

# Bioinformatics

Bi1X-2010

**Part II: Sequence alignment and BLAST**

Arbel Tadmor

# Overview

- Sequence alignment - basic concepts
- Local vs. global alignment
- BLAST –a local alignment tool
- Exercise: alignment of random sequences
- Exercise: finding homologs of HBA1 with BLAST

# Finding homologues sequences

- **Homology ≠ similarity**

**Homology:** two sequences are descended from a common ancestor, therefore in an alignment identical residues at a site are identical by descent

**Similarity:** merely reflects proportion of sites that are identical

- What changes could occur over time in a sequence?

# Finding homologues sequences

- **Homology ≠ similarity**

**Homology:** two sequences are descended from a common ancestor, therefore in an alignment identical residues at a site are identical by descent

**Similarity:** merely reflects proportion of sites that are identical

- What changes could occur over time in a sequence?
  - Changes that conserve length:
    - Substitutions (one nt mutation)
    - Inversions
  - Changes that do not conserve length :
    - Deletions (e.g. DNApol replication error, unequal crossover, transposon)
    - Insertions (e.g. DNApol replication error, unequal crossover, transposon HGT via transposable elements)

# Matches, Mismatches and Indels

Two aligned, identical characters in an alignment are a **match**.

Two aligned, unequal characters are a **mismatch**.

A character aligned with a **gap**, represents an **indel** (insertion/deletion).

```

A A C T A C T - C C T A A C A C T - -
- - C T C C T A C C T - - T A C T T T
      |                               |
  
```

the bars show mismatches

10 matches, 2 mismatches, 7 gaps  
 Total number of characters 10+2+7=19  
**Percent identity** = 10/19 = 53%

## Scoring scheme for an alignment – example:

$w(\text{match}) = +2$  or 4X4 nt substitution matrix  
 $w(\text{mismatch}) = -1$  or 4X4 nt substitution matrix  
 $w(\text{gap}) = -3$  or some other heuristic penalty

*Find alignment with maximum score*

# Substitution matrix for amino acids

## BLOSUM: Blocks Substitution Matrix

(used to calculate an alignment score)

- Values in matrix are empirical. Based on a large sample of verified aa pairwise alignments
- $i,j$  element = Log of probability ( $p_{ij}$ ) that amino acid  $i$  mutates into amino acid  $j$  in a homologous sequence normalized by the probability of this match by chance given the frequency of the amino acids ( $q_i, q_j$ ) -> **positive: better than chance, negative: worst than chance**

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	2	5											
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

$$S_{ij} = \left( \frac{1}{\lambda} \right) \log \left( \frac{p_{ij}}{q_i * q_j} \right)$$

Positive for chemically similar substitutions (more likely than chance)

Common aa have a low score

Rare aa have a high score

Unlikely substitution compared with chance

• We will come back to substitutions matrices later when we talk about phylogeny

# Local alignment vs. global alignment

- **Global alignment** – attempts to align every residue in every sequence

```
P-ELICAN--  
COELICANTH
```

- **Local alignment** – find best subsequence alignment (useful for finding similar exons in two genomes, finding similar functional regions in a protein, etc.)

```
ELICAN  
ELICAN
```

# BLAST- Basic Local Alignment Search Tool

## The Google of biology (*google blast*)

### Basic BLAST

---

Choose a BLAST program to run.

nt vs. nt

[nucleotide blast](#)

Search a **nucleotide** database using a **nucleotide** query  
*Algorithms: blastn, megablast, discontinuous megablast*

aa vs. aa

[protein blast](#)

Search **protein** database using a **protein** query  
*Algorithms: blastp, psi-blast, phi-blast*

tras nt vs. aa database

[blastx](#)

Search **protein** database using a **translated nucleotide** query

aa vs. trans nt database

[tblastn](#)

Search **translated nucleotide** database using a **protein** query

tras nt vs. trans nt database

[tblastx](#)

Search **translated nucleotide** database using a **translated nucleotide** query

### Specialized BLAST

---

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript and genomic libraries](#)
- Constraint Based Protein [Multiple Alignment Tool](#)
- Needleman-Wunsch [Global Sequence Alignment Tool](#)

Global alignment





# BLAST Scores and Statistics

- **Percent identity** is the fraction of identical characters
- **Percent similarity** (for amino acids) is the fraction of identical or chemically similar residues
- **Bit score** – A normalized alignment score ( $S'$ ). The bit score gives an indication of how good the alignment is; **the higher the score, the better the alignment.**
- **E value** – **The E-value is a test statistic that gives an indication of the statistical significance of a given pairwise alignment by comparing it to a model of random sequences.** It's the average number of sequences with this level of similarity (i.e. raw alignment score  $S$ ) or better expected to be in the database by chance. Reflects the size of the database and the scoring system. **Lower is better.** The threshold is usually placed at  $10^{-3}$ .
- **P value** = The probability of finding at least one such sequence in the database by chance =  $1 - e^{-E}$  (The E value is just the  $\lambda$  parameter in a Poisson distribution  $\lambda^n e^{-\lambda}/n! \dots$ )

Further reading: <http://www.ncbi.nlm.nih.gov/blast/tutorial/>

# Main BLAST window

## Basic BLAST

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms: blastp, psi-blast, phi-blast</i>
<i>click</i> <a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> query

## Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript and genomic libraries](#)
- Constraint Based Protein [Multiple Alignment Tool](#)
- Needleman-Wunsch [Global Sequence Alignment Tool](#)

# Main BLAST window

Choose BLAST tool

Paste nt or aa sequence  
(=query) here

Click here to align  
two sequences  
one against the  
other

Choose database  
to BLAST against.  
Two important  
databases:  
**nr and refseq**

The screenshot shows the NCBI BLAST web interface. At the top, there is a navigation bar with 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. Below this is the 'BLAST' logo and the text 'Basic Local Alignment Search Tool'. The main content area is divided into several sections:

- Enter Query Sequence:** A large text input field for pasting a sequence. To its right are 'Clear' and 'Query subrange' (with 'From' and 'To' sub-inputs) buttons.
- Or, upload file:** A 'Browse...' button for uploading a file.
- Job Title:** A text input field for a descriptive title.
- Align two or more sequences:** A checkbox with the label 'Align two or more sequences'.
- Choose Search Set:** A dropdown menu for selecting a database. The 'Non-redundant protein sequences (nr)' option is highlighted. Other options include 'Reference proteins (refseq\_protein)', 'Swissprot protein sequences (swissprot)', 'Patented protein sequences (pat)', 'Protein Data Bank proteins (pdb)', and 'Environmental samples (env\_nr)'. There are also 'Exclude' and 'sample sequences' options.
- Program Selection:** Radio buttons for selecting an algorithm: 'blastp (protein-protein BLAST)', 'PSI-BLAST (Position-Specific Iterated BLAST)', and 'PHI-BLAST (Pattern Hit Initiated BLAST)'. A 'Choose a BLAST algorithm' link is also present.
- BLAST Button:** A large blue button labeled 'BLAST'.
- Footer:** A summary line: 'Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)' and a checkbox for 'Show results in a new window'.

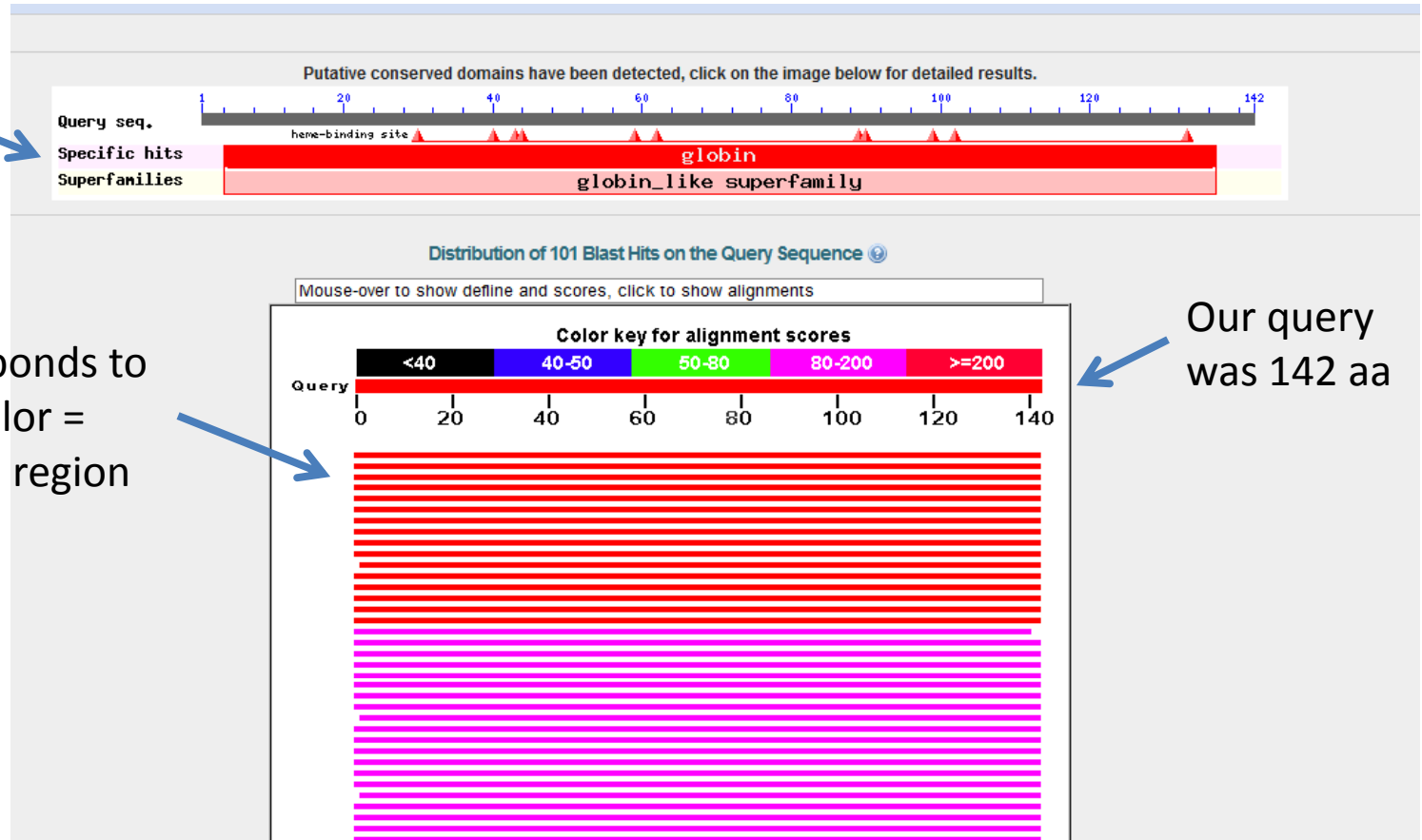
Blue arrows point from the text annotations to the corresponding elements in the interface: 'Choose BLAST tool' points to the 'blastp' tab; 'Paste nt or aa sequence (=query) here' points to the 'Enter Query Sequence' input field; 'Click here to align two sequences one against the other' points to the 'Align two or more sequences' checkbox; and 'Choose database to BLAST against. Two important databases: nr and refseq' points to the 'Non-redundant protein sequences (nr)' option in the 'Choose Search Set' dropdown.

Let's search now for homologs of human HBA1 in the refseq\_protein database

The screenshot shows the NCBI BLAST web interface. The main navigation bar includes 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The 'blastp' tab is selected. The 'Enter Query Sequence' section contains a text box with the following amino acid sequence:   
MVLSPADKTNVKAAGKVGSAHAGEYGAELERMFLSEFTTKTYFRHFDLSHGSAQVKGHGKKVA  
DRLTNAVAVHDDMPNALSALSDDLHAHKLRVDFVNEKLLSHCLLVTLAAHLPAETTPAVHASLDK  
FASVSTVLTISKYR  
Below the text box is a 'Browse...' button. The 'Choose Search Set' section has a dropdown menu for 'Database' set to 'Reference proteins (refseq\_protein)'. The 'Program Selection' section has 'blastp (protein-protein BLAST)' selected. At the bottom, a 'BLAST' button is highlighted with a red box and the word 'click' written below it. The search parameters at the bottom of the page are: 'Search database Reference proteins (refseq\_protein) using Blastp (protein-protein BLAST)'.

# Results

Conserved domains in the protein



# Example of local alignment result

Link to gene record

Bit score

E value

```
> ref|NP\_001013875.1| UG globin, alpha [Rattus norvegicus]
Length=142
```

```
GENE ID: 287167 LOC287167 | globin, alpha [Rattus norvegicus]
(10 or fewer PubMed links)
```

Our seq.

```
Score = 223 bits (568), Expect = 4e-57, Method: Compositional matrix adjust.
Identities = 106/142 (74%), Positives = 120/142 (84%) Gaps = 0/142 (0%)
```

Database seq.

```
Query 1 MVLSPADKTNVKAANGKVGAGHAGEYGAELERMFLSFPITTKTYFPHFDLSHGSAQVKGHG 60
Sbjct 1 MVLS DK N+K AW K+G HA E GAE + R+P+ FP++KTYFPHF+ S GS QVK HG 60

Query 61 KKVADALTNAVAHVDDMPNALSALSSDLHAHKL RVD P V N F K L L S H C L L V T L A A H L P A E F T P 120
Sbjct 61 KKVADALTNA +H+DD+P ALS LSDLHANKLRVDPVNEK LSHCLLVTLA+H P +FTP 120

Query 121 AVHASLDKFFLASVSTVLTISKYR 142
Sbjct 121 AMHASLDKFFLASVSTVLTISKYR 142
```

