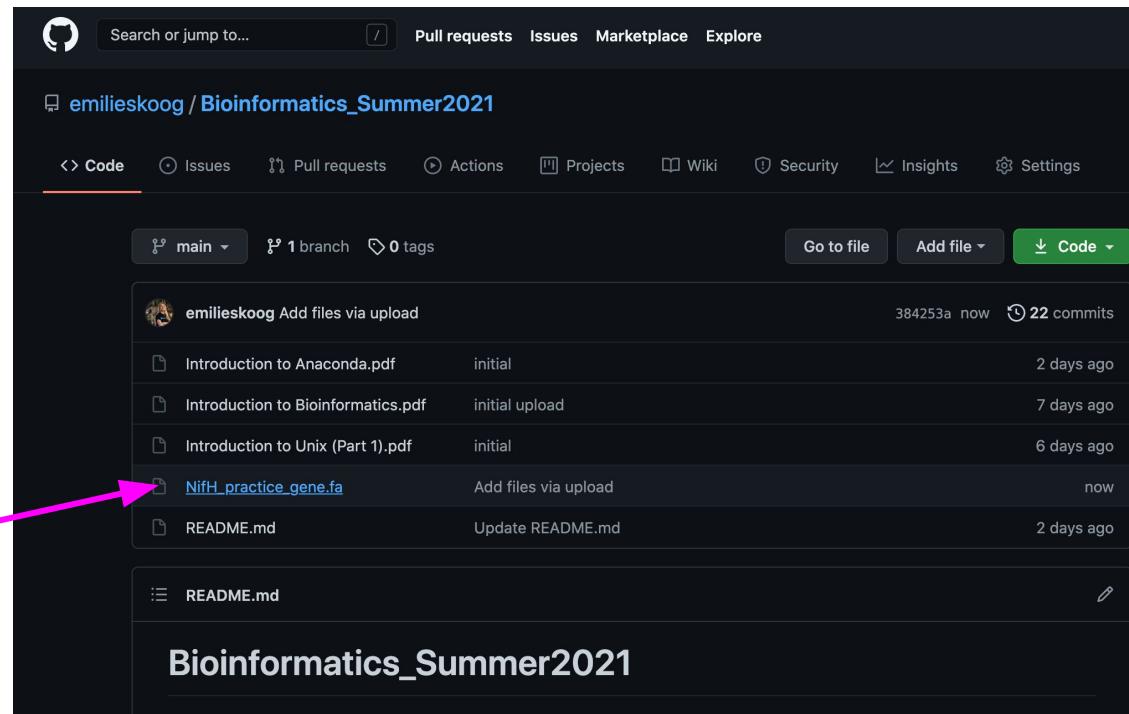
Homology search using BLAST

Multiple sequence alignment



# First: Acquire a gene of interest (provided)

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



The screenshot shows a GitHub repository page for 'emilieskoog / Bioinformatics\_Summer2021'. The 'Code' tab is selected. The repository has 1 branch and 0 tags. The commit history shows several files uploaded by 'emilieskoog': 'Introduction to Anaconda.pdf', 'Introduction to Bioinformatics.pdf', 'Introduction to Unix (Part 1).pdf', 'NifH\_practice\_gene.fa', and 'README.md'. A pink arrow points to the 'NifH\_practice\_gene.fa' file.

File	Description	Last Commit
Introduction to Anaconda.pdf	initial	2 days ago
Introduction to Bioinformatics.pdf	initial upload	7 days ago
Introduction to Unix (Part 1).pdf	initial	6 days ago
NifH_practice_gene.fa	Add files via upload	now
README.md	Update README.md	2 days ago

## 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]
MTDENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLHAKAKTTVLHLAAERGA
VEDLELHEVMLTGFGRGVRCVESGGPEPGVGCAGRIITAINFLEENGAYQDLDFVSYDVLGDVVCGGFAM
PIREGKAQEIQYIVTSGEMMAMYAANNIARGILKYAHSGGVRLGGLICNSRKTDREAElienlaerlntqm
IHFPVDNIVQHAELRRMTVNEYAPDSNQGQEYRALAKKIINNDKLTIPPTPIEMDELEALLIEYGILD
SKHAEIIGKPAEATK
```

# You will see a “FASTA file”:

```
>CAA83510.1 NifH [Nostoc sp. PCC 6720]
MTDENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLHAKAKTTVLHLAAERGA
VEDLELHEVMLTGFGRGVRCVESGGPEPGVGCAGRIITAINFLEENGAYQDLDFVSYDVLGDVVCGGFAM
PIREGKAQEIQYIVTSGEMMAMYAANNIARGILKYAHSGGVRLGGLICNSRKTDREAElienlaerlntqm
IHFPVDNIVQHAELRRMTVNEYAPDSNQGQEYRALAKKIINNDKLTIPPTPIEMDELEALLIEYGILD
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.

Unique identifier

Gene name

Gene's "host organism"

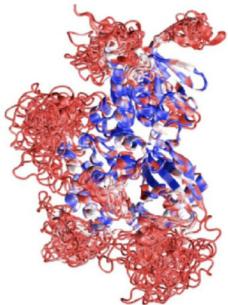
>CAA83510.1 NifH [Nostoc sp. PCC 6720]

MTDENIRQIAFYKGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLHAKAKTTVLHLAAERGA  
VEDLELHEVMLTGFRGVRCVESGGPEPGVGCAGRGIIITAINFLEENGAYQDLDFVSYDVLDVVCGGFAM  
PIREGKAQEIYIVTSGEMMAMYAANNIARGILKYAHSGGVRLGGLICNSRKTDRAEELIENLAERLNTQM  
IHFPVRDNIVQHAELRRMTVNEYAPDSNQGQEYRALAKKIINNDKLTIPPTIEMDELEALLIEYGILD  
SKHAEIIGKPAEATK

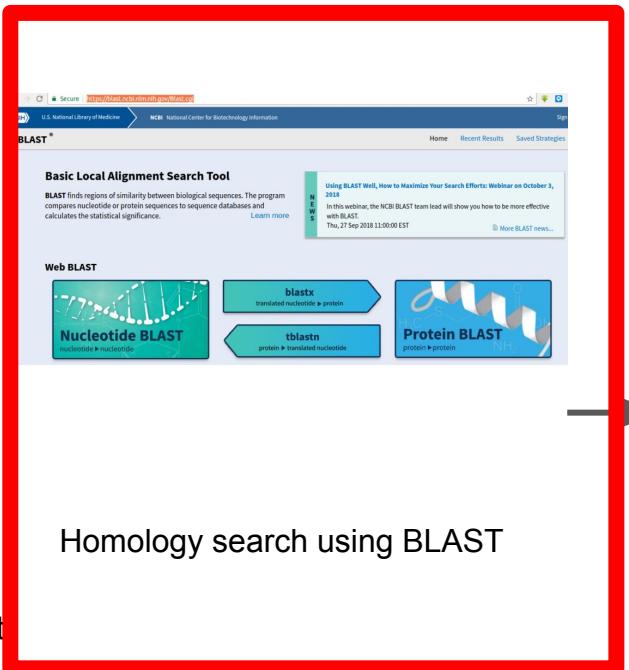
defline

# Overview of usual steps in creating a phylogenetic tree

## STEPS:



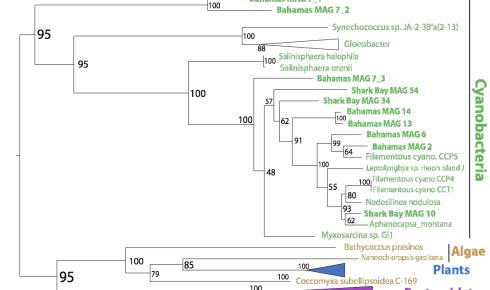
Identify amino acid  
(protein) or  
nucleotide (DNA)  
sequence of interest



Multiple sequence alignment

Accession	Length	Sequence
1033991_1_1945	1945	HSKKQF--TINKEELDAEVIFGQTRSPNIPWPAHAYI
1356861_1_326	326	HALPDP--PSMQLLEAEAVIFGQTRHNPWPKHAPYI
360116_9_42	42	MSL--MKEMLSAGVIFGKKAFTVNPNPKMCYI
117136_1_612	612	KHD--ITDROLLAEAVIFGQTRVNPWPKHAPYI
638300_3_40	40	MSK--VSMGELFLAEAVIFGQTRSPVNPWPKHAPYI
637389_7_3	3	MSPFI--VSMGALLAEAVIFGQTRVNPWPKHAPYI
933093_1_1630	1630	HNN--VSMGRELLEAEAVIFGHRERFPVNPWPKHAPYI
160383_1_1933	1933	TINKEELDAEVIFGQTRSPNIPWPAHAYI
380359_1_1993	1993	MPO--VTMQMLAEAVIFGQCRVNPWPKHAPYI
1214121_48_53	53	MPO--VTMQRMLAEAVIFGQTRVNPWPKHAPYI
153948_2_2974	2974	MS--VTMQRMLAEAVIFGQTRVNPWPKHAPYI
404462_1_2435	2435	VAN--VTMQRMLAEAVIFGQTRVNPWPKHAPYI
472759_2_2435	2435	VAN--VTMQRMLAEAVIFGQTRVNPWPKHAPYI
631362_8_478	478	NOVAEVOPKRKROWMD--VSMQRMLAEAVIFGQTRVNPWPKHAPYI
765912_1_699	699	MROMLAEAVIFGQTRVNPWPKHAPYI
765912_1_699	699	MROMLAEAVIFGQTRVNPWPKHAPYI
572471_1_2031	2031	MROMLAEAVIFGQTRVNPWPKHAPYI
768673_6_233	233	MROMLAEAVIFGQTRVNPWPKHAPYI
1142511_1_356	356	MHN--LS1RNIMGSQVIFGQTRVNPWPKHAPYI
346781_1_356	356	MHN--LS1RNIMGSQVIFGQTRVNPWPKHAPYI
395493_2_3193	3193	MA--VSMQRMLAEAVIFGQTRVNPWPKHAPYI
555778_1_1453	1453	MAS--VSMQRMLAEAVIFGQTRVNPWPKHAPYI
697282_1_766	766	MAA--VSMQRMLAEAVIFGQTRVNPWPKHATYL
1091494_1_1981	1981	MAA--VSMQRMLAEAVIFGQTRVNPWPKHATYL
1250000_1_1929	1929	MSPF--VSMQRMLAEAVIFGQTRVNPWPKHAPYI
519981_16_24	24	HPN--VSMQRMLAEAVIFGQTRVNPWPKHAPYI

Multiple sequence alignment



Generate phylogenetic tree

# Go to NCBI BLAST

Google ncbi blast

All Books News Videos Images More Settings Tools

About 90,100,000 results (0.68 seconds)

<https://blast.ncbi.nlm.nih.gov> :

**BLAST: Basic Local Alignment Search Tool**

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to ...

You've visited this page many times. Last visit: 6/16/21

**FASTA format**  
BLAST is a heuristic that works by finding word-matches between the query and ...

**Standard Protein BLAST**  
Standard Protein BLAST. BLASTP programs search protein ...

**US National Library of Medicine**  
Then use the BLAST button at the bottom of the page to align your ...

**Nucleotide BLAST**  
Homo sapiens (human) Nucleotide BLAST. BLASTN programs ...

[More results from nih.gov »](#)

# 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.

The screenshot shows the official BLAST homepage. At the top, there's a blue header with the NIH logo, the text "U.S. National Library of Medicine", "National Center for Biotechnology Information", and a "Log in" button. Below the header, the word "BLAST®" is prominently displayed. To the right, there are links for "Home", "Recent Results", "Saved Strategies", and "Help". A green vertical bar on the left contains the word "NEWS". On the right, there's a news box with the following text:

A new feature was added to Primer-BLAST.  
We now offer the ability for user to run primer-blast from NCBI assembly page..

Tue, 23 Feb 2021 12:00:00 EST

[More BLAST news...](#)

Below the news box, there's a section titled "Web BLAST" featuring three icons: "Nucleotide BLAST" (nucleotide to nucleotide), "blastx" (translated nucleotide to protein), and "tblastn" (protein to translated nucleotide). To the right of these is a "Protein BLAST" icon (protein to protein). At the bottom, there's a search bar for "BLAST Genomes" with fields for "Enter organism common name, scientific name, or tax id", "Human", "Mouse", "Rat", "Microbes", and a "Search" button.

# 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

The screenshot shows the NCBI BLAST homepage. At the top, there's a navigation bar with the NIH logo, "U.S. National Library of Medicine National Center for Biotechnology Information", a "Log in" button, and links for "Home", "Recent Results", "Saved Strategies", and "Help". A green vertical bar on the left says "NEWS". A message box says: "A new feature was added to Primer-BLAST. We now offer the ability for user to run primer-blast from NCBI assembly page..". Below this is a timestamp: "Tue, 23 Feb 2021 12:00:00 EST" and a link "More BLAST news...". The main content area is titled "Basic Local Alignment Search Tool". It describes BLAST as finding regions of similarity between biological sequences by comparing nucleotide or protein sequences to sequence databases and calculating statistical significance. There's a "Learn more" link. Below this, under "Web BLAST", there are three options: "Nucleotide BLAST" (nucleotide → nucleotide), "blastx" (translated nucleotide → protein), and "tblastn" (protein → translated nucleotide). To the right is "Protein BLAST" (protein → protein). Below these options is a "BLAST Genomes" section with a search bar and buttons for "Human", "Mouse", "Rat", and "Microbes". Three pink arrows point from the text labels on the left to their corresponding BLAST options: one arrow points to "Nucleotide BLAST", another to "blastx", and a third to "Protein BLAST".

# Search for homologous sequence using NCBI BLAST

1. Click on 'protein blast' (blastp)

NIH U.S. National Library of Medicine  
National Center for Biotechnology Information

Log in

BLAST®

Home Recent Results Saved Strategies Help

**Basic Local Alignment Search Tool**

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**NEWS**

A new feature was added to Primer-BLAST.  
We now offer the ability for user to run primer-blast from NCBI assembly page..

Tue, 23 Feb 2021 12:00:00 EST [More BLAST news...](#)

**Web BLAST**

**Nucleotide BLAST** nucleotide ▶ nucleotide

**blastx** translated nucleotide ▶ protein

**tblastn** protein ▶ translated nucleotide

**Protein BLAST** protein ▶ protein

**BLAST Genomes**

Enter organism common name, scientific name, or tax id  Search

Human Mouse Rat Microbes

>CAA83510.1 NifH [Nostoc sp. PCC 6720]  
MTDENIRQIAFYGKGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLHAKAKTTVLHLAAERGA  
VEDLELHEVMLTGFRGVRCVESGGPEPGVGCGAGRIITAINFLEENGAYQDLDVSYDVLGVCGGFAM  
PIREGKAQEYIYIVTSGEMMAMYAANNIARGILKYAHSGGVRLGLICNSRKTDRAEELIENLAERLNTQM  
IHFPVRDNIVQHAELRRMTVNEYAPDSNQGQEYRALAKKIINNDKLTIP TPMDEALLIEYGILDDD  
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'

Standard Protein BLAST

blastn    **blastp**    blastx    tblastn    tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)    Query subrange [?](#)

MTDENIRQIAFYKGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLH  
AKAKTTVLHLAAERGA  
**VEDDELHEVMLTGRGVRCESGGPEPGVGCAGRGIITAINFLEENGAYQDLDF**  
**SYDVVLGDGVCGGFAM**

From  To

Or, upload file  No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database  [?](#)

Organism [Optional](#) Enter organism name or id—completions will be suggested  exclude [Add organism](#)

Exclude [Optional](#) Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Models (XM/XP)  Non-redundant RefSeq proteins (WP)  Uncultured/environmental sample sequences

**Program Selection**

Algorithm  Quick BLASTP (Accelerated protein-protein BLAST)  
 **blastp** (protein-protein BLAST)  
 PSI-BLAST (Position-Specific Iterated BLAST)  
 PHI-BLAST (Pattern Hit Initiated BLAST)  
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

**BLAST**    Search database nr using Blastp (protein-protein BLAST)  
 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

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

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. [more...](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

MTDENIRQIAFYKGIGKSTTSQNTLAAMAEMGQRIMVGCDPKADSTRMLH  
AKAKTTVLHLAAERGA  
VEDLELHEVMLTGRGVRCVESGGPEPGVGCAGRGIIATINFEENGAYQLDF  
VSYDVLDGVVCGGFAM

From \_\_\_\_\_ To \_\_\_\_\_

Or, upload file [Choose File](#) No file chosen [?](#)

Job Title \_\_\_\_\_

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database Non-redundant protein sequences (nr) [?](#)

Organism [Optional](#) Enter organism name or id—completions will be suggested  exclude [Add organism](#)

Exclude [Optional](#) Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Models (XM/XP)  Non-redundant RefSeq proteins (WP)  Uncultured/environmental sample sequences

**Program Selection**

Algorithm  Quick BLASTP (Accelerated protein-protein BLAST)  
 blastp (protein-protein BLAST)  
 PSI-BLAST (Position-Specific Iterated BLAST)  
 PHI-BLAST (Pattern Hit Initiated BLAST)  
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

**BLAST** [Search database nr using Blastp \(protein-protein BLAST\)](#)  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)

Standard Protein BLAST

blastn    **blastp**    blastx    tblastn    tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)    Query subrange [?](#)

MTDENIRQIAFYKGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLH  
AKAKTTVLHLAAERGA  
VEDLELHEVMLTGRGVRCVESGGPEPGVGCAGRGIITAINFLEENGAYQDLDF  
VSYDVLDGVVCGGFAM

From  To

Or, upload file  No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database  [?](#)

Organism Optional  Enter organism name or id—completions will be suggested  exclude [Add organism](#) [?](#)

Exclude Optional  Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Models (XM/XP)  Non-redundant RefSeq proteins (WP)  Uncultured/environmental sample sequences [?](#)

**Program Selection**

Algorithm  Quick BLASTP (Accelerated protein-protein BLAST)  
 **blastp** (protein-protein BLAST)  
 PSI-BLAST (Position-Specific Iterated BLAST)  
 PHI-BLAST (Pattern Hit Initiated BLAST)  
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)  
Choose a BLAST algorithm [?](#)

**BLAST**

Search database nr using Blastp (protein-protein BLAST)  
 Show results in a new window

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

Standard Protein BLAST

blastn    **blastp**    blastx    tblastn    tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)    Query subrange [?](#)

MTDENIRQIAFYKGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLH  
AKAKTTVLHLAAERGA  
VEDLELHEVMLTGRGVRCVESGGPEPGVGCAGRGIITAINFLEENGAYQDLDVF  
VSYDVLDGVVCGGFAM

From \_\_\_\_\_ To \_\_\_\_\_

Or, upload file  No file chosen [?](#)

Job Title \_\_\_\_\_

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database  [?](#)

Organism [Optional](#) Enter organism name or id—completions will be suggested  exclude [Add organism](#)

Exclude [Optional](#) Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Models (XM/XP)  Non-redundant RefSeq proteins (WP)  Uncultured/environmental sample sequences

**Program Selection**

Algorithm  Quick BLASTP (Accelerated protein-protein BLAST)  
 **blastp** (protein-protein BLAST)  
 PSI-BLAST (Position-Specific Iterated BLAST)  
 PHI-BLAST (Pattern Hit Initiated BLAST)  
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

**BLAST**

Search database nr using Blastp (protein-protein BLAST)  
 Show results in a new window

# 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	<a href="#">CNDBG98Y013</a> Search expires on 06-18 10:37 am <a href="#">Download All</a>
Program	BLASTP <a href="#">?</a> <a href="#">Citation</a>
Database	nr <a href="#">See details</a>
Query ID	Icl Query_68648
Description	None
Molecule type	amino acid
Query Length	295
Other reports	<a href="#">Distance tree of results</a> <a href="#">Multiple alignment</a> <a href="#">MSA viewer</a> <a href="#">?</a>

## Filter Results

**Organism** only top 20 will appear  exclude

Type common name, binomial, taxid or group name  
[+ Add organism](#)

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

to   to   to

**Filter** **Reset**

[Descriptions](#) [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

Sequences producing significant alignments										<a href="#">Download</a>	<a href="#">Select columns</a>	Show 100	<a href="#">?</a>	
<input checked="" type="checkbox"/> select all 100 sequences selected										<a href="#">GenPept</a>	<a href="#">Graphics</a>	<a href="#">Distance tree of results</a>	<a href="#">Multiple alignment</a>	<a href="#">MSA Viewer</a>
	Description			Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession			
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ...	Nostoc sp. PCC ...	606	606	100%	0.0	100.00%	295	<a href="#">Q51296.1</a>					
<input checked="" type="checkbox"/>	MULTISPECIES: nitrogenase iron protein [Nostocaceae]	Nostocaceae	605	605	100%	0.0	99.66%	295	<a href="#">WP_011320610.1</a>					
<input checked="" type="checkbox"/>	nitrogenase iron protein [Anabaena subtrlica]	Anabaena subtr...	601	601	100%	0.0	98.98%	295	<a href="#">WP_190408029.1</a>					
<input checked="" type="checkbox"/>	nitrogenase iron protein [Nostoc sp. TCL26-01]	Nostoc sp. TCL2...	601	601	100%	0.0	98.31%	295	<a href="#">WP_179049315.1</a>					
<input checked="" type="checkbox"/>	TPA: nitrogenase iron protein [Anabaena sp.]	Anabaena sp.	598	598	100%	0.0	98.31%	295	<a href="#">HFS09990.1</a>					
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ...	Anabaena sp. L...	594	594	99%	0.0	98.29%	294	<a href="#">P33178.1</a>					

# Our results:

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 Icl|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 E value Query Coverage

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
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ... Nostoc sp. PCC...	Nostocaceae	606	606	100%	0.0	100.00%	295	Q51296.1
<input checked="" type="checkbox"/>	MULTISPECIES: nitrogenase iron protein [Nostocaceae]	Nostocaceae	605	605	100%	0.0	99.66%	295	WP_011320610.1
<input checked="" type="checkbox"/>	nitrogenase iron protein [Anabaena subtrlica]	Anabaena subtr...	601	601	100%	0.0	98.98%	295	WP_190408029.1
<input checked="" type="checkbox"/>	nitrogenase iron protein [Nostoc sp. TCL26-01]	Nostoc sp. TCL2...	601	601	100%	0.0	98.31%	295	WP_179049315.1
<input checked="" type="checkbox"/>	TPA: nitrogenase iron protein [Anabaena sp.]	Anabaena sp.	598	598	100%	0.0	98.31%	295	HFS09990.1
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ... Anabaena sp. L...	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   Icl|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