+ indicates similar aa

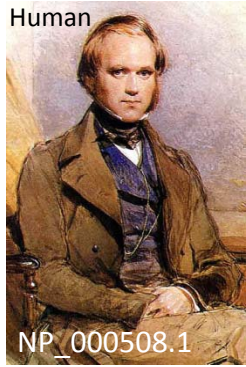
mismatch

106 identities + 14 "+" = 120 positive hits

# Results

Here's our sequence  
NP\_000549.1

Accession	Description	Max score	Total score	Query coverage	E value	Links
NP_000549.1	hemoglobin subunit alpha (Homo sapiens) > ref NP_000549.1  hemoglobin subunit alpha (Homo sapiens) > ref NP_001036092.1  hemoglobin subunit alpha (Pan troglodytes) > ref NP_001036091.1	286	286	100%	3e-78	<a href="#">U</a> <a href="#">G</a>
XP_001055934.1	PREDICTED: similar to alpha 2 globin (Macaca mulatta)	275	276	100%	3e-73	<a href="#">U</a> <a href="#">G</a>
NP_001038189.1	theta 1 globin (Macaca mulatta)	276	276	100%	5e-73	<a href="#">U</a> <a href="#">G</a>
NP_001162287.1	hemoglobin, alpha 1 (Psolo anulis)	262	262	100%	9e-69	<a href="#">U</a> <a href="#">G</a>
NP_001072624.1	hemoglobin alpha, adult chain 2 (Mus musculus) > ref NP_032244.2  hemoglobin subunit alpha (Mus musculus)	255	255	100%	1e-66	<a href="#">U</a> <a href="#">G</a>
NP_001070890.2	hemoglobin subunit alpha (Bos taurus) > ref XP_001788728.1  PREDICTED: similar to hemoglobin alpha chain (Bos taurus)	252	252	100%	8e-66	<a href="#">U</a> <a href="#">G</a>
NP_001078901.1	hemoglobin subunit alpha (Equus caballus)	251	251	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001078900.1	alpha 2 globin (Equus caballus)	250	250	100%	2e-65	<a href="#">U</a> <a href="#">G</a>
NP_001078899.1	hemoglobin subunit alpha-1/2 (Oryctolagus cuniculus)	249	249	100%	7e-65	<a href="#">U</a> <a href="#">G</a>
NP_001013874.1	globin, alpha (Rattus norvegicus)	233	233	100%	4e-57	<a href="#">U</a> <a href="#">G</a>
XP_001517140.1	PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha chain) (Alpha-globin) (Ornithomynchus anatinus)	216	216	99%	5e-55	<a href="#">U</a> <a href="#">G</a>
NP_001007723.1	hemoglobin alpha 2 chain (Rattus norvegicus)	211	211	100%	2e-53	<a href="#">U</a> <a href="#">G</a>
NP_001028158.1	hemoglobin subunit alpha (Monodelphis domestica)	210	210	100%	2e-53	<a href="#">U</a> <a href="#">G</a>
NP_037228.1	hemoglobin alpha, adult chain 2 (Rattus norvegicus)	209	209	100%	8e-53	<a href="#">U</a> <a href="#">G</a>
NP_001004374.1	hemoglobin subunit alpha-A (Gallus gallus)	207	207	100%	2e-52	<a href="#">U</a> <a href="#">G</a>
XP_002186132.1	PREDICTED: putative hemoglobin alpha (Taeniopygia guttata)	206	206	100%	5e-52	<a href="#">U</a> <a href="#">G</a>
XP_356933.3	PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha chain) (Alpha-globin) (Mus musculus) > ref XP_800048.2  PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha	196	196	98%	7e-49	<a href="#">U</a> <a href="#">G</a>
XP_002186164.1	PREDICTED: hemoglobin, zeta (Taeniopygia guttata)	191	191	100%	2e-47	<a href="#">U</a> <a href="#">G</a>
NP_001079746.1	hemoglobin, alpha 2 (Xenopus laevis)	185	185	100%	9e-46	<a href="#">U</a> <a href="#">G</a>
XP_00274712.1	PREDICTED: mCG1039741-like (Rattus norvegicus) > ref XP_002727804.1  PREDICTED: mCG1039741-like (Rattus norvegicus)	184	184	100%	2e-45	<a href="#">U</a> <a href="#">G</a>
NP_001028153.1	hemoglobin, theta T1 (Mus musculus)	184	184	100%	2e-45	<a href="#">U</a> <a href="#">G</a>
NP_001162288.1	hemoglobin, theta 1 (Psolo anulis)	184	184	100%	2e-45	<a href="#">U</a> <a href="#">G</a>
NP_001164888.1	hemoglobin subunit zeta (Oryctolagus cuniculus)	184	184	100%	3e-45	<a href="#">U</a> <a href="#">G</a>
XP_001510398.1	PREDICTED: similar to zeta-1-globin (Ornithomynchus anatinus)	182	182	100%	6e-45	<a href="#">U</a> <a href="#">G</a>
NP_001004373.1	hemoglobin subunit alpha-D (Gallus gallus)	182	182	99%	8e-45	<a href="#">U</a> <a href="#">G</a>
NP_001164886.1	zeta globin (Oryctolagus cuniculus)	182	182	100%	8e-45	<a href="#">U</a> <a href="#">G</a>
NP_001004374.1	hemoglobin subunit pi (Gallus gallus)	182	182	100%	8e-45	<a href="#">U</a> <a href="#">G</a>
NP_005322.1	hemoglobin subunit theta-1 (Homo sapiens)	182	182	100%	1e-44	<a href="#">U</a> <a href="#">G</a>
XP_001915677.1	PREDICTED: similar to Hemoglobin subunit theta-1 (Hemoglobin theta-1 chain) (Theta-1-globin) (Equus caballus)	181	181	100%	1e-44	<a href="#">U</a> <a href="#">G</a>
NP_001164973.1	hemoglobin, zeta (Oryctolagus cuniculus)	181	181	100%	2e-44	<a href="#">U</a> <a href="#">G</a>
NP_001079797.1	hemoglobin alpha 3 subunit (Xenopus laevis)	178	178	100%	1e-43	<a href="#">U</a> <a href="#">G</a>
XP_002186147.1	PREDICTED: putative hemoglobin alpha-D chain (Taeniopygia guttata)	177	177	99%	2e-43	<a href="#">U</a> <a href="#">G</a>
NP_001079749.1	hemoglobin subunit alpha-4 (Xenopus laevis)	177	177	100%	2e-43	<a href="#">U</a> <a href="#">G</a>
NP_001164887.1	theta 1 globin (Oryctolagus cuniculus) > ref NP_001164972.1  hemoglobin subunit theta-1 (Oryctolagus cuniculus)	176	176	100%	4e-43	<a href="#">U</a> <a href="#">G</a>
NP_001135724.1	hemoglobin, alpha 2 (Xenopus (Silurana) tropicalis)	176	176	100%	5e-43	<a href="#">U</a> <a href="#">G</a>
NP_001108014.1	hemoglobin subunit zeta (Equus caballus)	176	176	100%	6e-43	<a href="#">U</a> <a href="#">G</a>
NP_005323.1	hemoglobin subunit zeta (Homo sapiens)	176	176	100%	6e-43	<a href="#">U</a> <a href="#">G</a>
NP_001166316.1	hemoglobin, zeta (Rattus norvegicus) > ref XP_002724711.1  PREDICTED: hemoglobin, zeta (Rattus norvegicus) > ref XP_002727803.1  PREDICTED: hemoglobin, zeta (Rattus norvegicus)	176	176	100%	7e-43	<a href="#">U</a> <a href="#">G</a>
XP_001474223.1	PREDICTED: similar to pol polyprotein (Monodelphis domestica)	176	348	100%	2e-42	<a href="#">U</a> <a href="#">G</a>
NP_026535.1	hemoglobin subunit zeta (Mus musculus)	173	173	100%	4e-42	<a href="#">U</a> <a href="#">G</a>
NP_001135556.1	hypothetical protein LOC100216102 (Xenopus (Silurana) tropicalis)	172	172	100%	8e-42	<a href="#">U</a> <a href="#">G</a>
NP_001015904.1	hemoglobin alpha 3 subunit (Xenopus (Silurana) tropicalis)	172	172	100%	1e-41	<a href="#">U</a> <a href="#">G</a>
XP_001788743.1	PREDICTED: similar to Hemoglobin subunit zeta (Hemoglobin zeta chain) (Zeta-globin) (Bos taurus)	171	171	100%	1e-41	<a href="#">U</a> <a href="#">G</a>
NP_001073802.1	hemoglobin, theta 1 (Bos taurus)	171	171	98%	2e-41	<a href="#">U</a> <a href="#">G</a>
NP_001081493.1	hemoglobin, alpha 1 (Xenopus laevis)	169	169	100%	5e-41	<a href="#">U</a> <a href="#">G</a>
NP_888860.1	hemoglobin, alpha 1 (Xenopus (Silurana) tropicalis)	166	166	100%	4e-40	<a href="#">U</a> <a href="#">G</a>
XP_880707.3	PREDICTED: similar to Hemoglobin subunit zeta (Hemoglobin zeta chain) (Zeta-globin) (Bos taurus)	165	165	98%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001165373.1	alpha globin larval-3 (Xenopus (Silurana) tropicalis)	165	165	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001061675.2	PREDICTED: hemoglobin, theta 1 (Rattus norvegicus) > ref XP_347267.4  PREDICTED: hemoglobin, theta 1 (Rattus norvegicus)	166	164	100%	2e-39	<a href="#">U</a> <a href="#">G</a>
XP_001151282.1	PREDICTED: similar to zeta-1-globin isoform 1 (Pan troglodytes) > ref XP_001151351.1  PREDICTED: similar to zeta-1-globin isoform 2 (Pan troglodytes)	163	163	100%	4e-39	<a href="#">U</a> <a href="#">G</a>
NP_001165374.1	alpha globin larval-4 (Xenopus (Silurana) tropicalis)	163	163	100%	4e-39	<a href="#">U</a> <a href="#">G</a>
NP_001016009.1	alpha globin larval-5 (Xenopus (Silurana) tropicalis)	163	163	100%	4e-39	<a href="#">U</a> <a href="#">G</a>
NP_001002092.1	hemoglobin, zeta (Xenopus (Silurana) tropicalis)	162	162	100%	1e-38	<a href="#">U</a> <a href="#">G</a>
NP_001118023.1	alpha-globin IV (Oncorhynchus mykiss)	158	158	100%	2e-37	<a href="#">U</a> <a href="#">G</a>
NP_001107321.1	alpha globin larval-2 (Xenopus (Silurana) tropicalis)	155	155	100%	1e-36	<a href="#">U</a> <a href="#">G</a>
NP_001082702.1	hypothetical protein LOC734769 (Xenopus laevis)	154	154	100%	3e-36	<a href="#">U</a> <a href="#">G</a>
XP_002723855.1	PREDICTED: zeta globin-like (Oryctolagus cuniculus)	153	153	98%	5e-36	<a href="#">U</a> <a href="#">G</a>
NP_571332.2	hemoglobin alpha adult-1 (Danio rerio)	153	153	100%	5e-36	<a href="#">U</a> <a href="#">G</a>
NP_001013479.1	hypothetical protein LOC497166 (Danio rerio)	152	152	100%	8e-36	<a href="#">U</a> <a href="#">G</a>
NP_001028265.1	alpha globin-like (Danio rerio)	152	152	100%	8e-36	<a href="#">U</a> <a href="#">G</a>
XP_002723856.1	PREDICTED: zeta globin-like (Oryctolagus cuniculus)	151	151	98%	2e-35	<a href="#">U</a> <a href="#">G</a>
NP_778165.1	hemoglobin, theta T2 (Mus musculus)	151	151	88%	2e-35	<a href="#">U</a> <a href="#">G</a>
NP_001117134.1	hemoglobin subunit alpha (Salmo salar)	150	150	100%	3e-35	<a href="#">U</a> <a href="#">G</a>
NP_001085638.1	hemoglobin, zeta (Xenopus laevis)	150	150	100%	4e-35	<a href="#">U</a> <a href="#">G</a>

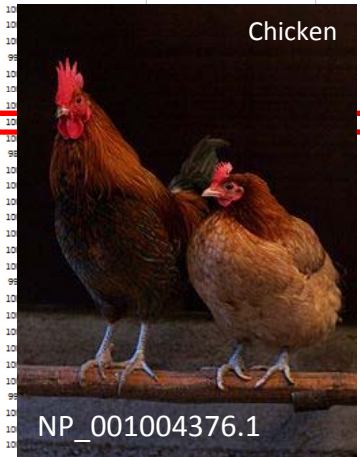
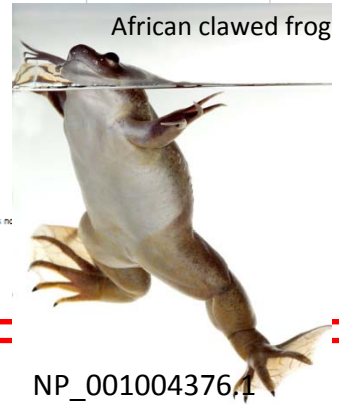