Identity	E value	Query Coverage
<input type="text"/>	<input type="text"/> to <input type="text"/>	<input type="text"/> to <input type="text"/>

**Filter**   **Reset**

Our query (or protein) sequence was 295 amino acids

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
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ... <a href="#">Nostoc sp. PCC...</a>	<a href="#">Nostocaceae</a>	606	606	100%	0.0	100.00%	295	<a href="#">Q51296.1</a>
<input checked="" type="checkbox"/>	MULTISPECIES: nitrogenase iron protein [Nostocaceae]	<a href="#">Nostocaceae</a>	605	605	100%	0.0	99.66%	295	<a href="#">WP_011320610.1</a>
<input checked="" type="checkbox"/>	nitrogenase iron protein [Anabaena subtrlica]	<a href="#">Anabaena subtr...</a>	601	601	100%	0.0	98.98%	295	<a href="#">WP_190408029.1</a>
<input checked="" type="checkbox"/>	nitrogenase iron protein [Nostoc sp. TCL26-01]	<a href="#">Nostoc sp. TCL2...</a>	601	601	100%	0.0	98.31%	295	<a href="#">WP_179049315.1</a>
<input checked="" type="checkbox"/>	TPA: nitrogenase iron protein [Anabaena sp.]	<a href="#">Anabaena sp.</a>	598	598	100%	0.0	98.31%	295	<a href="#">HFS09990.1</a>
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ... <a href="#">Anabaena sp. L...</a>	<a href="#">Anabaena sp. L...</a>	594	594	99%	0.0	98.29%	294	<a href="#">P33178.1</a>

# 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	<a href="#">CNDBG98Y013</a> Search expires on 06-18 10:37 am <a href="#">Download All</a>
Program	BLASTP <a href="#">Citation</a>
Database	nr <a href="#">See details</a>
Query ID	Icl Query_68648
Description	None
Molecule type	amino acid
Query Length	295
Other reports	<a href="#">Distance tree of results</a> <a href="#">Multiple alignment</a> <a href="#">MSA viewer</a> <a href="#">?</a>

[Descriptions](#) Graphic Summary Alignments Taxonomy

## Sequences producing significant alignments

Download [New](#)

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. Iden	Acc. Len	Accession
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ... <a href="#">Nostoc sp. PCC...</a>	<a href="#">Nostoc sp. PCC...</a>	606	606	100%	0.0	100.00%	295	<a href="#">Q51296.1</a>
<input checked="" type="checkbox"/>	MULTISPECIES: nitrogenase iron protein [Nostocaceae]	<a href="#">Nostocaceae</a>	605	605	100%	0.0	99.66%	295	<a href="#">WP_011320610.1</a>
<input checked="" type="checkbox"/>	nitrogenase iron protein [Anabaena subtrlica]	<a href="#">Anabaena subtr...</a>	601	601	100%	0.0	98.98%	295	<a href="#">WP_190408029.1</a>
<input checked="" type="checkbox"/>	nitrogenase iron protein [Nostoc sp. TCL26-01]	<a href="#">Nostoc sp. TCL2...</a>	601	601	100%	0.0	98.31%	295	<a href="#">WP_179049315.1</a>
<input checked="" type="checkbox"/>	TPA: nitrogenase iron protein [Anabaena sp.]	<a href="#">Anabaena sp.</a>	598	598	100%	0.0	98.31%	295	<a href="#">HFS09990.1</a>
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ... <a href="#">Anabaena sp. L...</a>	<a href="#">Anabaena sp. L...</a>	594	594	99%	0.0	98.29%	294	<a href="#">P33178.1</a>

## Filter Results

Organism only top 20 will appear  exclude

Type common name, binomial,

+ Add organism

Percent Identity E value

 to 

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.

Subject

A A B D E F G H I J



Percent identity: 100%

Query

A A A A A B D E F G H I J K L M N O P Q

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!

Subject

A A B D E F G H I J  
| | | | | | | | | |

Percent identity: 100%

Query

A A A A A B D E F G H I J K L M N O P Q

Query cover: 50%

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](#)

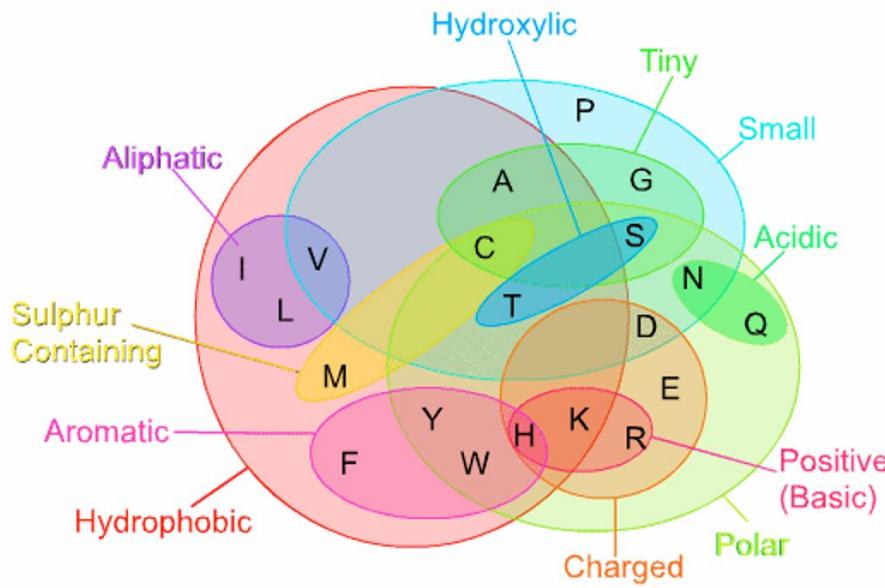
▼ [Next Match](#) ▲ [Previous](#)

	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	TDENIRQIAFYGKGGIGKSTTSONTLAAMAEAMGORIMIVGCDPKADSTRILMLHAKAKTTV				61
Sbjct	3	TD NIRQIAFYGKGGIGI				H+KA+TTV
Query	62	TDGNIRQIAFYGKGGIGI				HSKAQTTV
Sbjct	63	LHLAAERGAVEDLELHE				
Query	62	L LAAERGAVEDLELHE				EENGAYQD
Sbjct	63	LSLAAE				EENGAYQD
Query	122	LDFVSY	GPF			EENGAYQD
Sbjct	123	LDFVSYDVLCDDVVCGGFAMPIREGKAQEIIYIVTSGEMMAMYAANNIARG+LK				121
Query	122	LDFVSYDVLCDDVVCGGFAMPIREGKAQEIIYIVTSGEMMAMYAANNIARG+LK				YAHSGGVR
Sbjct	123	YAHSGGVR				181
Query	182	YAHSGGVR				
Sbjct	183	LGGLICNSRKTDRDREELIENLAERLNTQMIHVPRDNIVQHAELRRMTVNEYAPDSNQGQ				182
Query	182	LGGLICNSRKTDRDREELIENLAERLNTQMIHVPRDNIVQHAELRRMTVNEYAPDSNQGQ				241
Sbjct	183	LGGLICNSRNVDREIELIETLAKRLNTQMIHYVPRDNIVQHAELRRMTVNEYAPDSQGN				242
Query	242	LGGLICNSRNVDREIELIETLAKRLNTQMIHYVPRDNIVQHAELRRMTVNEYAPDSQGN				
Sbjct	243	EYRALAKKIINNDKLTIPTPIEMDELEALLIEYGILDSSKHAEIIGKPA				291
		EYR LAKKIINN LTIPTPIEM+ELE LLIE+GIL+ D +++GK A				
		EYRTLAKKIINNKNLTIPTPIEMEELEELLIEFGILESDENTEKLVGKAA				292

Base change but 'new' amino acid has physiochemically similar properties compared to previous amino acid

GAP

# '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** tryptophan (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](#) ▲ [Previous](#)

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		TDENIRQIAFYGKGGIGKSTTSQNTLAAMAEQMQRIMIVGCDPKADSTRMLHAKAKTTV			61
		TD NIRQIAFYGKGGIGKSTTSQNTLAAMAEQMQRIMIVGCDPKADSTRMLH+KA+TTV			
Sbjct 3		TDGNIRQIAFYGKGGIGKSTTSQNTLAAMAEQMQRIMIVGCDPKADSTRMLHSKAQTTV			62
Query 62		LHLAAERGAVEDLELHEVMLTGFRGVRCVESGGPEPGVGCAGRGIIITAINFLEENGAYQD			121
		L LAAERGAVEDLELHEVMLTGFRGV+CVESGGPEPGVGCAGRGIIITAINFLEENGAYQD			
Sbjct 63		LSLAAERGAVEDLELHEVMLTGFRGVKCVCESGGPEPGVGCAGRGIIITAINFLEENGAYQD			122

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

472759\_2\_2435  
631362\_8\_478



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
<input checked="" type="checkbox"/>	<a href="#">TPA: MAG TPA: ParA [Caudovirales sp.]</a>	<a href="#">Caudovirales sp.</a>	47.4	47.4	80%	9e-04	20.32%	242	<a href="#">DAF22519.1</a>
<input type="checkbox"/>	<a href="#">AAA family ATPase [Enterococcus phage EF62phi]</a>	<a href="#">Enterococcus phage EF62phi</a>	42.7	42.7	25%	0.025	28.75%	267	<a href="#">YP_006218698.1</a>
<input type="checkbox"/>	<a href="#">TPA: MAG TPA: ParA [Myoviridae sp.]</a>	<a href="#">Myoviridae sp.</a>	42.4	42.4	24%	0.032	33.33%	256	<a href="#">DAE74853.1</a>

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 KGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLHAKATTVVLHAAERGAVED 73  
KG + G + TT + QN A + ++++++ D + + T L K K + L++

Query	74	LELHEVMLTGFRGVRC <del>VESGGPEPGVGCAGRGIIITAINFL--EENGAYQDLDVFVSYDV--</del> +EV + VE G +N L E D D++ D	129
Sbjct	63	<del>--NEVDINT-----AVEKDFIAGSKFLVGENTEIELNLKLNELQQLKLTDYDIIIDTPP</del>	114

Query 187 CN--SRKTDREALIENL---AERLNTOQMIFHVPRDNI-VQHAEELRRMTVNEYAPDSNQG 240  
++T +. AE+INT++++ RD+I ++ ++ + + +VA S G

Subject 169 VTRFNKRTILGRTMLN

Query	241	QEYRALAKKII	251
		++YRAL K+I+	
Sbjct	229	RDYRALVKETI	239

# Pairwise alignment vs multiple sequence alignment

- BLAST is pairwise alignment (limited to comparing 2 sequences)
- A multiple sequence alignment compares well... multiple sequences (necessary for building phylogenetic trees)

472759\_2\_2435 -----MAN-----VITMRQMLEAG  
631362\_8\_478 MGVAEVQPKRSRQMID-----VSMRQMLQAG

VS

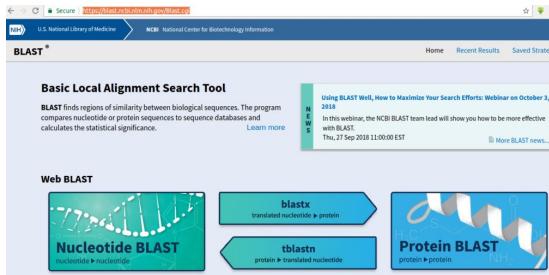
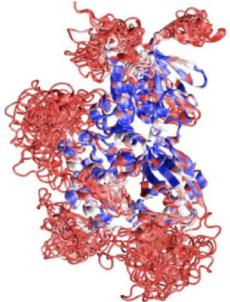
1003202\_1\_49 -----MSKISP-----INIKELLDAGVHF  
1033991\_1\_1545 -----MALPD-----FSMRQOLLEAGVHF  
1356861\_1\_326 -----MSL-----MKEMLSAGVHF  
360116\_9\_42 -----MTD-----ITMRQOLLEAGVHF  
1172194\_4\_612 -----MSN-----ITMRQOLLEAGVHF  
638300\_3\_40 -----MSK-----VSMRELFEAGAHF  
637389\_71\_3 -----MSSPN-----VSMRALLEAGAHF  
933093\_1\_1630 -----MNN-----VSMRELLEAGAHF  
1304275\_18\_53 -----MSE-----ISMROLLEAGVHF  
380358\_1\_1993 -----MPQ-----VTMRQMLEAGVHF  
1214121\_48\_53 -----MPQ-----VTMRQMLEAGVHF  
153948\_2\_2974 -----MS-----VTMRQMLEAGVHF  
428406\_4\_1339 -----MS-----VTMRQMLEAGVHF  
472759\_2\_2435 -----MAN-----VTMRQMLEAGVHF  
631362\_8\_478 MGVAEVQPKRSRQMID-----VSMRQMLQAGVHF  
765912\_1\_699 -----MRQMLEAGVHF  
765910\_1\_112 -----MRQMLEAGVHF  
572477\_1\_2031 -----MRQMLEAGVHF  
768671\_6\_233 -----MRQMLEAGVHF  
1142511\_1\_356 -----MIN-----LSIRNMIQSGVHF  
36870\_2\_414 -----MLN-----LSMKDMIQSGVHF  
395493\_2\_3193 -----MA-----VSMRQMLEAGVHF  
555778\_1\_1453 -----MAS-----VSMRQMLEAGVHF  
697282\_1\_766 -----MAA-----VSMRQMLEAGVHF  
1091494\_1\_2981 -----MAA-----VSMRQMLEAGVHF  
1255043\_1\_1929 -----MSF-----VTMRQMLEAGVHF  
519989\_16\_24 -----MPN-----VTMRQMLEAGVHF

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.

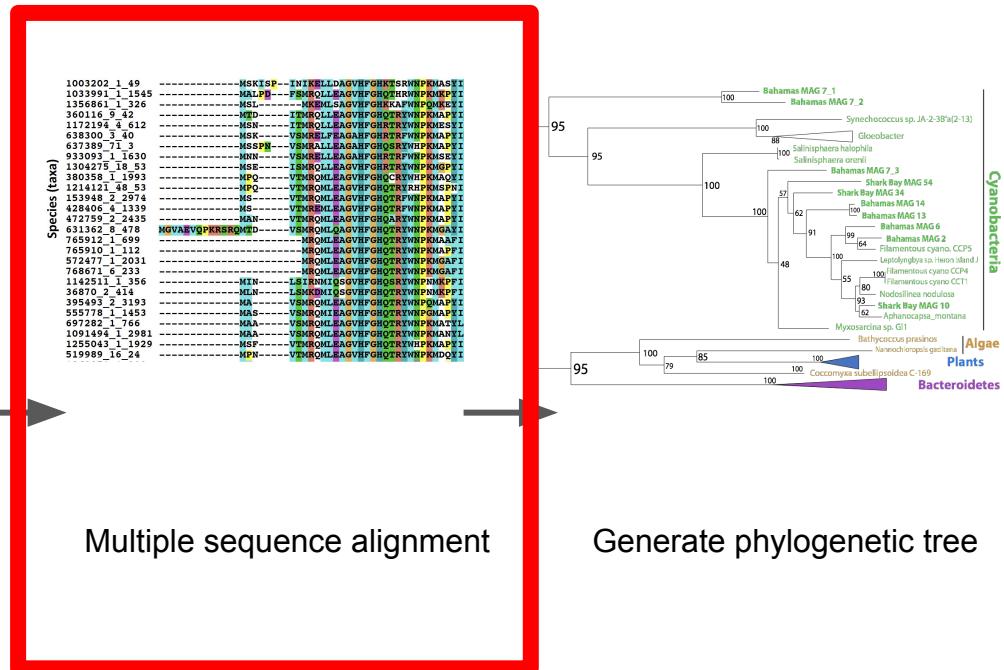
# Overview of usual steps in creating a phylogenetic tree

STEPS:



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

Homology search using BLAST



# Gather sequences for multiple sequence alignment

## 1. Uncheck 'select all'

◀ 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 E value Query Coverage

to to to to

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments

Download New Select columns Show 100 ?

select all 10 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	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
<input type="checkbox"/>	MULTISPECIES: nitrogenase iron protein [Nostocaceae]	Nostocaceae	605	605	100%	0.0	99.66%	295	WP_011320610.1
<input checked="" type="checkbox"/>	nitrogenase iron protein [Anabaena subtropica]	Anabaena subtr...	601	601	100%	0.0	98.98%	295	WP_190408029.1
<input type="checkbox"/>	nitrogenase iron protein [Nostoc sp. TCL26-01]	Nostoc sp. TCL2...	601	601	100%	0.0	98.31%	295	WP_179049315.1
<input type="checkbox"/>	TPA: nitrogenase iron protein [Anabaena sp.]	Anabaena sp.	598	598	100%	0.0	98.31%	295	HFS09990.1
<input type="checkbox"/>	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

# 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)

Screenshot of a BLAST search results page showing how to gather sequences for multiple sequence alignment.

The page includes:

- Job Title: Protein Sequence
- RID: CNDBG98Y013
- Program: BLASTP
- Database: nr
- 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 section:

- Organism: only top 20 will appear
- Type common name, binomial, taxid or group name
- + Add organism
- Percent Identity, E value, Query Coverage filters
- Filter and Reset buttons

Table of sequences producing significant alignments:

Sequences producing significant alignments										
		Download		Select columns		Show		MSA Viewer		?
		GenPept		Graphics		Distance tree of results		Multiple alignment		New MSA Viewer
Descriptions		Graphic Summary		Alignments		Taxonomy				
<input type="checkbox"/> select all 10 sequences selected										
	Description		Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	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
<input type="checkbox"/>	MULTISPECIES: nitrogenase iron protein [Nostocaceae]		Nostocaceae	605	605	100%	0.0	99.66%	295	WP_011320610.1
<input checked="" type="checkbox"/>	nitrogenase iron protein [Anabaena subtropica]		Anabaena subtr...	601	601	100%	0.0	98.98%	295	WP_190408029.1
<input type="checkbox"/>	nitrogenase iron protein [Nostoc sp. TCL26-01]		Nostoc sp. TCL2...	601	601	100%	0.0	98.31%	295	WP_179049315.1
<input type="checkbox"/>	TPA: nitrogenase iron protein [Anabaena sp.]		Anabaena sp.	598	598	100%	0.0	98.31%	295	HFS09990.1
<input type="checkbox"/>	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

A pink arrow points to the "select all" checkbox at the top left of the table.

# 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 (lower values)!

The screenshot shows a BLAST search results page with the following details:

- 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** section:

- 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 [ ]

**Descriptions** tab is selected. The table below lists "Sequences producing significant alignments".

Sequences producing significant alignments									
<input type="checkbox"/> select all	10 sequences selected								
	Description		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
<input checked="" type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ...		Nostoc sp. PCC...	606	606	100%	0.0	100.00%	295
<input type="checkbox"/>	MULTISPECIES: nitrogenase iron protein [Nostocaceae]		Nostocaceae	605	605	100%	0.0	99.66%	295
<input checked="" type="checkbox"/>	nitrogenase iron protein [Anabaena subtropica]		Anabaena subtr...	601	601	100%	0.0	98.98%	295
<input type="checkbox"/>	nitrogenase iron protein [Nostoc sp. TCL26-01]		Nostoc sp. TCL2...	601	601	100%	0.0	98.31%	295
<input type="checkbox"/>	TPA: nitrogenase iron protein [Anabaena sp.]		Anabaena sp.	598	598	100%	0.0	98.31%	295
<input type="checkbox"/>	RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ...		Anabaena sp. L...	594	594	99%	0.0	98.29%	294

# 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)'

The screenshot shows a BLAST search results page in a web browser. The search parameters are:

- Program: BLASTP
- Database: nr
- Query ID: Icl|Query\_68648
- Description: None
- Molecule type: amino acid
- Query Length: 295

The 'Descriptions' tab is selected. Below it, a table titled "Sequences producing significant alignments" lists 10 sequences. The first few entries are:

- RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe protein; AltName: Full=Nitrogenase ...
- MULTISPECIES: nitrogenase iron protein [Nostocaceae]
- nitrogenase iron protein [Anabaena subtropica]
- nitrogenase iron protein [Nostoc sp. TCL26-01]
- TPA: nitrogenase iron protein [Anabaena sp.]

An arrow points from the "Download" button to the "FASTA (complete sequence)" option in the dropdown menu. The "Select columns" and "Show 100" buttons are also visible.