# Results

Here's our sequence  
NP\_000549.1



Accession	Description	Max score	Total score	Query coverage	Evalue	Links
NP_000508.1	hemoglobin subunit alpha (Homo sapiens) > ref NP_000549.1  hemoglobin subunit alpha (Homo sapiens) > ref NP_001036092.1  hemoglobin subunit alpha (Pan troglodytes) > ref NP_001036091.1	286	286	100%	3e-76	<a href="#">U</a> <a href="#">G</a>
XP_001025604.1	PREDICTED: similar to alpha 2 globin (Macaca mulatta)	276	276	100%	3e-73	<a href="#">U</a> <a href="#">G</a>
NP_001038189.1	theta 1 globin (Macaca mulatta)	276	276	100%	3e-73	<a href="#">U</a> <a href="#">G</a>
P_001162287.1	hemoglobin, alpha 1 (Psolo anubis)	262	262	100%	9e-69	<a href="#">U</a> <a href="#">G</a>
NP_001072924.1	hemoglobin alpha, adult chain 2 (Mus musculus) > ref NP_032244.2  hemoglobin subunit alpha (Mus musculus)	253	253	100%	1e-66	<a href="#">U</a> <a href="#">G</a>
NP_001070890.2	hemoglobin subunit alpha (Bos taurus) > ref XP_001788728.1  PREDICTED: similar to hemoglobin alpha chain (Bos taurus)	252	252	100%	8e-66	<a href="#">U</a> <a href="#">G</a>
NP_001078901.1	hemoglobin subunit alpha (Equus caballus)	251	251	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001078900.1	alpha 2 globin (Equus caballus)	250	250	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001078899.1	hemoglobin subunit alpha-1/2 (Oryctolagus cuniculus)	249	249	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001013874.1	globin, alpha (Rattus norvegicus)	233	233	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
XP_001517140.1	PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha chain) (Alpha-globin) (Ornithomychnus anatinus)	216	216	99%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001007723.1	hemoglobin alpha 2 chain (Rattus norvegicus)	211	211	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001028158.1	hemoglobin subunit alpha (Monodelphis domestica)	210	210	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_037228.1	hemoglobin alpha, adult chain 2 (Rattus norvegicus)	209	209	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001004374.1	hemoglobin subunit alpha-A (Gallus gallus)	207	207	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
XP_000195134.1	PREDICTED: putative hemoglobin alpha (Taeniopygia guttata)	206	206	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
XP_356935.3	PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha chain) (Alpha-globin) (Mus musculus) > ref XP_800048.2  PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha	196	196	98%	1e-65	<a href="#">U</a> <a href="#">G</a>
XP_002196164.1	PREDICTED: hemoglobin, zeta (Taeniopygia guttata)	191	191	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001079746.1	hemoglobin, alpha 2 (Xenopus laevis)	185	185	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_002724712.1	PREDICTED: mCG1039741-like (Rattus norvegicus) > ref XP_002727804.1  PREDICTED: mCG1039741-like (Rattus norvegicus)	184	184	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001028153.1	hemoglobin, theta T1 (Mus musculus)	184	184	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001162288.1	hemoglobin, theta 1 (Psolo anubis)	184	184	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001164888.1	hemoglobin subunit zeta (Oryctolagus cuniculus)	184	184	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
XP_001510398.1	PREDICTED: similar to zeta-1-globin (Ornithomychnus anatinus)	182	182	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001004373.1	hemoglobin subunit alpha-D (Gallus gallus)	182	182	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001164889.1	zeta globin (Oryctolagus cuniculus)	182	182	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001004374.1	hemoglobin subunit pi (Gallus gallus)	182	182	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_005332.1	hemoglobin subunit theta-1 (Homo sapiens)	182	182	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
XP_001915677.1	PREDICTED: similar to Hemoglobin subunit theta-1 (Hemoglobin theta-1 chain) (Theta-1-globin) (Bos caballus)	181	181	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001164973.1	hemoglobin, zeta (Oryctolagus cuniculus)	181	181	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001079737.1	hemoglobin alpha 3 subunit (Xenopus laevis)	178	178	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
XP_002196147.1	PREDICTED: putative hemoglobin alpha-D chain (Taeniopygia guttata)	177	177	99%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001079749.1	hemoglobin subunit alpha-4 (Xenopus laevis)	177	177	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001164887.1	theta 1 globin (Oryctolagus cuniculus) > ref NP_001164972.1  hemoglobin subunit theta-1 (Oryctolagus cuniculus)	176	176	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001135724.1	hemoglobin, alpha 2 (Xenopus (Silurana) tropicalis)	176	176	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001108014.1	hemoglobin subunit zeta (Equus caballus)	176	176	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_005333.1	hemoglobin subunit zeta (Homo sapiens)	176	176	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
NP_001166316.1	hemoglobin, zeta (Rattus norvegicus) > ref XP_002724711.1  PREDICTED: hemoglobin, zeta (Rattus norvegicus) > ref XP_002727803.1  PREDICTED: hemoglobin, zeta (Rattus no	176	176	100%	1e-65	<a href="#">U</a> <a href="#">G</a>
XP_001472423.1	PREDICTED: similar to pol polyprotein (Monodelphis domestica)	348	348	100%	2e-42	<a href="#">U</a> <a href="#">G</a>
NP_024535.1	hemoglobin subunit zeta (Mus musculus)	173	173	100%	4e-42	<a href="#">U</a> <a href="#">G</a>
NP_001135556.1	hypothetical protein LOC100216102 (Xenopus (Silurana) tropicalis)	172	172	100%	8e-42	<a href="#">U</a> <a href="#">G</a>
NP_001015904.1	hemoglobin alpha 3 subunit (Xenopus (Silurana) tropicalis)	172	172	100%	1e-41	<a href="#">U</a> <a href="#">G</a>
XP_001788743.1	PREDICTED: similar to Hemoglobin subunit zeta (Hemoglobin zeta chain) (Zeta-globin) (Bos taurus)	171	171	100%	1e-41	<a href="#">U</a> <a href="#">G</a>
NP_001073807.1	hemoglobin, theta 1 (Bos taurus)	171	171	98%	2e-41	<a href="#">U</a> <a href="#">G</a>
NP_001011893.1	hemoglobin, alpha 1 (Xenopus laevis)	169	169	100%	5e-41	<a href="#">U</a> <a href="#">G</a>
NP_388850.1	hemoglobin, alpha 1 (Xenopus (Silurana) tropicalis)	166	166	100%	4e-40	<a href="#">U</a> <a href="#">G</a>
XP_580707.3	PREDICTED: similar to Hemoglobin subunit zeta (Hemoglobin zeta chain)	165	165	98%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001165373.1	alpha globin larval-3 (Xenopus (Silurana) tropicalis)	165	165	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001061675.2	PREDICTED: hemoglobin, theta 1 (Rattus norvegicus) > ref XP_347267.4	164	164	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
XP_001151282.1	PREDICTED: similar to zeta-1-globin isoform 1 (Pan troglodytes) > ref XP_001151282.1	163	163	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001165374.1	alpha globin larval-4 (Xenopus (Silurana) tropicalis)	163	163	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001016009.1	alpha globin larval-5 (Xenopus (Silurana) tropicalis)	163	163	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001005092.1	hemoglobin, zeta (Xenopus (Silurana) tropicalis)	162	162	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001118023.1	alpha-globin IV (Oncorhynchus mykiss)	158	158	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001107321.1	alpha globin larval-2 (Xenopus (Silurana) tropicalis)	155	155	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001082702.1	hypothetical protein LOC734769 (Xenopus laevis)	154	154	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_002723855.1	PREDICTED: zeta globin-like (Oryctolagus cuniculus)	153	153	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_571332.2	hemoglobin alpha adult-1 (Danio rerio)	153	153	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_000101459.2.4	hypothetical protein LOC104911049 (Danio rerio)	152	152	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_001028265.1	alpha globin-like (Danio rerio)	152	152	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
XP_002723856.1	PREDICTED: zeta globin-like (Oryctolagus cuniculus)	151	151	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_778165.1	hemoglobin, theta T2 (Mus musculus)	151	151	100%	1e-39	<a href="#">U</a> <a href="#">G</a>
NP_000117134.1	hemoglobin subunit alpha (Salmo salar)	150	150	100%	3e-35	<a href="#">U</a> <a href="#">G</a>





# Before we align...

- Should we align the amino acid sequences or the nucleotide sequences of a protein coding gene?



# Before we align...

- Should we align the amino acid sequences or the nucleotide sequences of a protein coding gene?

## Amino acid

- Amino acids are more conserved
- Aligning nts can lead to placing gaps inside codons

# Before we align...

- Should we align the amino acid sequences or the nucleotide sequences of a protein coding gene?

## Amino acid

- Amino acids are more conserved
  - Aligning nts can lead to placing gaps inside codons
- Then should we download for alignment the amino acid sequence or the nt sequence?

# Before we align...

- Should we align the amino acid sequences or the nucleotide sequences of a protein coding gene?

## Amino acid

- Amino acids are more conserved
- Aligning nts can lead to placing gaps inside codons
- Then should we download for alignment the amino acid sequence or the nt sequence?

## Nucleotide! **contains more information**

- The nucleotide sequence gives us the ability to
  - Detect silent mutations (codon bias)
  - Measure selection pressure
  - Detect frame shifts (rare but can occur in defunct genes)
- No information is lost (but its always worth comparing hypothetical translation to annotated version if it exists)

# How to obtain nt sequence?

## Sequences producing significant alignments:

Accession	Description
<a href="#">NP_000508.1</a>	hemoglobin, alpha 1 [Papio anubis]
<a href="#">XP_001094404.1</a>	PREDICTED protein
<a href="#">NP_001038189.1</a>	theta-1 globin [Papio anubis]
<a href="#">NP_001162287.1</a>	hemoglobin, alpha 1 [Papio anubis]
<a href="#">NP_001038189.1</a>	theta-1 globin [Papio anubis]
<a href="#">NP_001038189.1</a>	theta-1 globin [Papio anubis]
<a href="#">NP_001038189.1</a>	theta-1 globin [Papio anubis]

click

click

click

click

click

click

click

```
>gi|281183149:38-466 Papio anubis hemoglobin, alpha 1 (HBA1), mRNA
ATGGTGTCTGCTCCTGACGACAAGAAACAGCTCAAGCGCCGCTGGGTAAAGTCCGGCAGCAGCTGGG
AGTATGGTGGGAGCCCTCGGAGAGGATGTTCTGCTCTCCCAACACCAAGACCTACTCCGCCACTT
CGACTGAGCCACGCTCTGACAGCTTAACAACACGCGCAAGAGTGGCCGACGCTGACCCCTGCC
GTGGGACAGTGGACGACATGCCCGGCGCTGTTCAAGCTGAGCGACTGCACCGCCACAGCTTCGGG
TGGACCGGCTCACTTCAAGCTCCTGAGCCACTGCTGCTGCTGACTGTGGCCGCTCACTCCCGCCGA
GTTACCCTCGGTCAGCCTCCCTGGACAAGTTCTGGCTTCTGTGACACGCTGCTGACTCCAAA
TACCGTTAA
```

# Create a FASTA file (must have extension “.fasta”)

HBA1\_nt.fasta

## For MEGA

- Use unique names
- Use ‘\_’ instead of spaces
- Use only letters, numbers and the characters ‘\_’ and ‘.’
- Names should be <10 characters
- Use meaningful names
- Use the file extension ‘.fasta’

```
>gi|14456711:38-466 Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA (NM_000558.3)
ATGGTCTCTCTCTGACGACAAGACCAACGCTCAAGGCCGCTGGGGTAAGGTCGGCGCGCAAGCCGCTG
AGTATGTGCGGAGGCCCTGGAGAGGATGTTCTGCTTCCCAACCAAGACCTCTCCGACCTT
CGACCTGAGCCACGGCTCTGCCAGGTTAAGGGCCACGGCAAGAAGTGGCCGACGGCTGALAAAGCC
GTGGGCGACGTGACGACATGCCCAACGGCTGTCCGCCCTGAGCGACCTGACCGCCCAAGCTCGGG
TGGACCCGGTCAACTTCAAGCTCTCAAGCCACTGCTGCTGGTGACCTGCGCCGCCCACTCCCCGCCA
GTTCAACCCCTGCGGTGACGCTCTCTGGACAAGTCTGCTGCTGTGAGCACCGTGTGACCTCAAA
TACCGTTAA

>gi|281183149:38-466 Papio anubis hemoglobin, alpha 1 (HBA1), mRNA (NM_001168816.1)
ATGGTCTCTCTCTGACGACAAGAAACACGCTCAAGGCCGCTGGGGTAAGGTCGGCGAGCCGCTG
AGTATGTGCGGAGGCCCTGGAGAGGATGTTCTGCTTCCCAACCAAGACCTCTCCGACCTT
CGACCTGAGCCACGGCTCTGCCAGGTTAACAACAACGGCAAGAAGTGGCCGACGGCTGACCCGCTG
GTGGGCGACGTGACGACATGCCCAACGGCTGTCCGCCCTGAGCGACCTGACCGCCCAAGCTCGGG
TGGACCCGGTCAACTTCAAGCTCTGAGCCACTGCTGCTGGTGACTCTGCGCCGCTCACTCCCCGCCA
GTTCAACCCCTGCGGTGACGCTCTCTGGACAAGTCTGCTGCTGTGAGCACCGTGTGACCTCAAA
TACCGTTAA

>gi|52345402:16-444 Gallus gallus hemoglobin, alpha 1 (HBA1), mRNA (NM_001004376.2)
ATGGTCTCTCTCTGACGACAAGAAACACGCTCAAGGCCGCTGGGGTAAGGTCGGCGAGCCGCTG
AGTATGTGCGGAGGCCCTGGAGAGGATGTTCAACCACTCCCAACCAAGACCTCTCCGACCTT
CGACCTGAGCCACGGCTCTGCCAGGTTAACAACAACGGCAAGAAGTGGCCGACGGCTGACCCGCTG
GCAACCAACATGATGACATGCCCGCACCTCTCAAGCTCAGCGACCTCCATGCCCAAGCTCCGCG
TGGACCCGTCAACTTCAAACTCTGGGCCAATGCTCTGCTGGTGTTGGTGGCCATCCACCACTGCTG
CTGACCCCGGAGGTCCATGCTCTCTGGACAAGTCTGCTGCGCCGTTGGGCACTGTGTCAGCCGCAAG
TACCGTTAA

>gi|147902602:32-460 Xenopus laevis hemoglobin, alpha 1 (hba1), mRNA (NM_001088024.1)
ATGCTTCTTTCAGCTGATGACAAGAAACACATCAAGGCAATTATGCTTCCATAGCCGCTCATGGCAG
AATTTGGTGGAGAGCTTTGTACAGGATGTTCTGGTTAACCTAAGACCAAAACCTACTTCTAGTTT
TGACTTCCACCAATTCAAAACAGATCACTTCTATGGCAAGAAAGTGTGATGCTGATGAATGAAGCT
GCCAACCAATTTGGATAACATTTGCTGGAAAGCATGAGCAAGCTGAGCGACCTCCATGCTATGACCTGAGAG
TGGATCCGGGCAACTTCCCAATGCTGGCTCATAAATTTGCTGGTGGTTGTGCTATGCACTTCCCAAGCA
GTTTGATCTGCAACCCATAAGGCCCTGGATAAGTCTGCTGGCTACCGTATCTACTGTTCTGACTTCCAAA
TATCGTTAA

>gi|185132477:52-483 Salmo salar hemoglobin subunit alpha 1 (hba), mRNA (NM_001123662.1)
ATGAGTCTGACAGCAAGGGACAATCTGTGGTCAATGCTTCTGGGGCAAGATTAAAGGAAAGGCAAT
TCGTGCGGCTGAGGCTTTGGGAAGGATGCTGACTGCTTACCCCGAGCTAAGACCTACTTCCCACTG
GGCTGACCTGAGCCCGGCTCTGCCCACTGAAAGAAGCATGGAGCGCTCATCATGGGTGCAATGGTAA;
GCTGTCGACTGATGGACGACCTCTGGGGGGAATGAGTGTCTCAGCGATCTGACGCGCTTCAAGCTGC
GGTGTGACCTGGAAACTTCAAGATTCTGCCCAACATCCTTGTCAACCTGGCTTACTTCCCTGCG
GGATTTCACTCCCGAAGTGACATTTGCTGTGGATAAATCTTGCAGCCTTGTCCGCTGCCCTGGCTGAC
AAATACAGATAA

>gi|47271416:49-480 Danio rerio hemoglobin alpha adult-1 (hbaa1), mRNA (NM_131257.2)
ATGAGTCTCTGATACGACAAGGCTGTTGTTAAGGCCACTGGGCTAAGATCAGCCCCAAGGCCGAT
AAATGGTGCTGAAGCCCTCGCCAGAAATGCTGACCGCTCAACCTCAGACCAAGACCTAATTTCTCTCA
TGGTGAAGTGAAGCCCTGGGCTGCTGCCGTGAAGAAGCACGGAAGACTATCATGGGTGCGCTGGCGA
GCTGTTCAAAAATAGACGACCTTGTGGGAGGACTGGCCCGCTGAGCGAACTCCATGCTTCAAGCTGC
GTGTTGACCCGGCCAACTTCAAGATCTGTGACACAATGTCATTGGTCAATGCCATGCTCTTCCCTGC
AGACTTCAACCCCTGAGGTTACAGTGTGACGACAAGTCTTAACTTAACTTGGCCCTGGCTCTCTGAG
AAGTACCGCTAA
```