GenBank (complete sequence)	Its	Multiple alignment	New MSA Viewer
100% 0.0 100.00% 295 Q51296_1			
Hit Table (text)	100% 0.0 99.66% 295 WP_011320610_1		
Hit Table (CSV)	100% 0.0 98.98% 295 WP_190408029_1		
Text	100% 0.0 98.31% 295 WP_179049315_1		
Descriptions Table (CSV)	100% 0.0 98.31% 295 HFS09990_1		
XML	100% 0.0 98.29% 294 P33178_1		
ASN.1	100% 0.0 97.63% 295 WP_193883645_1		
Nostocaceae	592 592 100% 0.0 97.97% 295 WP_190698764_1		
Nostocaceae	591 591 100% 0.0 96.61% 295 WP_010995626_1		
Nostocales	589 589 99% 0.0 96.94% 296 WP_096647180_1		
Nostoc sp. PCC ...	588 588 100% 0.0 96.95% 295 WP_0151141		

# Gather sequences for multiple sequence alignment

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

```
seqdump (2).txt
>051296.1 RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe
protein; AltName: Full=Nitrogenase component II; AltName: Full=Nitrogenase reductase
[Nostoc sp. PCC 6720]
MTDENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHAKAKTTVLHLLAERGADEDLEHEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDLDFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLLGICNSRKTDREAEELIENLAERLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPMEMDEALLIEYEYGIIDDDSKHAEIIGKPAEATK
>WP_190408029.1 nitrogenase iron protein [Anabaena subtropica]
MTDENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHAKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDLDFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLLGICNSRKTDREAEELIENLAERLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPMEMDEALLIEYEYGIIDDDSKHAEIIGKPAEATK
>WP_045868724.1 MULTISPECIES: nitrogenase iron protein [Nostocales]
MTDENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHAKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDLDFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLLGICNSRKTDREAEELIENLAERLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPMEMDEALLIEYEYGIIDDDSKHAEIIGKPAEASAK
>WP_096622974.1 nitrogenase iron protein [Calothrix sp. NIES-3974]
MSDEKIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDLDFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLLGICNSRKVDREAEELIENLAERLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNEKLTIP TPMEMDEALLIEYEYGIIDDDSKHAEIIGKPAEAK
>WP_026731053.1 nitrogenase iron protein [Fischerella sp. PCC 9605]
MSDDKIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDLDFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLLGICNSRKVDREAEELIENLAERLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNEKLTIP TPMEMDEALLIEYEYGIIDDDSKHAEIIGKPAEASAK
>RUR73888.1 nitrogenase iron protein [Chlorogloeopsis fritschii PCC 6912]
MTENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDLDFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMMA
YAANNIARGILKYAHSGGVRGLLGICNSRKVDREAEELIENLAERLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNKLTIP TPMEMDEALLIEYEYGIIDDDSKHAEIIGKPAEATK
>WP_196812786.1 MULTISPECIES: nitrogenase iron protein [Halosiphonaceae]
MMTENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LAGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDLDFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLLGICNSRKVDREAEELIENLAERLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNNQLQIPTPMEMDEELLEELLIEFGILESEENAAKLGKPAESTAK
>WP_167722396.1 nitrogenase iron protein [Tolyphothrix sp. PCC 7910]
MSIDSKIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
MLTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDVFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLLGICNSRNVDREIELIETLAKRLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
SNEYRALAKKIINNNQLNLIPTPMEMDEELLEELLIEFGILESEENAAKLGKPAESTAK
>WP_190561792.1 nitrogenase iron protein [Anabaena sphaerica]
MATDGNIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEHEV
MLTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDVFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLLGICNSRNVDREIELIETLAKRLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSDQ
GNEYRTLAKKIINNNQLNLIPTPMEMDEELLEELLIEFGILESEENAAKLGKPAEAVPKK
>WP_190561792.1 nitrogenase iron protein [Anabaena sphaerica]
MATDGNIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEHEV
MLTGFRGVRCVSESGGPEPGVGCAcRGIIITAIFLEENGAYQDVFVSYDVLGDVVCGGFAMPIREGKAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLLGICNSRNVDREIELIETLAKRLNTOMIHFVPRDNIVQHAELRRMTVNEYAPDSDQ
GNEYRTLAKKIINNNQLNLIPTPMEMDEELLEELLIEFGILESEENAAKLGKPAEAVPKK
```

# 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!

```
seqdump (2).txt
>P_051296.1 RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe
protein; AltName: Full=Nitrogenase component II; AltName: Full=Nitrogenase reductase
[Nostoc sp. PCC 6720]
MTDENIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHAKAKTTVLHLLAERGADEDLEHEVM
LTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPIEMDEALLELIYEYGILDQSKHAEIIGKPAEATK
>WP_190408029.1 nitrogenase iron protein [Anabaena subtropica]
MTDENIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHAKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPIEMDEALLELIYEYGILDQSKHAEIIGKPAEATK
>WP_045868724.1 MULTISPECIES: nitrogenase iron protein [Nostocales]
MTDENIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHAKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPIEMDEALLELIYEYGILDQSKHAEIIGKPAEASAK
>WP_096622974.1 nitrogenase iron protein [Calothrix sp. NIES-3974]
MSDEKIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNEKLTIPPTPIEMDEALLELIYEYGILDQSKHAEIIGKPAEAK
>WP_026731053.1 nitrogenase iron protein [Fischerella sp. PCC 9605]
MSDDKIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHAKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNTKLTIPTPIEMDEALLELIYEYGILDQSKHAEIIGKPAEASAK
>RUR738811 nitrogeanase iron protein [Chlorogloeopsis fritschii PCC 6912]
MTENIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNTKLTIPTPIEMDEALLELIYEYGILDQSKHAEIIGKPAEASAK
>WP_196812786.1 MULTISPECIES: nitrogenase iron protein [Hapalosiphonaceae]
MMTENIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LAGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNNQLQIPTPIEMDEALLEELLIFFGILESEENAAKLGKPAESTAK
>WP_167722396.1 nitrogeanase iron protein [Tolyphothrix sp. PCC 7910]
MSIDKIROIAFYGKGGIGKSTSNTLAAMAEQMRILIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
MLTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDVFVSYDVLGVVCGGFAMPIREKNAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLGLICNSRNVDREIELIETLAKRLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
SNEYRALTAKIINNNQLQIPTPIEMDEALLEELLIFFGILESEENAAKLGKPAESTAK
>WP_200321811.1 nitrogenase iron protein [Sphaerotilus sphaeroides]
MTTDANIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEHEV
MLTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLGLICNSRNVDREIELIETLAKRLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
GNEYRILAKKIINNNNLNIPITPIEMEELLEELLIFFGILESEENAAKLGKPAEVPKK
>WP_190561792.1 nitrogeanase iron protein [Anabaena sphaerica]
MATDGNIROIAFYGKGGIGKSTSNTLAAMAEQMRIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEHEV
MLTGFRGVCKVESGGPEPGVGCAGRGIIITAINFOLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLGLICNSRNVDREIELIETLAKRLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSDQ
GNEYRTLAKKIINNNNLNIPITPIEMEELLEELLIFFGILESEENAAKLGKPAEAVPKK
```

# 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'

```
seqdump (2).txt
>051296.1 RecName: Full=Nitrogenase iron protein; AltName: Full=Nitrogenase Fe
protein; AltName: Full=Nitrogenase component II; AltName: Full=Nitrogenase reductase
[Nostoc sp. PCC 6720]
MTDENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHAKAKTTVLHLLAERGADEDLEHEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPMEMDEALLIEYGILODDDSKHAEIIGKPAEATK
>WP_045408029.1 nitrogenase iron protein [Anabaena subtropica]
MTDENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHAKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPMEMDEALLIEYGILODDDSKHADIIIGKPAEATK
>WP_045868724.1 MULTISPECIES: nitrogenase iron protein [Nostocales]
MTDENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHAKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPMEMDEALLIEYGILODDDSKHADIIIGKPAEASAK
>WP_096622974.1 nitrogenase iron protein [Calothrix sp. NIES-3974]
MSDEKIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNEKLTIP TPMEMDEALLIEYGILODDDSKHADIIIGKPAEAK
>WP_026731053.1 nitrogenase iron protein [Fischerella sp. PCC 9605]
MSDDKIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNTKLTIPTPMEMDEALLIEYGILODDDSKHADIIIGKPAEASAK
>RUR73888.1 nitrogenase iron protein [Chlorogloeopsis fritschii PCC 6912]
MTENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
YAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNDKLTIPTPMEMDEALLIEYGILODDDSKHADIIIGKPAEATK
>WP_196812786.1 MULTISPECIES: nitrogenase iron protein [Halosiphonaceae]
MMTENIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
LAGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMMA
MYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIENLAERLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
QEYRALAKKIINNTKLTIPTPMEMDEALLIEYGILODDDSKHADIIIGKPAEASAK
>WP_167722396.1 nitrogenase iron protein [Tolyphothrix sp. PCC 7910]
MSIDSKIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEEEVM
MLTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIETLAKRLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
SNEYRALAKKIINNNQLQIPTPMEMDEELLEELIEFGILESEENAAKLIGKPAESTAK
>WP_200321811.1 nitrogenase iron protein [Sphaerospermopsis aphaniopisoides]
MTTDANIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEHEV
MLTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIETLAKRLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSNQ
GNEYRILAKKIINNNNLNIPITPMEELLEELIEFGILESDENTEVGKPAIEVPVKK
>WP_190561792.1 nitrogenase iron protein [Anabaena sphaerica]
MATDGNIROIAFYGKGGIGKSTSNTLAAAMAEMGORIMIVGCDPKADSTRMLHLSKAQTTVLHLLAERGADEDLEHEV
MLTGFRGVRCVSESGGPEPGVGCAcRGIIITAINFLEENGAYQDLDFVSYDVLGVVCGGFAMPIREGKAQEIYIVTSGEMM
AMYAANNIARGILKYAHSGGVRGLGLICNSRKTDREALELIETLAKRLNTOMIHFPVRDNIVQHAELRRMTVNEYAPDSDQ
GNEYRTLAKKIINNNKLTIPTPMEMEELEELIEFGILESDENTKELVGKAAATEAPVKK
```

# Terminal time!!

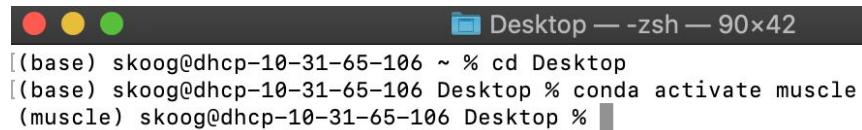
1. Open your terminal and ‘cd’ into your Desktop directory



```
[(base) skoog@dhcp-10-31-65-106 ~ % cd Desktop  
(base) skoog@dhcp-10-31-65-106 Desktop % ]
```

# Terminal time!!

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



```
Desktop — zsh — 90x42
[(base) skoog@dhcp-10-31-65-106 ~ % cd Desktop
[(base) skoog@dhcp-10-31-65-106 Desktop % conda activate muscle
(muscle) skoog@dhcp-10-31-65-106 Desktop %
```

# Terminal time!!

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

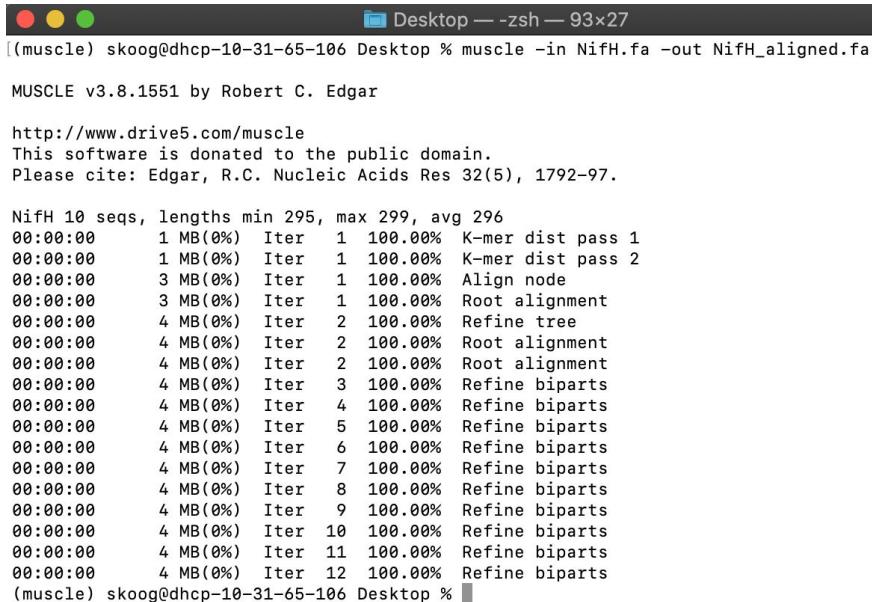
```
skoog — -zsh — 80x24
(muscle) skoog@dhcp-10-31-65-106 ~ % muscle -in NifH.fa -out NifH_aligned.fa
```