# Bioinformatics

## Bi1X-2010

**Part III:** Multiple alignment and  
phylogenetic analysis

Arbel Tadmor

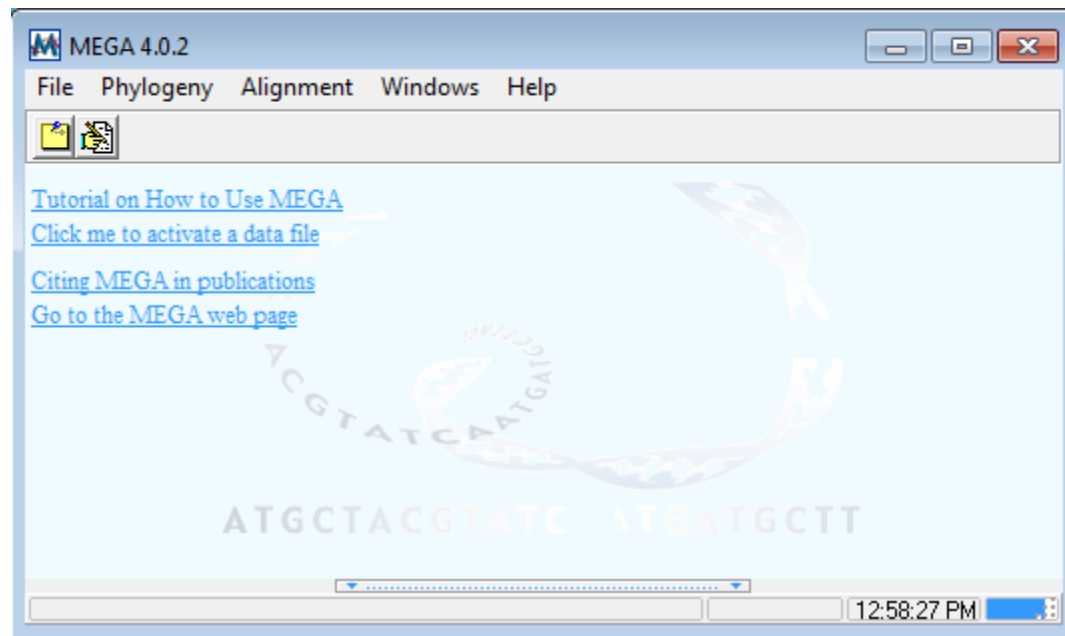
# Overview

- Multiple alignment (HBA1 homologs)
- Phylogenetic trees
- Measuring evolutionary distance
- Building a neighbor joining tree with MEGA (HBA1 example)
- The case of rRNA sequences



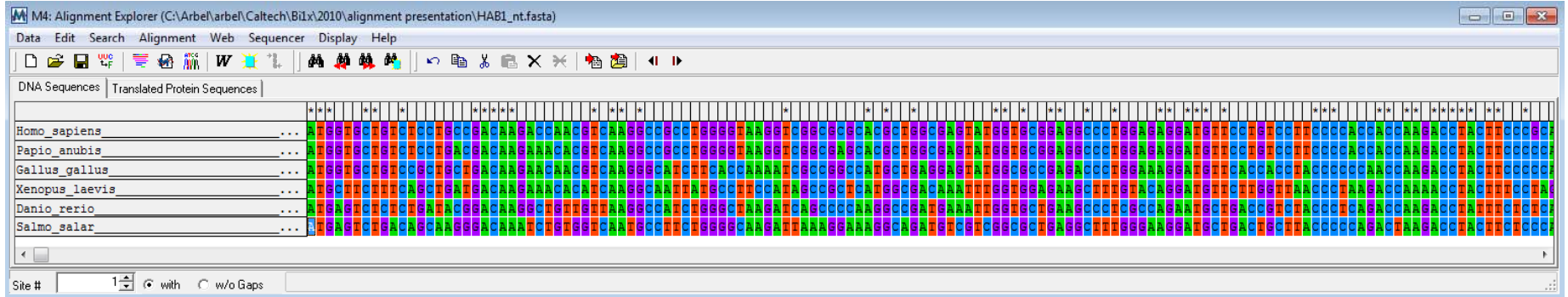
# Multiple Alignment in MEGA4

- Google MEGA4 and install
- Launch MEGA4
- Drag the FASTA file into MEGA4



# Multiple Alignment in MEGA4

nt view



aa view



# ClustalW:

## Popular multiple alignment algorithm

Algorithm overview:

- Global alignment on all sequence pairs to find the **distance** between all pairs of sequences
- Uses distances to create a **guide tree**
- Align the closest sequences in the guide tree, followed by adding more sequences to the initial alignment

Read more at <http://www.ebi.ac.uk/2can/tutorials/nucleotide/clustalw.html>

# What can we do with a multiple alignment?

- Identify conserved regions within protein
  - Signifies conserved function
  - Useful for primer design
- Identify variable regions within protein
  - Functionally not important
  - Important but under positive selection pressure or rapidly changing
  - Identify non-silent mutations (e.g. leading to disease, due to adaptation, etc.)
- Construct a phylogenetic tree (discuss later)

# Alignment using ClustalW

*Hint: be sure to align in the protein pane*

The screenshot shows the M4: Alignment Explorer interface. The main window displays a sequence alignment of several protein sequences. A context menu is open over the alignment, with the 'Align by ClustalW' option selected. A red arrow points from this menu to the 'M4: ClustalW Parameters' dialog box. The dialog box is set to the 'Protein' tab and shows the following parameters:

- Pairwise Alignment:
  - Gap Opening Penalty: 10
  - Gap Extension Penalty: 0.1
- Multiple Alignment:
  - Gap Opening Penalty: 3 (highlighted with a red box)
  - Gap Extension Penalty: 1.8 (highlighted with a red box)
- Protein Weight Matrix: Gonnet
- Residue-specific Penalties: ON
- Hydrophilic Penalties: ON
- Gap Separation Distance: 4
- End Gap Separation: OFF
- Use Negative Matrix: OFF
- Delay Divergent Cutoff (%): 30
- Keep Predefined Gaps

At the bottom of the dialog box are buttons for '? Help', 'OK', and 'Cancel'. A red arrow points from the text 'Change default setting to those best suited for protein alignments' to the '3' and '1.8' values in the Multiple Alignment section.

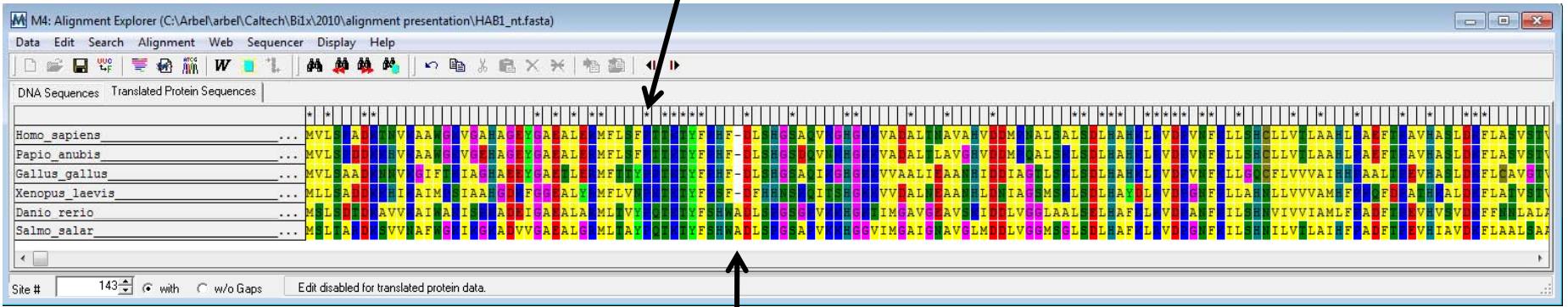
# Manual inspection of alignment

- Pay attention to the edges
- Are there any obviously wrong sequences that did not align well?
  - Sequences too divergent? (must have  $\geq 20\%$  aa identity)
  - Reverse complement?
  - Frame shift?

# Alignment using ClustalW

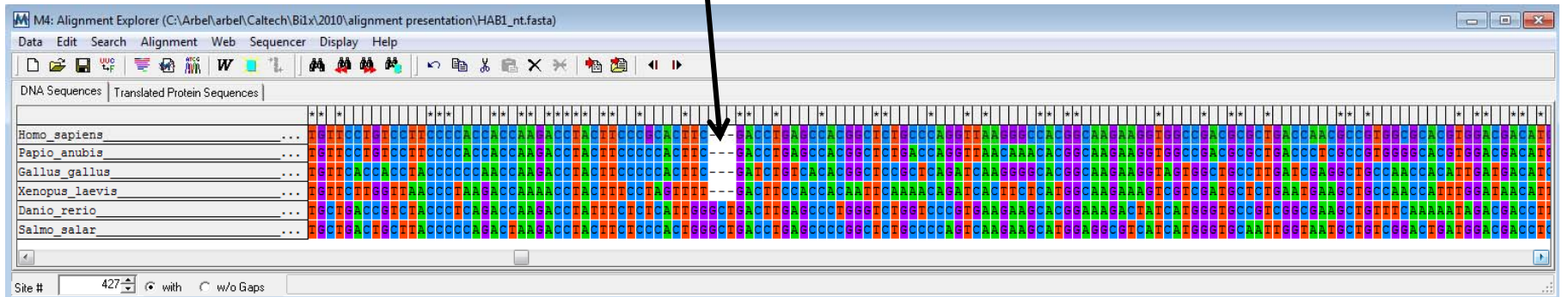
aa view

Star indicates conserved character

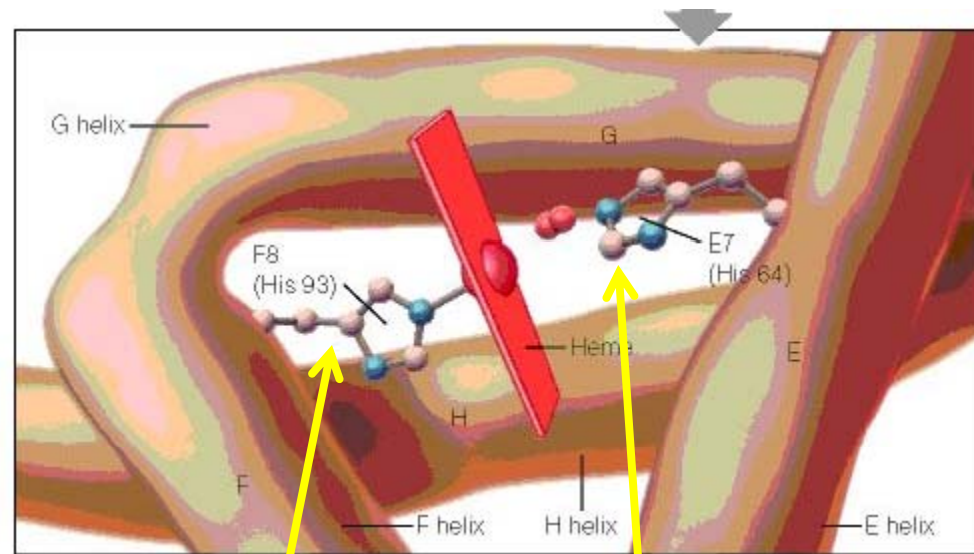
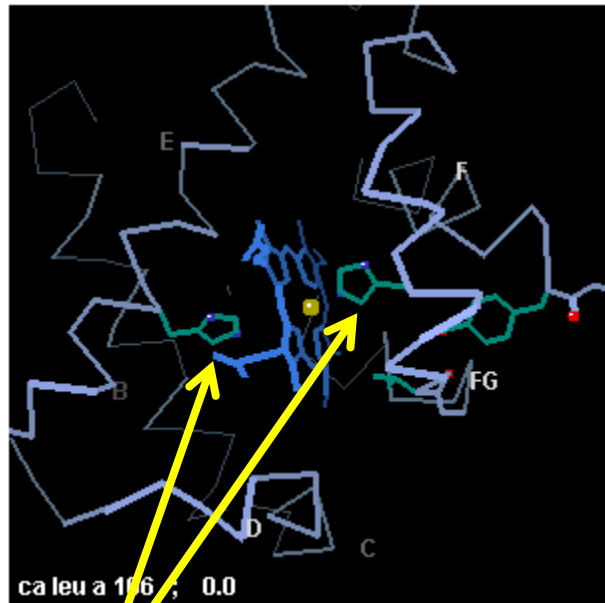


nt view

Deletion/insertion event



# Example of residues conserved due to function: His87 and His58 maintain the heme and oxidation state of the iron



Mathews et al. 2000. Biochemistry 3rd edition

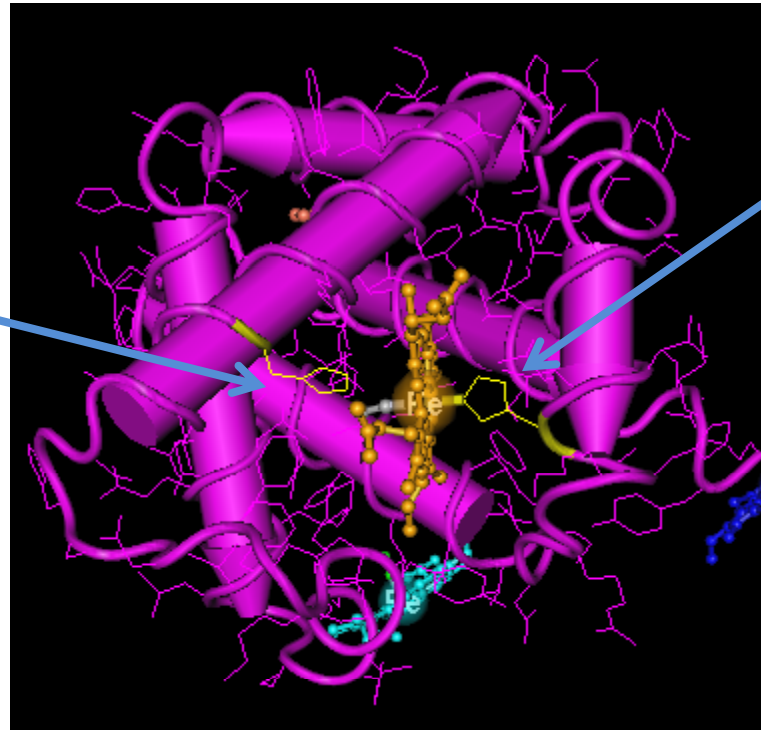
**Proximal His:** Anchoring of the heme is facilitated by a nitrogen from a histidine that binds to the iron.