File going in

Name of outputted file  
(you pick the name)

# Terminal time!!

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



A screenshot of a terminal window titled "Desktop — zsh — 93x27". The window shows the command "muscle -in NifH.fa -out NifH\_aligned.fa" being run. The output is the MUSCLE v3.8.1551 alignment process, which includes a header about the software being donated to the public domain, citation information, and a detailed log of the alignment steps. The log shows the number of sequences (NifH 10 seqs), lengths (min 295, max 299, avg 296), memory usage (4 MB(0%)), iterations (Iter 1 to Iter 12), and progress (100.00% completion) for K-mer dist pass, Align node, Root alignment, Refine tree, and Refine biparts.

```
(muscle) skoog@dhcp-10-31-65-106 Desktop % muscle -in NifH.fa -out NifH_aligned.fa
MUSCLE v3.8.1551 by Robert C. Edgar
http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

NifH 10 seqs, lengths min 295, max 299, avg 296
00:00:00 1 MB(0%) Iter 1 100.00% K-mer dist pass 1
00:00:00 1 MB(0%) Iter 1 100.00% K-mer dist pass 2
00:00:00 3 MB(0%) Iter 1 100.00% Align node
00:00:00 3 MB(0%) Iter 1 100.00% Root alignment
00:00:00 4 MB(0%) Iter 2 100.00% Refine tree
00:00:00 4 MB(0%) Iter 2 100.00% Root alignment
00:00:00 4 MB(0%) Iter 2 100.00% Root alignment
00:00:00 4 MB(0%) Iter 3 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 4 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 5 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 6 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 7 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 8 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 9 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 10 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 11 100.00% Refine biparts
00:00:00 4 MB(0%) Iter 12 100.00% Refine biparts
(muscle) skoog@dhcp-10-31-65-106 Desktop %
```

# 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)

gaps

```
>DAF22519.1 TPA: MAG TPA: ParA [Caudovirales sp.]
MEIISIINO-----KGGVGKTTAQNLTAGLRLQNKVLLLDLDAQCNLTLLQQATKYKY
NILNVLKNEVDI-NTAVEKDFIAGSKFL-----VGENTEIEENKLKNELQK
LKTDYDYIIIDTPPALSNITINALT-----ASTDIIITITADLLPIQGIVDLYKTVQSI
QKTSNKNLNLIKILVTRFNKRTILGRTMLNSLIDIAEKLNTKVLNTKIR-SISIKESQAK
MTDIFKYARYSTAGRDYRALVKEILEDK-----
```

```
>WP_200321811.1 nitrogenase iron protein [Sphaerospermopsis aphanizomenoides]
MTTDANIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLHSKAQT
TVLSLAAERGAVEDLELHEVMLTGFRGVRCVESGGPEPGVGCAGRGIIITAINFLEEN--G
AYQDLDLDFVSYD---VLGDVVC CGGFAMPIREGKAQEIYIVTSGEMMAMYAANNIARGV--L
KYAHTGGVRLGLICN--SRNV DREIELIETL---AKRLNTQMIHYVPRDNI-VQHAELR
RMTVNEYAPDSNQGNEYRILAKKIINNTNLNIPPTPIEMEELEELLIEFGILESDENTEK
VGKPAIEVPVKK
>WP_190561702.1 nitrogenase iron protein [Anabaena sphaerica]
MATDGNIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLHSKAQT
TVLSLAAERGAVEDLELHEVMLTGFRGVKCVESGGPEPGVGCAGRGIIITAINFLEEN--G
AYQDLDLDFVSYD---VLGDVVC CGGFAMPIREGKAQEIYIVTSGEMMAMYAANNIARGV--L
KYAHSGGVRLGLICN--SRNV DREIELIETL---AKRLNTQMIHYVPRDNI-VQHAELR
RMTVNEYAPDSDQGNEYRTLAKKIINNNKNTIPTPIEMEELEELLIEFGILESDENTEK
VGKAATEAPVKK
```

# Terminal time!!

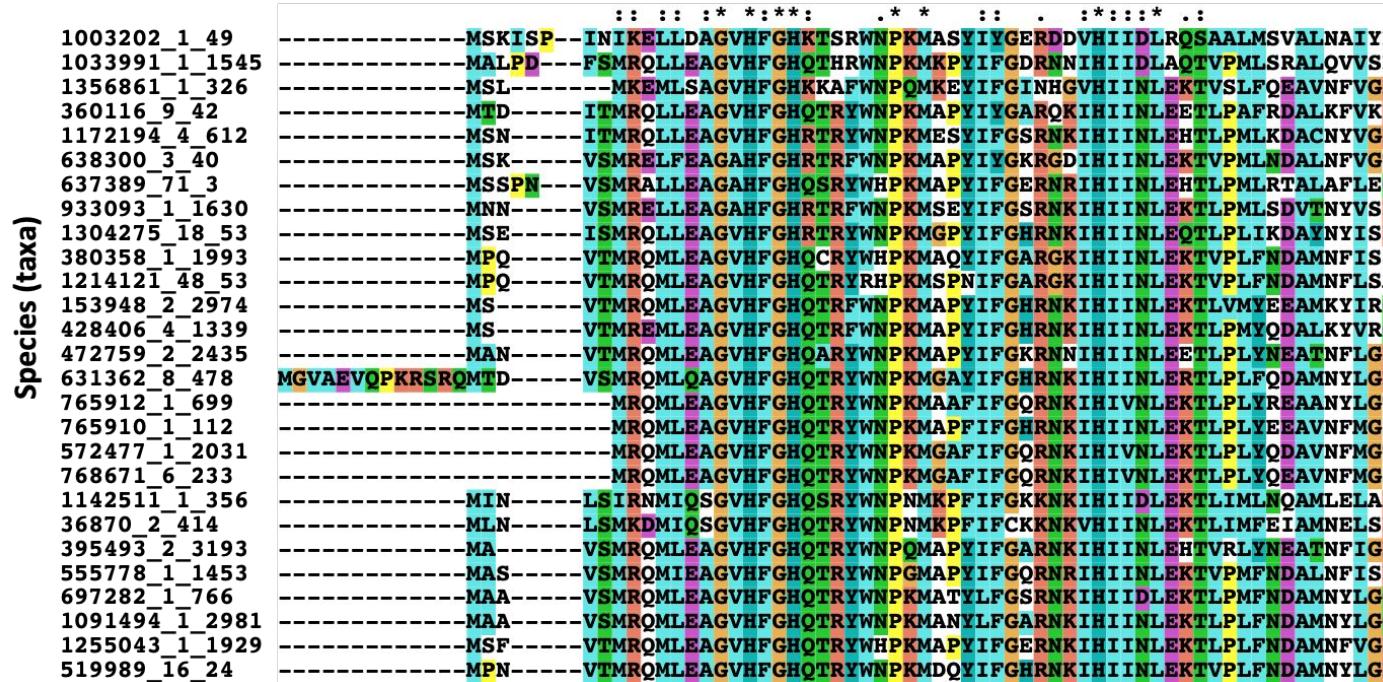
1. Open your terminal and ‘cd’ into your Desktop directory
2. Enter your ‘muscle’ environment
3. Now align those sequences using muscle!
4. Take a look at the alignment! (open the file)
5. Last thing (will come in handy later during tree viewing): remove the first section of the defline (leave the >)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

```
>RUR73888.1 nitrogenase iron protein [Chlorogloeopsis fritschii PCC 6912]
--MTENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLHSKAQT
TVLHLAAERGAVEDLELEEVMLTGFRGVKCVESGGPEPGVGCAGRGIITAINFLEENGAY
QDLDLFSYDVLDVVCGGFAMPIREGKAQEIYIVTSGEMMAMYAANNIARGILKYAHSGG
VRLGGLICNSRKVDREAEELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
SQEYRALAKKIINNDKLTIP TPMEMDDLEALLIEYGILD DDTKHAEIIGK-PAEATSK-
```

↓

```
>Chlorogloeopsis fritschii PCC 6912]
--MTENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRMLHSKAQT
TVLHLAAERGAVEDLELEEVMLTGFRGVKCVESGGPEPGVGCAGRGIITAINFLEENGAY
QDLDLFSYDVLDVVCGGFAMPIREGKAQEIYIVTSGEMMAMYAANNIARGILKYAHSGG
VRLGGLICNSRKVDREAEELIENLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQ
SQEYRALAKKIINNDKLTIP TPMEMDDLEALLIEYGILD DDTKHAEIIGK-PAEATSK-
```

Multiple sequence alignment viewing programs exist



# Multiple ways to align sequences (parameters matter)



Alternative alignment?

**KQ-R-----GEKGKR**  
**KQQR-----KER--- ?**  
---**QRTAAAKEK---**

# How do we determine which sites should be aligned?

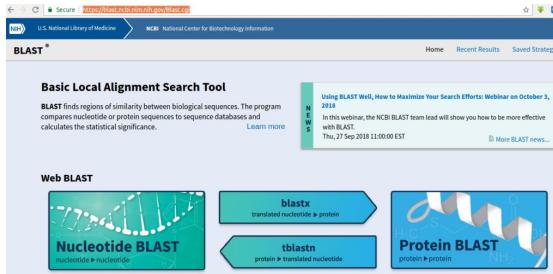
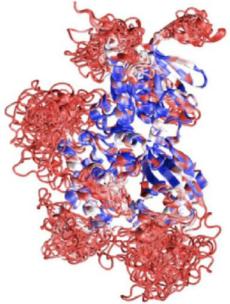
- Alignment Scores: Count well-aligned sites in sequences, scoring matches and penalizing mismatches, and penalizing gaps;
- Scoring matches: better alignments have more sites across homologs with the same residues. Sites with similar amino acids are also more likely to be homologous. Regions with these 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.

# 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

A thick, dark gray arrow pointing to the right, indicating a continuation or next step.