**Distal His:** The bound oxygen can be in two states, dioxygen (bound to  $\text{Fe}^{2+}$ ) and superoxide (bound to  $\text{Fe}^{3+}$ ). Oxygen must be released in the former because the latter is both harmful and leaves the iron in a state that cannot bind oxygen. The distal histidine binds more strongly to superoxide and the oxygen is therefore less likely to be released.



# Cn3D demo of Homo sapiens hemoglobin subunit $\alpha$

His58  
Distal His



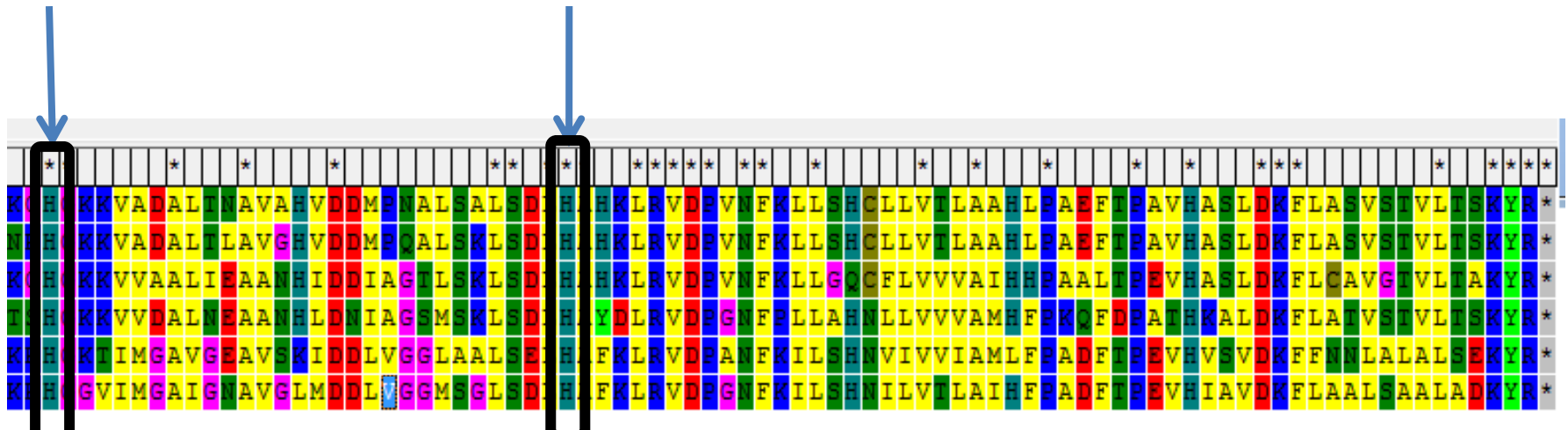
His87  
Proximal His

```
3D17 - Sequence/Alignment Viewer
View Edit Mouse Mode Unaligned Justification Imports
3D17_A phfdlshgsaqvkggkkvadaltnavahvddmpnalsalsdlhahklrvdpvnfkllshcllvtlaahlpaeftpavhasl
3D17_B sfgdlstpdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlseihcdklhvdpenfrllgnvlvcvlahhfgkeftpp
3D17_C phfdlshgsaqvkggkkvadaltnavahvddmpnalsalsdlhahklrvdpvnfkllshcllvtlaahlpaeftpavhasl
3D17_D sfgdlstpdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlseihcdklhvdpenfrllgnvlvcvlahhfgkeftpp
```

# Histidines are conserved

Conserved His58

Conserved His87



*	*	
C	F	C
C	F	C
C	F	C
C	F	C
C	F	C
C	F	C

*	*	
C	F	C
C	F	C
C	F	H
C	F	H
C	F	H
C	F	H
C	F	C

Selection pressure is on the amino acid sequence

The Genetic Code

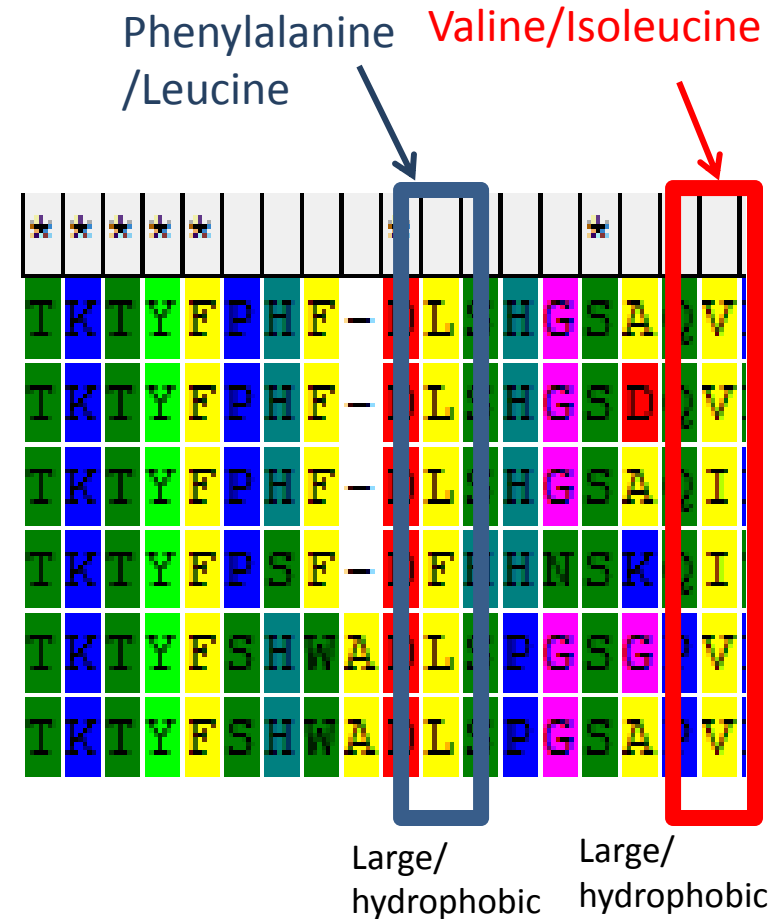
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	Hs	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
5' GCA	CGA	AAC	GAC	UGC	CAA	GAA	GGA	CAC	AUA	CUA	AAA	AUG	UUC	CCA	UCA	ACA	UGG	UAC	GUA
	C	U	U	U	G	G	C	U	C	C	A	U	U	C	U	C	G	U	C
	G	U	U	U	U	G	U	U	U	G	G	U	U	U	U	U	U	U	G
		U													U	U			U
		or												U		U			
		AGA														AGC			
		G													U				

# Some residues mutate to chemically similar residues

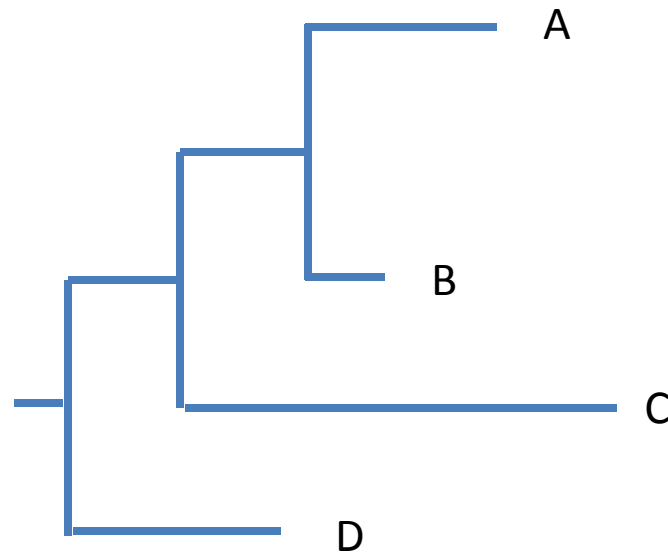
nonpolar polar basic acidic (stop codon)

The table shows the 64 codons and the amino acid for each. The direction of the mRNA is 5' to 3'.

		2nd base			
		U	C	A	G
U	UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine	
	UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine	
	UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (Stop)	UGA Opal (Stop)	
	UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (Stop)	UGG (Trp/W) Tryptophan	
	C	CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine
		CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine
CUA (Leu/L) Leucine		CCA (Pro/P) Proline	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine	
A	CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine	
	AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine	
	AUC (Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine	
G	AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine	
	AUG <sup>(M)</sup> (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine	
	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine	
G	GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine	
	GUA (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine	
	GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine	



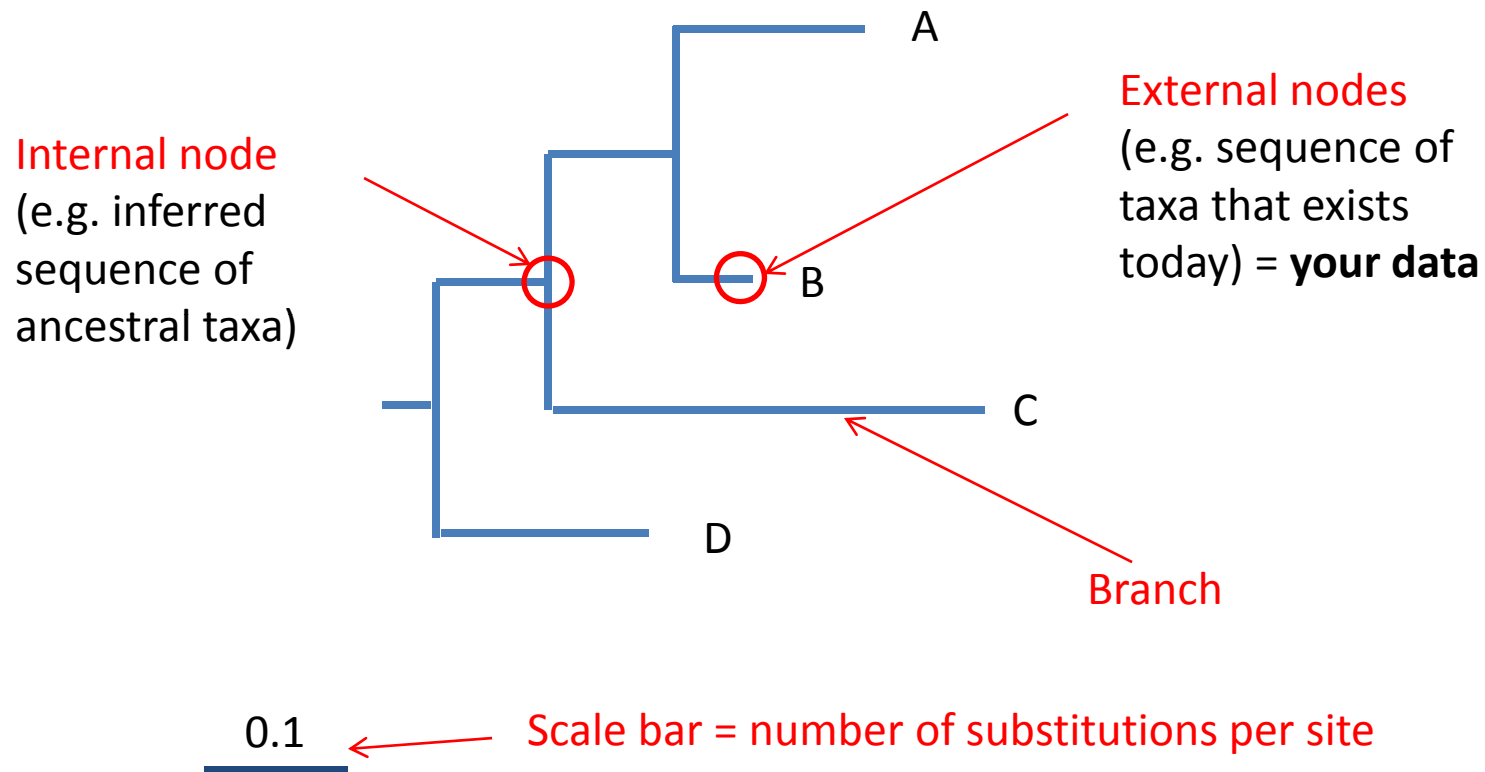
# Phylogenetic analysis



## What are trees good for?

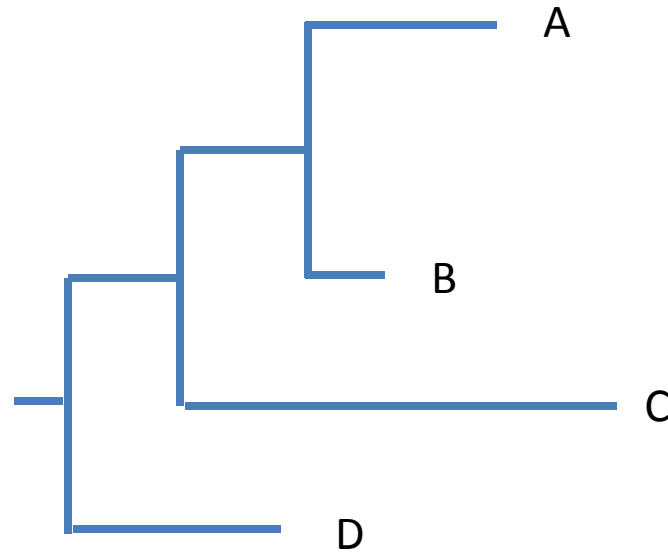
- Identify close relatives
- Determine evolutionary relationship between sequences