## 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?

# 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

# 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



A screenshot of a macOS terminal window titled "skoog — -zsh — 80x24". The window shows the command `conda deactivate` being run. The terminal output shows the prompt changing from `(muscle)` to `(base)`. A pink arrow points from a callout box to the word "base".

```
(muscle) skoog@dhcp-10-31-65-106 ~ % conda deactivate  
(base) skoog@dhcp-10-31-65-106 ~ %
```

Notice change back to 'base'

# IQ-TREE: Create conda environment

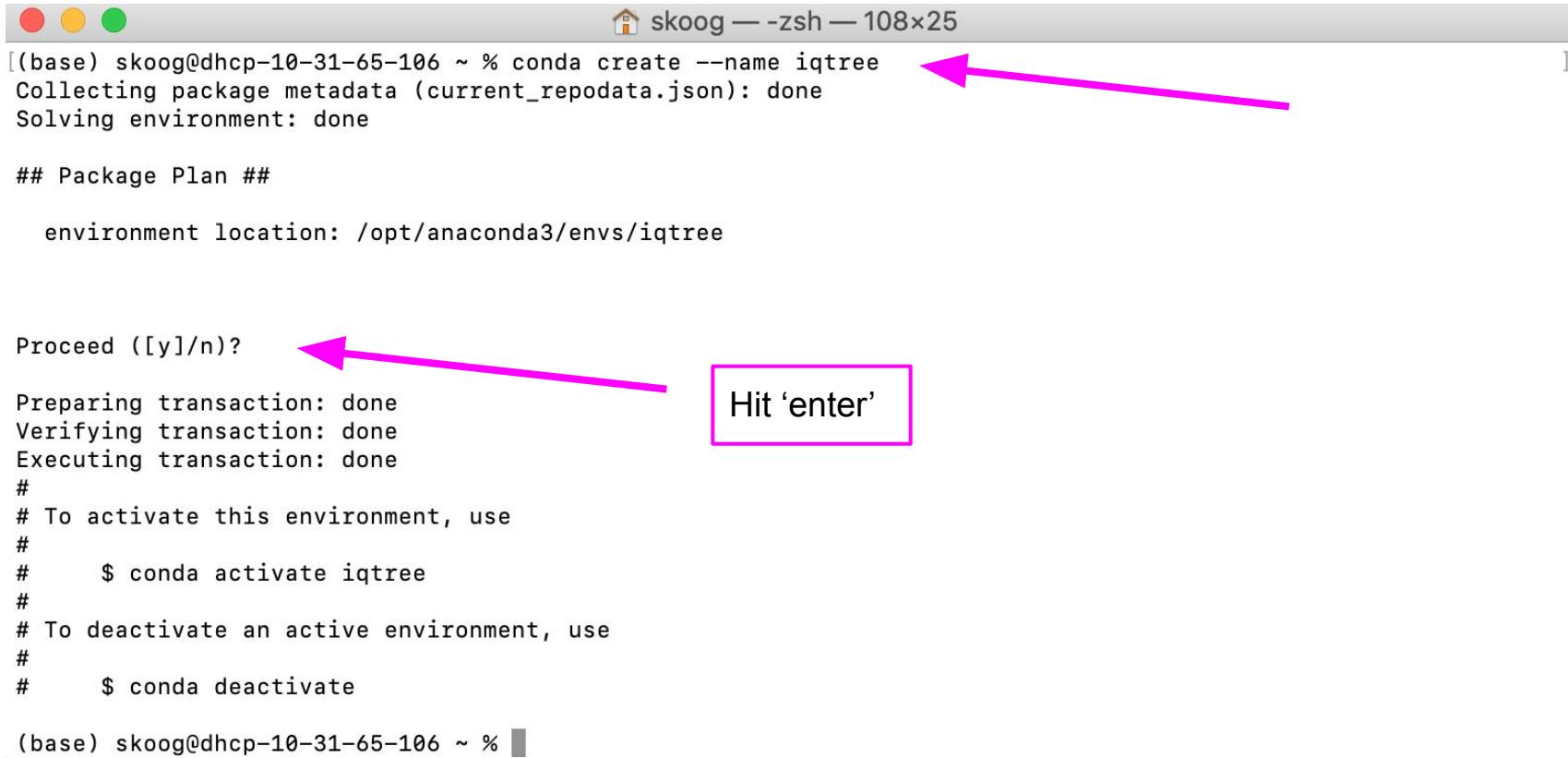
```
skoog — -zsh — 108x25
[(base) skoog@dhcp-10-31-65-106 ~ % conda create --name iqtree
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

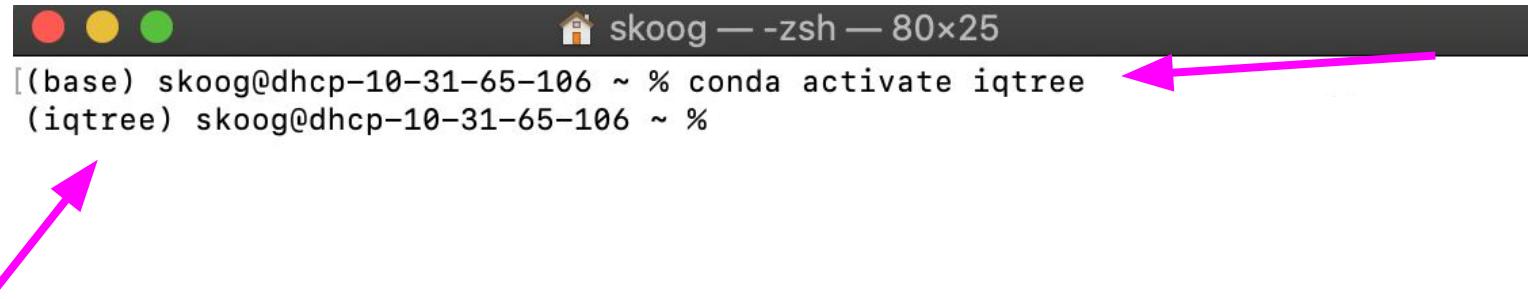
environment location: /opt/anaconda3/envs/iqtree

Proceed ([y]/n)? Hit 'enter'
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
#
# To activate this environment, use
#
#     $ conda activate iqtree
#
# To deactivate an active environment, use
#
#     $ conda deactivate

(base) skoog@dhcp-10-31-65-106 ~ %
```



# MUSCLE: Activate environment



The image shows a terminal window with a dark gray header bar. In the top left corner of the header bar are three colored circles: red, yellow, and green. To the right of these circles is a house icon followed by the text "skoog — -zsh — 80x25". Below the header bar, the main window area contains two lines of text. The first line starts with "(base)" followed by the user's command: "skoog@dhcp-10-31-65-106 ~ % conda activate iqtreetree". The second line shows the result of the command: "(iqtreetree) skoog@dhcp-10-31-65-106 ~ %". Two pink arrows point from the bottom-left towards the terminal window, highlighting the user's input and the resulting environment name.

```
(base) skoog@dhcp-10-31-65-106 ~ % conda activate iqtreetree  
(iqtreetree) skoog@dhcp-10-31-65-106 ~ %
```

# IQ-TREE: Install program inside of environment



skoog — -zsh — 80x25

```
(base) skoog@dhcp-10-31-65-106 ~ % conda activate iqtreetree  
(iqtreetree) skoog@dhcp-10-31-65-106 ~ % conda install -c bioconda iqtreetree
```



# IQ-TREE: Install program inside of environment

```
skoog ~ zsh - 90x42
(iqtreet) skoog@dhcp-10-31-65-106 ~ % conda install -c bioconda iqtreet

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/iqtreet

added / updated specs:
- iqtreet

The following packages will be downloaded:

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

The following NEW packages will be INSTALLED:

  iqtreet           bioconda/osx-64::iqtreet-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
iqtreet-2.0.3          | 3.1 MB      | #####| 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
(iqtreet) skoog@dhcp-10-31-65-106 ~ %
```

Hit 'enter'



# Run IQ-TREE!

```
(iqtree) skoog@dhcp-10-31-65-106 Desktop % iqtree -s NifH_aligned.fa -B 1000
```

Aligned file

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

```
Desktop — zsh — 112x39
(iqtree) skoog@dhcp-10-31-65-106 Desktop % iqtree -s NifH_aligned.fa -B 1000
```