# Some tree terminology



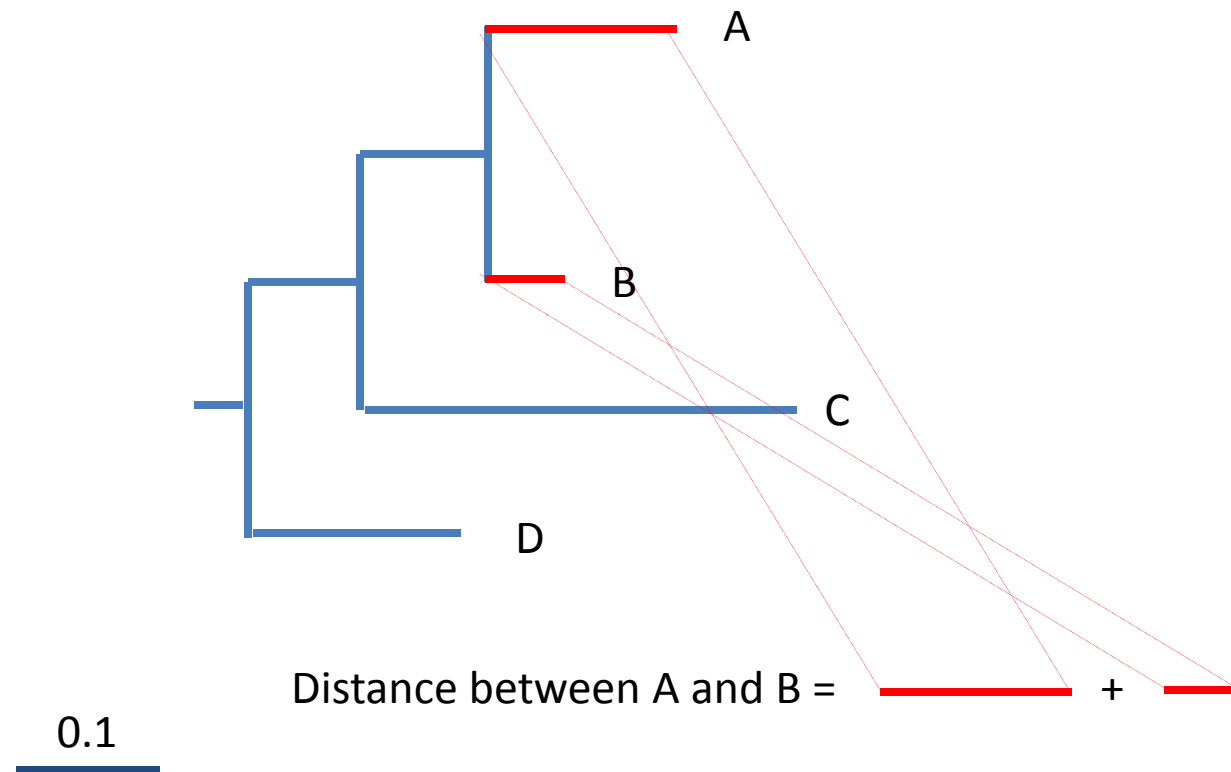
We will discuss only **bifurcating** trees: each node has only two immediate descendant lineages, i.e. we assume evolutionary speciation is a binary process

# Trees have two elements: branch lengths and branching order



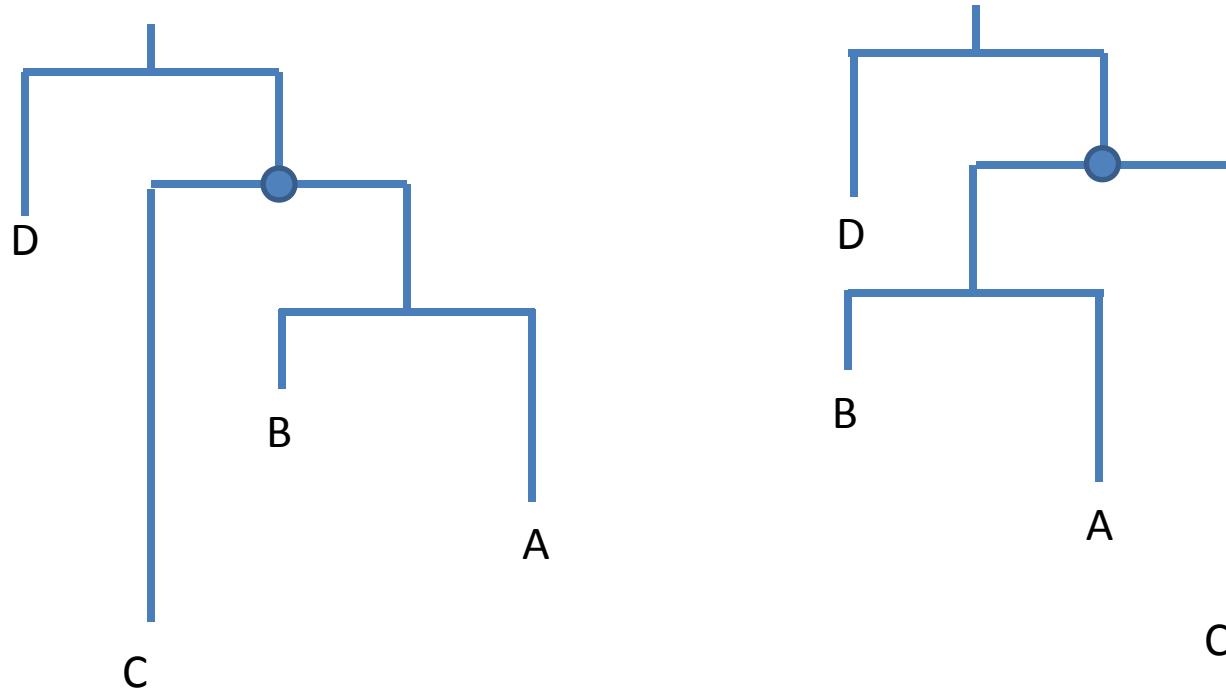
- **Topology** = branching order
- To estimate a tree you need to estimate
  - Branch lengths (simple problem)
  - Branching order (difficult for many seq)

# Branch lengths



- **Evolutionary distance** = accumulated horizontal distance between two external nodes = estimated number of substitution per site that differ between the two sequences = *d*

Branch order:  
note that trees are like mobiles...

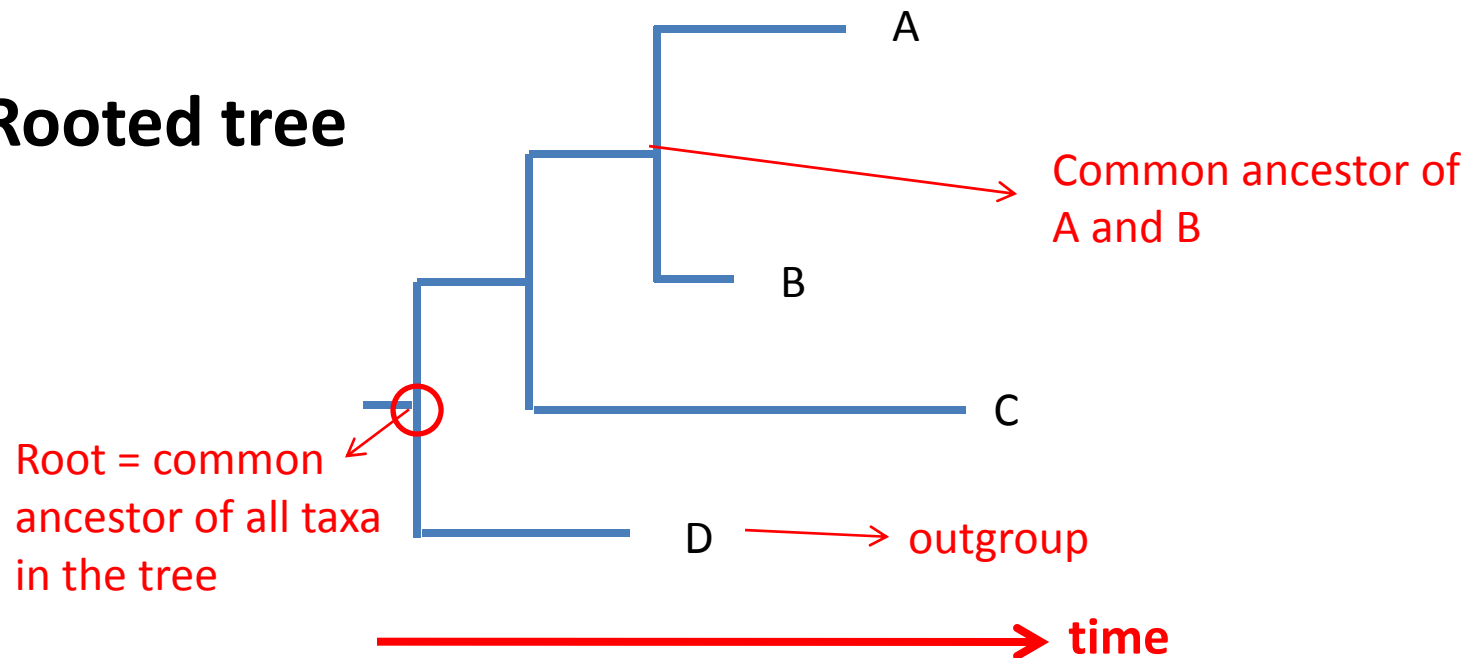


These two trees are equivalent in every way



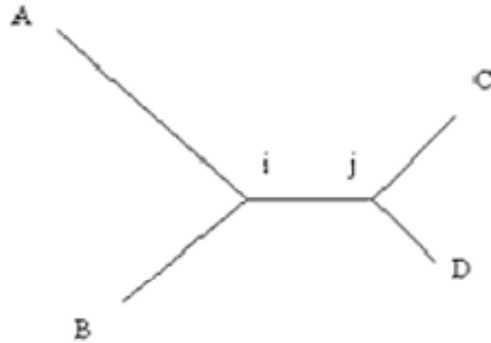
# Rooted trees

## Rooted tree

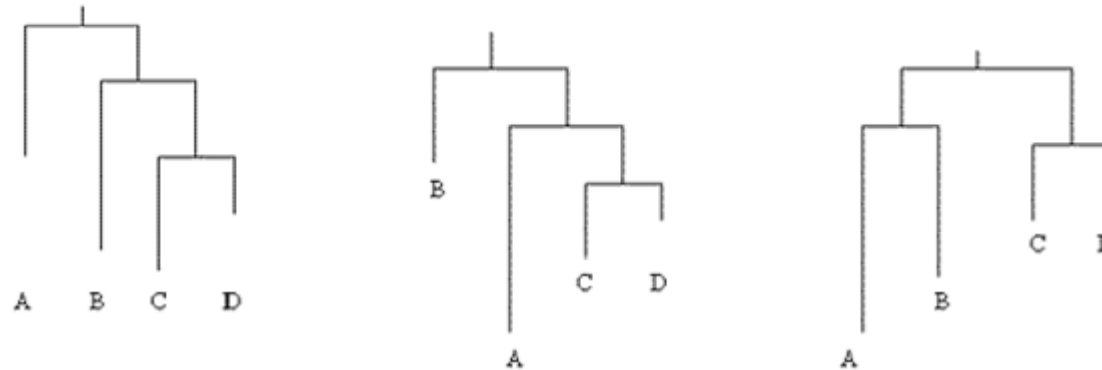


- **If** a tree is rooted, root = left most internal node
- By selecting a node to be a **root** you set a time arrow
- **The more recently species share a common ancestor, the more closely related they are** (e.g. A and C are more closely related than A and D)
- Which node is the root? You “break the symmetry” by adding additional information: e.g. you know D is more distantly related to the ingroup sequences than the ingroup sequences are related to each other

# Unrooted trees



Three ways this tree can be rooted:



- If it is not explicitly said that the tree is rooted assume it is unrooted
- unrooted trees do not specify an evolutionary pathway (who descended from whom) only relationships among taxa

# How to construct a tree?

Algorithmic methods (distance based)	Tree-searching methods (character based)
Use algorithm to construct a single tree from the data	Construct many trees then use some criterion to decide which is the best tree
<ol style="list-style-type: none"><li>1. Multiple alignment</li><li>2. Calculate <b>distance matrix</b> = matrix of evolutionary distances between all pairs of aligned sequences</li><li>3. Calculate tree topology</li></ol>	<ol style="list-style-type: none"><li>1. Multiple alignment</li><li>2. Compare characters at each column in the alignment and give each topology a score</li><li>3. Choose the topology with the best score</li></ol>
Example: Neighbor Joining (many others)	Examples: <ul style="list-style-type: none"><li>• Maximum Likelihood methods</li><li>• Maximum Parsimony methods</li><li>• Bayesian methods</li></ul>

# Measuring evolutionary distance between two sequences

- The evolutionary distance between two sequences  $d$  is the total number of number of aa/nt substitutions per site between the two sequences
- **$d=2rt$**   
 *$t$ =time in years,  $r$ = substitution rate per year per site*
- Branch length in tree =  $d$
- How can we estimate  $d$  from the sequences?

# Measuring evolutionary distance between two sequences

- **p distance:**  $p = n_d/n$  = number of different aa/nt between two aligned sequences of length  $n$ 
  - Doesn't account for multiple hits:     A → C → T
  - Doesn't account for back mutations:    A → C → A
  - Doesn't account for parallel mutations: A → C; A → C
  - Underestimates  $d$
  - Saturates at  $p=0.75$  (not a good estimate of  $d$  when  $p$  is high)
  - Can result in wrong topology
- **Estimation of  $d$  based on a stochastic model:** aa/nt substitutions are modeled as a stochastic process.
  - Different stochastic models make different assumptions regarding the probability of aa/nt substitutions
  - Different models assume different **substitution matrices**

# Calculating distances in MEGA4: nt p distance

The screenshot shows the MEGA 4.0.2 interface. The main window displays a DNA sequence alignment for five species: Homo sapiens (human), Papio anubis (olive baboon), Gallus gallus (chicken), Xenopus laevis (frog), Danio rerio (zebra fish), and Salmo salar (atlantic salmon). The alignment is shown in a grid format with columns representing individual nucleotide positions. The 'Distances' menu is open, and the 'Compute Pairwise...' option is selected. A red arrow points from this menu item to the 'Analysis' dialog box. In the dialog box, the 'Substitution Model' section is expanded, and 'Nucleotide: p-distance' is selected and highlighted with a red box. The 'Compute' button is visible at the bottom of the dialog.

Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Homo_sapiens_(human)	T	C	A	A	C	T	T	C	A	A	G	C	T	C	C	T	A	A		
Papio_anubis_(olive_baboon)	T	C	A	A	C	T	T	C	A	A	G	C	T	C	C	T	G	A		
Gallus_gallus_(chicken)	T	C	A	A	C	T	T	C	A	A	A	C	T	C	C	T	G	G		
Xenopus_laevis_(frog)	G	C	A	A	C	T	T	C	C	A	T	T	G	C	T	G	G			
Danio_rerio_(zebra_fish)	C	C	A	A	C	T	T	C	A	A	G	A	T	C	C	T	G	T		
Salmo_salar_(atlantic_salmon)	G	A	A	A	C	T	T	C	A	A	G	A	T	T	C	T	G	T		

Analysis

Analysis	Pairwise distance calculation
->Compute	Distances only <input checked="" type="checkbox"/>
Include Sites	<input type="checkbox"/>
->Gaps/Missing Data	Complete Deletion <input checked="" type="checkbox"/>
->Codon Positions	1st+2nd+3rd+Noncoding <input checked="" type="checkbox"/>
Substitution Model	<b>Nucleotide: p-distance</b> <input type="checkbox"/>
->Model	<input type="checkbox"/>
->Substitutions to Include	4-Transitions+Transversions <input checked="" type="checkbox"/>
->Pattern among Lineages	Same (Homogeneous) <input type="checkbox"/>
->Rates among sites	Uniform rates <input type="checkbox"/>

Compute Cancel Help

# Calculating distances in MEGA4: **p distance**

p distance =  $100 - \text{percent identity}/100$

	1	2	3	4	5	6
1. Homo sapiens (human) ...						
2. Papio anubis (olive baboon) ...	0.0513					
3. Gallus gallus (chicken) ...	0.2587	0.2681				
4. Xenopus laevis (frog) ...	0.4056	0.3869	0.3660			
5. Danio rerio (zebra fish) ...	0.3846	0.4009	0.3776	0.4336		
6. Salmo salar (atlantic salmon) ...	0.3963	0.3963	0.3986	0.4522	0.2960	

Note that p distances < 0.75

# Calculating distances in MEGA4: JC correction

The image shows the MEGA 4.0.2 interface. On the left, a DNA alignment explorer window displays a sequence alignment for five species: Homo sapiens (human), Papio anubis (olive baboon), Gallus gallus (chicken), Xenopus laevis (frog), Danio rerio (zebra fish), and Salmo salar (atlantic salmon). The alignment is color-coded by nucleotide (A, C, G, T). On the right, the 'Distances' menu is open, and 'Compute Pairwise...' is selected. A red arrow points from this menu item to the 'Compute Pairwise...' dialog box. In this dialog, the 'Substitution Model' is set to 'Nucleotide: Jukes-Dantor', which is highlighted with a red box. Other settings include 'Data type: Nucleotide (Loading)', 'Analysis: Pairwise distance calculation', and 'Rates among sites: Uniform rates'. The 'Compute' button is visible at the bottom of the dialog.

Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Homo_sapiens_(human)	T	C	A	A	C	T	T	C	A	A	G	C	T	C	C	T	A	A		
Papio_anubis_(olive_baboon)	T	C	A	A	C	T	T	C	A	A	G	C	T	C	C	T	G	A		
Gallus_gallus_(chicken)	T	C	A	A	C	T	T	C	A	A	A	C	T	C	C	T	G	G		
Xenopus_laevis_(frog)	G	C	A	A	C	T	T	C	C	A	T	T	G	C	T	G	G			
Danio_rerio_(zebra_fish)	C	C	A	A	C	T	T	C	A	A	G	A	T	C	C	T	G	T		
Salmo_salar_(atlantic_salmon)	G	A	A	A	C	T	T	C	A	A	G	A	T	T	C	T	G	T		



# Calculating distances in MEGA4: Evolutionary distance

## JC method

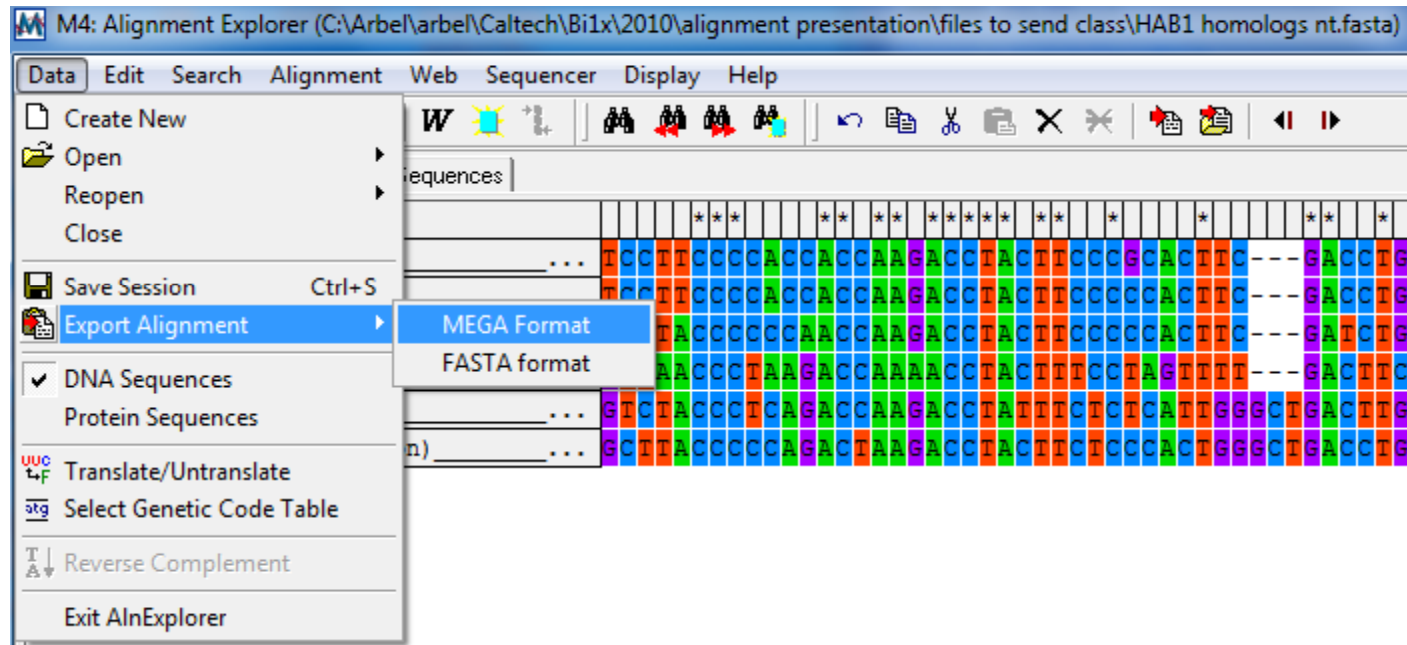
	1	2	3	4	5	6
1. Homo sapiens (human) ...						
2. Papio anubis (olive baboon) ...	0.0531					
3. Gallus gallus (chicken) ...	0.3173	0.3317				
4. Xenopus laevis (frog) ...	0.5837	0.5441	0.5020			
5. Danio rerio (zebra fish) ...	0.5393	0.5736	0.5251	0.6472		
6. Salmo salar (atlantic salmon) ...	0.5637	0.5637	0.5686	0.6928	0.3765	

## p distance

	1	2	3	4	5	6
1. Homo sapiens (human) ...						
2. Papio anubis (olive baboon) ...	0.0513					
3. Gallus gallus (chicken) ...	0.2587	0.2681				
4. Xenopus laevis (frog) ...	0.4056	0.3869	0.3660			
5. Danio rerio (zebra fish) ...	0.3846	0.4009	0.3776	0.4336		
6. Salmo salar (atlantic salmon) ...	0.3963	0.3963	0.3986	0.4522	0.2960	

# Building a Neighbor joining tree

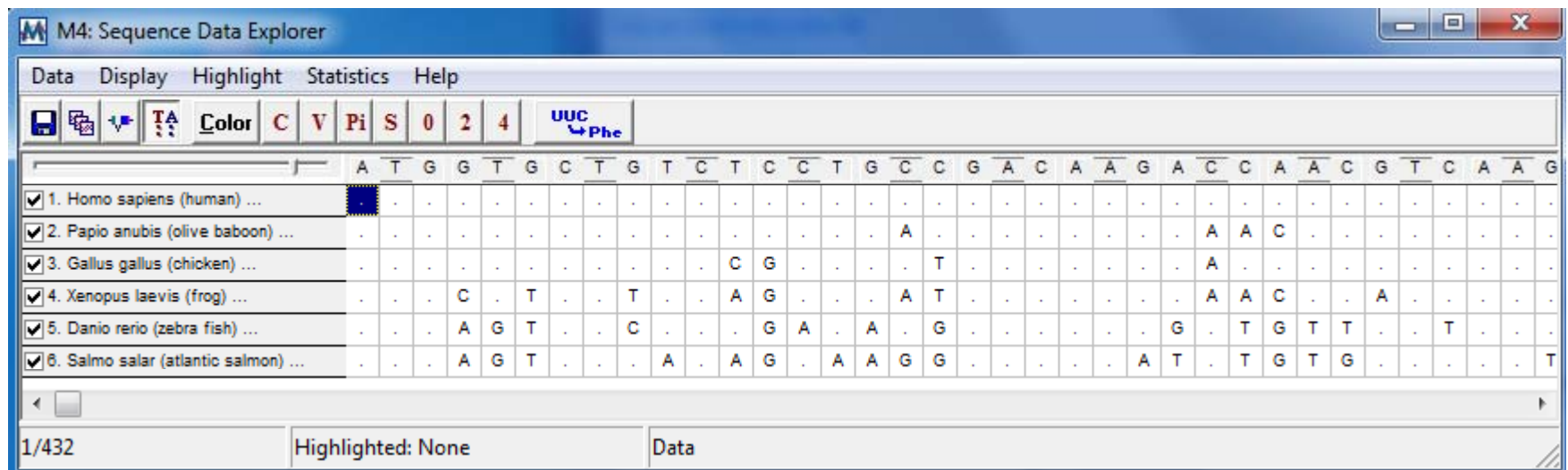
Export nt alignment in MEGA format



# Building a Neighbor joining tree

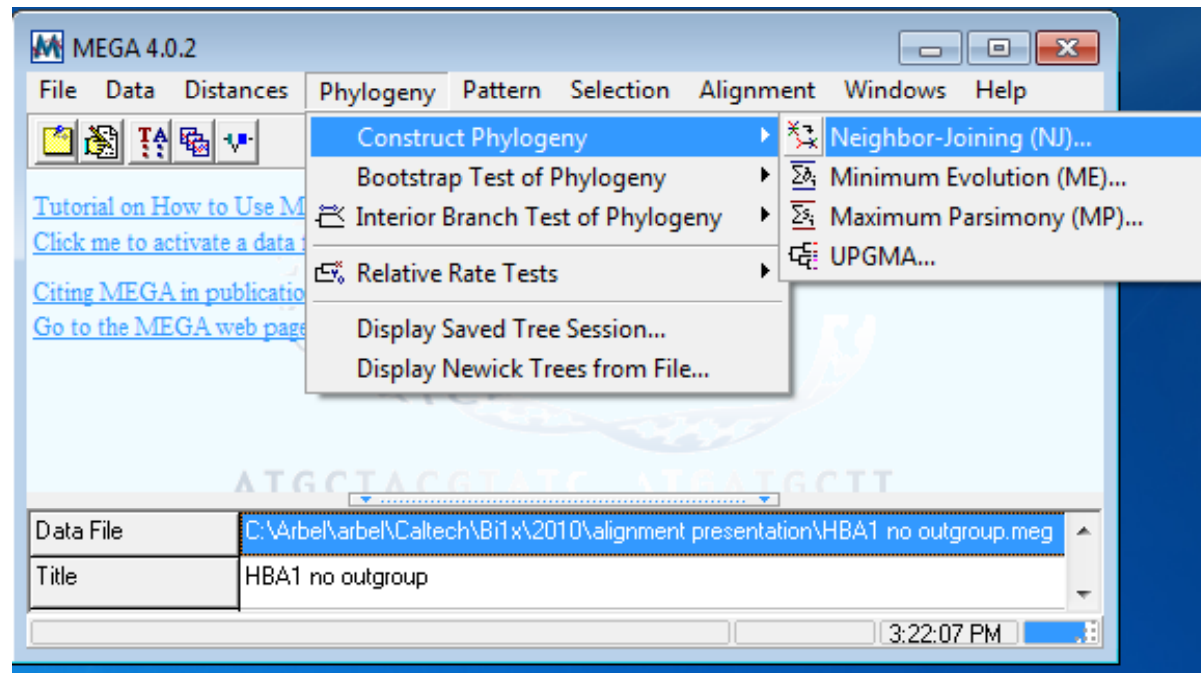
Close MEGA4 and open your MEGA file.

You should get this window with a nt sequence:



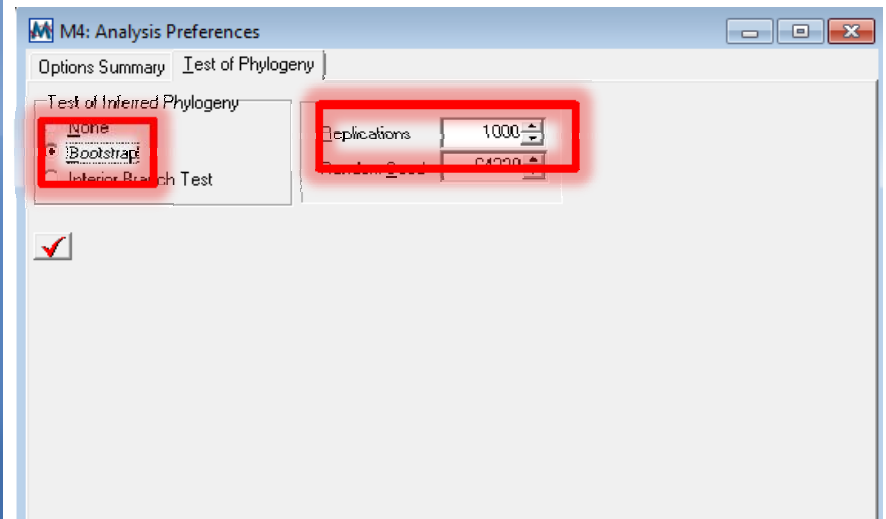
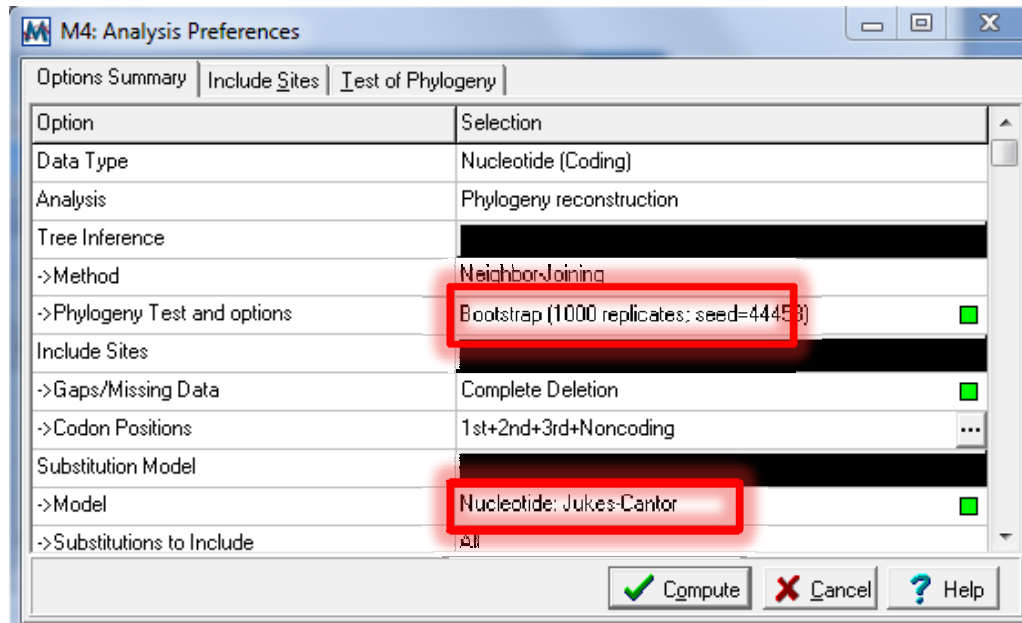
# Building a Neighbor joining tree