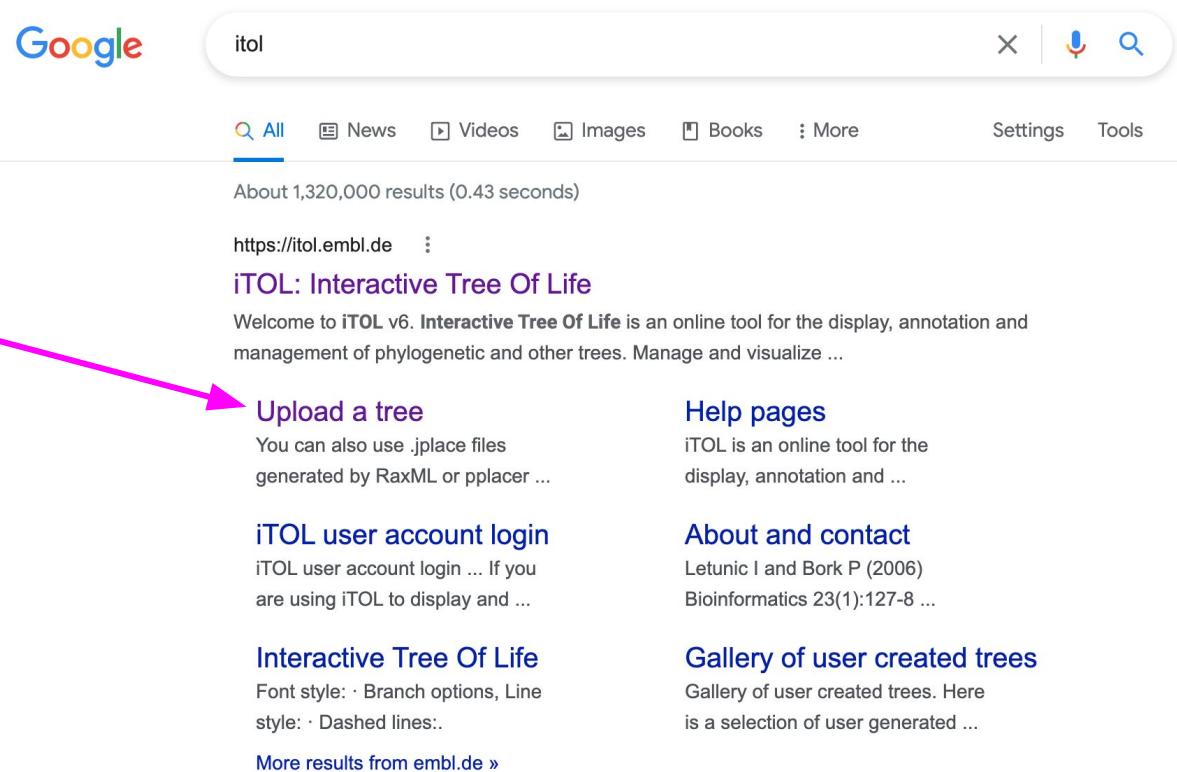
# Run IQ-TREE

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

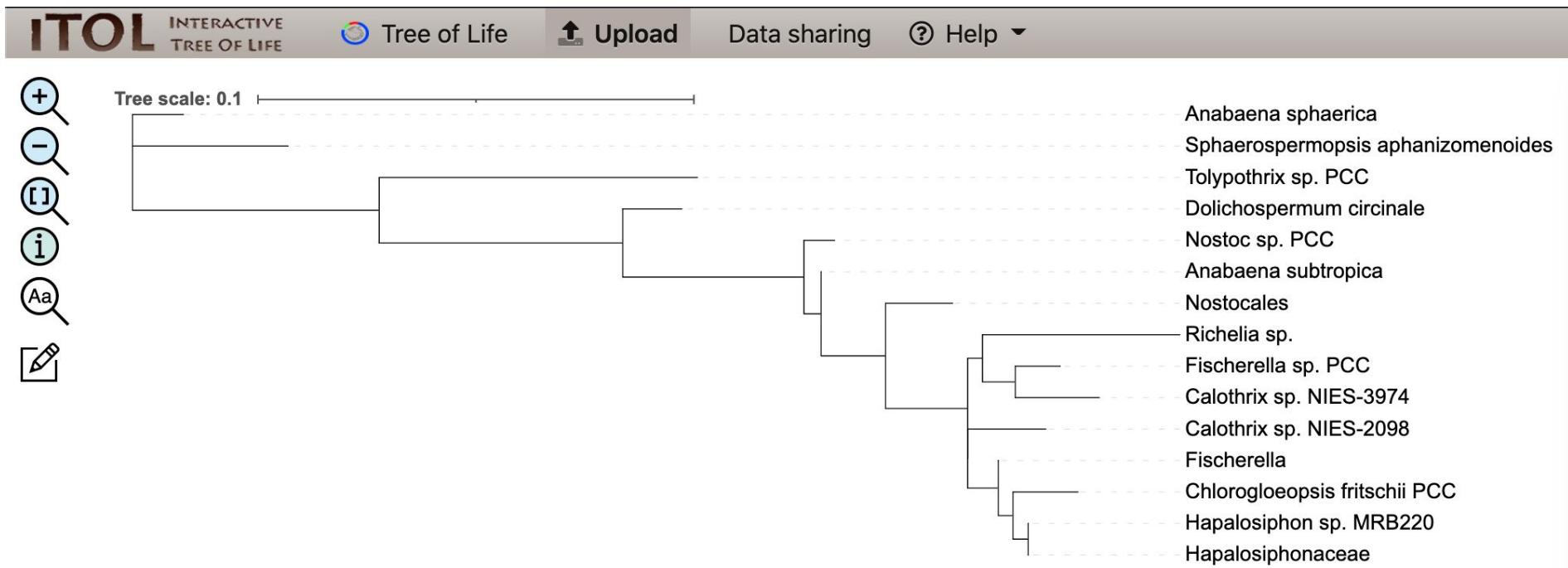


A screenshot of a Google search results page for the query "itol". The search bar at the top contains "itol". Below it, there are several search filters: All (selected), News, Videos, Images, Books, More, Settings, and Tools. The search results section shows the following information:

- https://itol.embl.de** :: **iTOL: Interactive Tree Of Life**  
Welcome to iTOL v6. **Interactive Tree Of Life** is an online tool for the display, annotation and management of phylogenetic and other trees. Manage and visualize ...
- Upload a tree**  
You can also use .jplace files generated by RaxML or pplacer ...
- iTOL user account login**  
iTOL user account login ... If you are using iTOL to display and ...
- Interactive Tree Of Life**  
Font style: · Branch options, Line style: · Dashed lines::  
[More results from embl.de »](#)
- Help pages**  
iTOL is an online tool for the display, annotation and ...
- About and contact**  
Letunic I and Bork P (2006)  
Bioinformatics 23(1):127-8 ...
- Gallery of user created trees**  
Gallery of user created trees. Here is a selection of user generated ...

A large pink arrow points from the text "Upload a tree" in the search results towards the "Upload a tree" link in the "iTOL user account login" section.

# View tree in iTOL



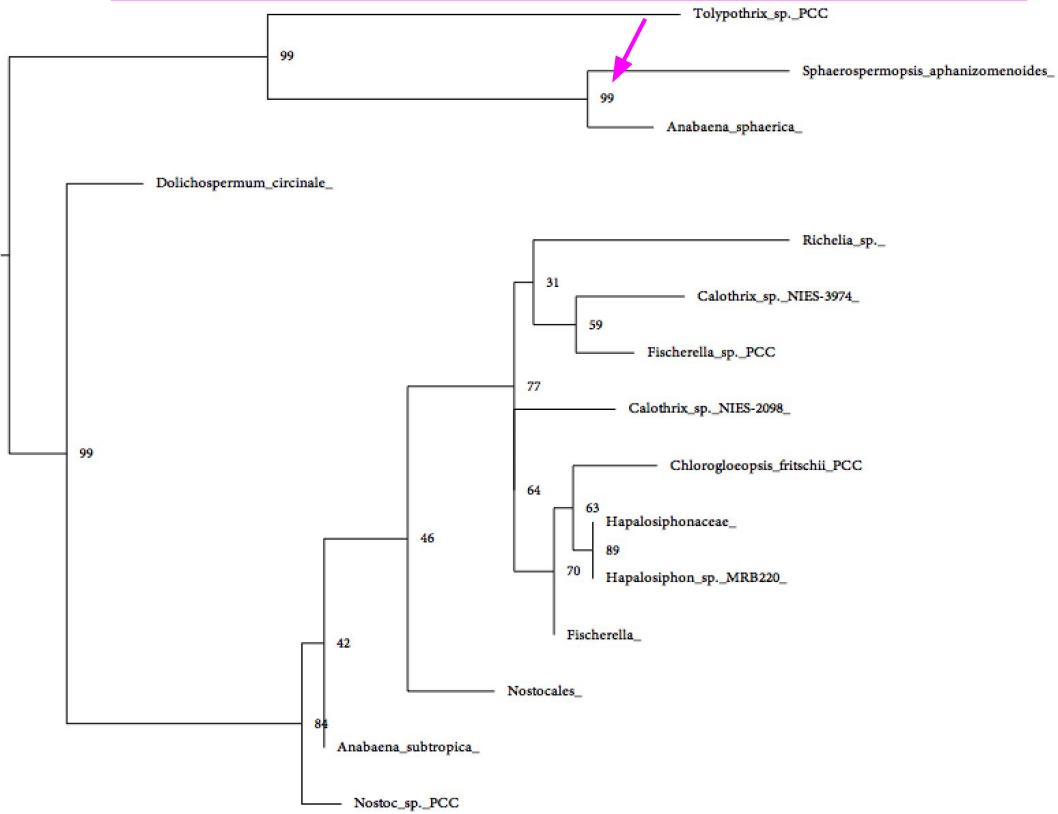
# Other tree visualization softwares

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 (waaaaaaay cheaper than Illustrator (but just as awesome (in my opinion))))))

# Same tree in FigTree

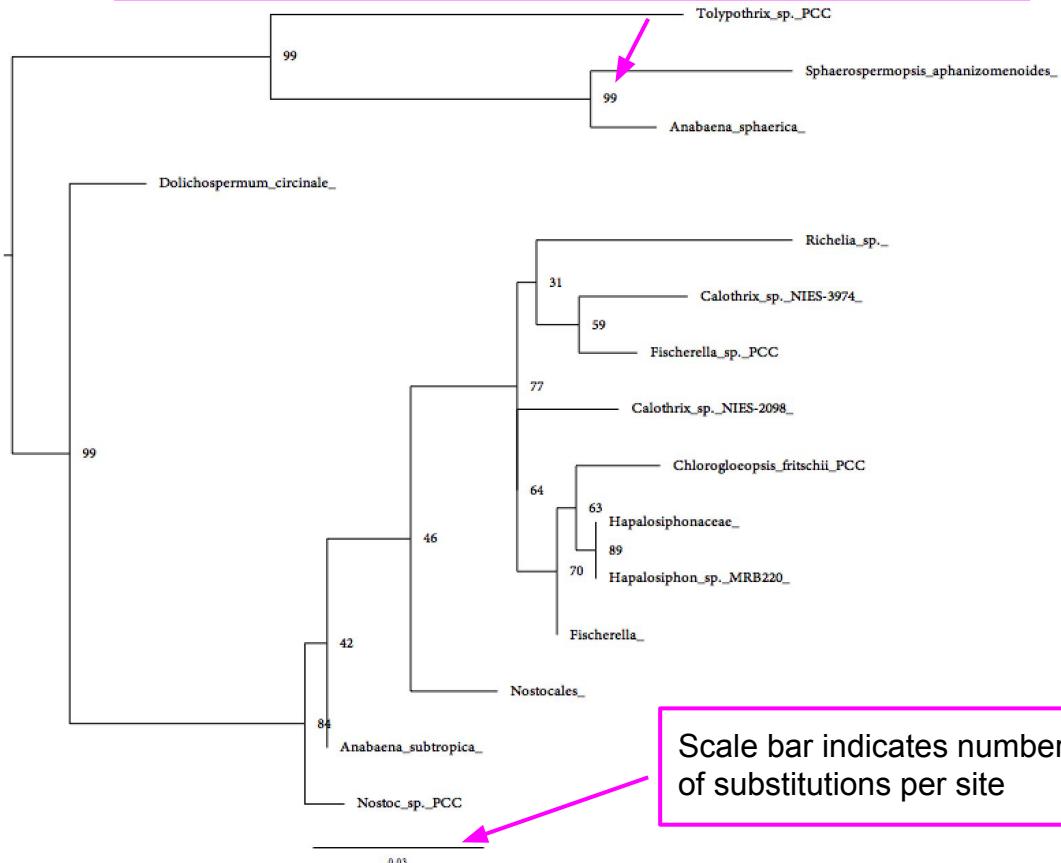
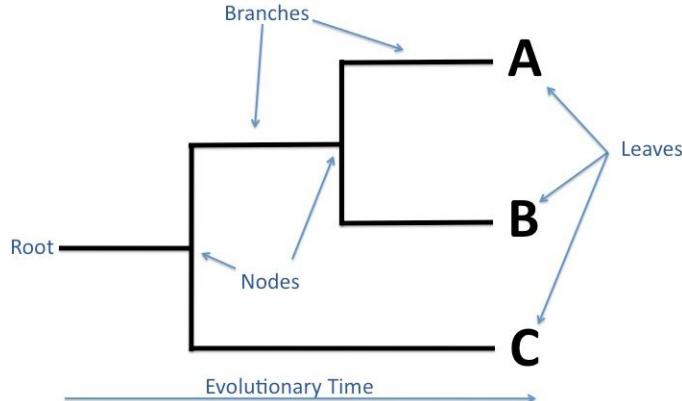
**Bootstrap values:** indicates how many times out of 100, the same branch was observed when repeating the phylogenetic reconstruction (you can be more confident in branches where node bootstrap values are closer to 100)



# 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

ncbi protein database

X |

All Images Videos News Shopping More Settings Tools

About 89,100,000 results (0.72 seconds)

<https://www.ncbi.nlm.nih.gov> › NCBI › Proteins

**Home - Protein - NCBI**

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, ...

**Protein**  
The Protein database is a collection of sequences from ...

**Standard Protein BLAST**  
Standard Protein BLAST. BLASTP programs search protein ...

**Proteins**  
Protein - Protein Clusters - Identical Protein Groups - ...  
[More results from nih.gov »](#)

**Entrez Sequences Quick Start**  
This is a quick start guide for the Entrez Protein, Nucleotide ...

**1796318598|ref|YP\_009724390**  
... syndrome coronavirus 2] 1273 aa protein YP\_009724390.1 GI ...

**Advanced search**  
Protein Advanced Search Builder. Use the builder below to create ...



NCBI Resources How To Sign in to NCBI Help

Protein  Advanced

**COVID-19 Information**  
[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

You did not provide any items for retrieving

**Protein**  
The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

**Using Protein**  
[Quick Start Guide](#)  
[FAQ](#)  
[Help](#)  
[GenBank FTP](#)  
[RefSeq FTP](#)

**Protein Tools**  
[BLAST](#)  
[LinkOut](#)  
[E-Utilities](#)  
[Batch Entrez](#)

**Other Resources**  
[GenBank Home](#)  
[RefSeq Home](#)  
[CDD](#)  
[Structure](#)

## 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