Now let's build a NJ tree



# Building a Neighbor joining tree

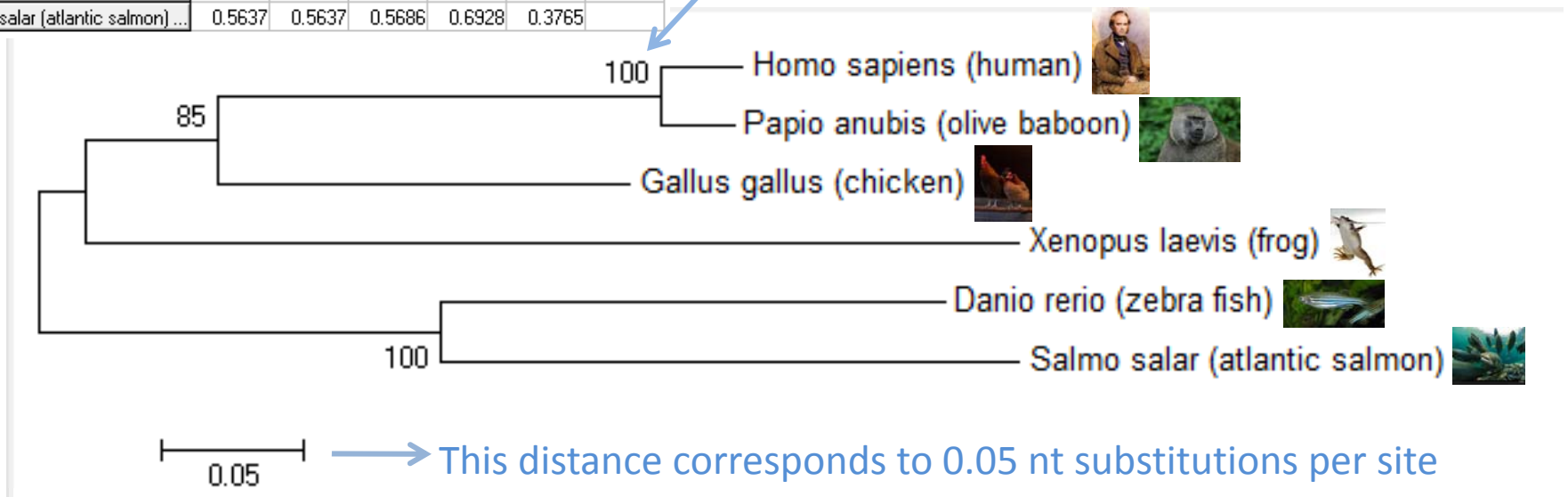
- Use JC nt model
- Calculate bootstrap support with 1000 replications



# Building a Neighbor joining tree (unrooted nt tree)

	1	2	3	4	5	6
1. Homo sapiens (human) ...						
2. Papio anubis (olive baboon) ...	0.0531					
3. Gallus gallus (chicken) ...	0.3173	0.3317				
4. Xenopus laevis (frog) ...	0.5837	0.5441	0.5020			
5. Danio rerio (zebra fish) ...	0.5393	0.5736	0.5251	0.6472		
6. Salmo salar (atlantic salmon) ...	0.5637	0.5637	0.5686	0.6928	0.3765	

Well supported node



- Human and primates group together
- Fish group together

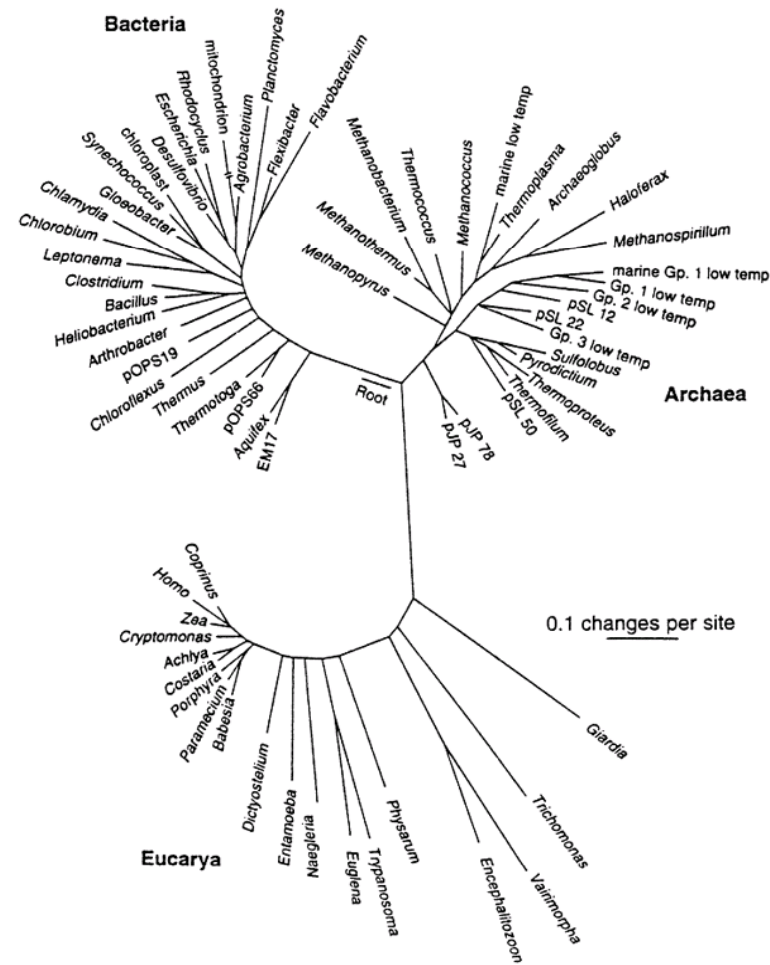
# Bioinformatics

Bi1X-2010

**Part IV:** Phylogenetic analysis of rRNA  
sequences - an exercise in class

Arbel Tadmor

# Universal phylogenetic tree based on small-subunit (SSU) rRNA sequences



- NR Pace, Science 1997



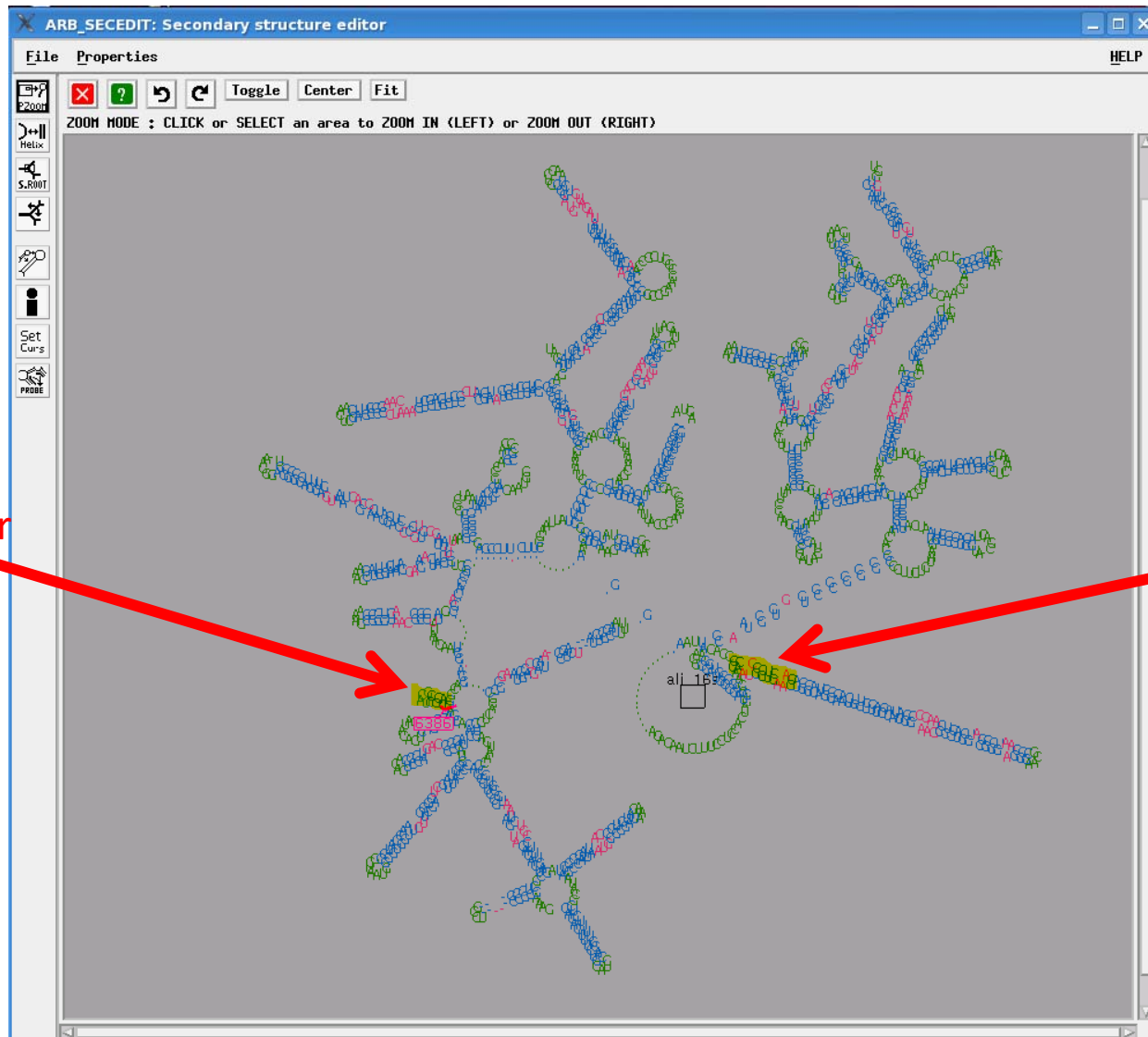
# How are SSU rRNA sequences different?

- rRNA genes are universal genes that are highly conserved
- Used as phylogenetic markers for species

## Some technical points

- No aa sequence
- Selection is directly on the nt sequence
- Alignment should take into account secondary structure of rRNA molecule
- We will therefore use a dedicated website called **green genes** to align the sequences and analyze the alignment in MEGA

# rRNA secondary structure



357F  
Forward primer

1492RL2D  
Reverse primer

# Program for today

- Learn more about the rRNA gene and the nature of the amplicon you generated
- Read the **green genes** website tutorial
- Convert your traces to nt using **Sequence Scanner**
- Align sequences with green genes
- Check for chimeras using green genes
- Import alignment into MEGA
- Calculate distance matrix
- Build a NJ tree
- Identify closest relatives of your sequences
- Is it likely that you found your phylotypes in the pond?

Additional slides

# Example 1: Global alignment of two random 300bp nucleotide sequences

We will generate random sequences of 300 nt in Matlab:

```
>> rand_int = floor(4*rand([1,300])+1);  
>> rand_nt = int2nt(rand_int)
```

rand~U(0,1)

```
rand_nt =  
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGC  
GTACAGTAGGGCGAGGCACGCTACTGTTACGAGATTCCTACCGAAGAA  
AAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATG  
AAATACTACGGGCCTTCCCCCAACATAAGGCAACTCATGCGGGGATAC  
ACATGCGCCTCGGTCCGATATGATTGCCGCATTTTCACGGTTGCCTCA  
TCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCAC  
ATCGCTATCTGT
```

# Example 1: Global alignment of two random 300bp nucleotide sequences

>Random seq 1

TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG  
 TTACGAGATTCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA  
 CTACGGGCCTTCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGCCTCGGTCCGATATGATTGCCGCATT  
 TTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCTG

>Random seq 2

CCGTAACGTCGTACAGAAGGCGTGCGAGCACCAATGTCTATCATCGGTCACTTGTGTTAGGTGTACCAAAGCTG  
 AGAGTCGTCTATCTTATCTTCTAATAGTACCTCTATTAAGATTGAGTTGTTGACCCTACAAGCAAATCGTCGCTC  
 CTCAAACCTTGTGCTCTATCAACTCTAGAATGTTGTCTAAGCCGACAGGACCGAACGGTCAATGTGGCGTTCAG  
 ATTCAGGCATTATACAAGGCCAATCGTGCGGATGCGGCAGGGGCCCTTCTAACGAAGCGGGGTGCTGGATAT

**Basic BLAST**

Choose a BLAST program to run.

[nucleotide blast](#) Search a nucleotide database using a nucleotide query  
Algorithms: blastn, megablast, discontinuous megablast

[protein blast](#) Search protein database using a protein query  
Algorithms: blastp, psi-blast, phi-blast

[blastx](#) Search protein database using a translated nucleotide query

[tblastn](#) Search translated nucleotide database using a protein query

[tblastx](#) Search translated nucleotide database using a translated nucleotide query

**Specialized BLAST**

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript and genomic libraries](#)
- [Constraint Based Error: Multiple Alignment Tool](#)
- [Needleman-Wurisch Global Sequence Alignment Tool](#)

>1cl|58213  
 Length=300

NW Score = -243  
 Identities = 165/333 (49%), Gaps = 66/333 (19%)  
 Strand=Plus/Plus

```

Query 1   TCGAAGGCGCTCGGTAGAGTACGTGTCCCA--ACTGTTGCCTAAGCGCGGTACAGTAGG  58
          ||| | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct 1   CCGTAA-CG-TCGTACAGAAGGCGTGCGAGCACCAATGTCTATCATCG-GT-CACTTGT  56

Query 59   GCGAGGCACGCTACTGTTACGAGATTCTACCGAAG----AAAAGTAAGCCCCTCG  111
          ||| | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct 57   GTTAGGTGTACCAAAGCT--GAGAGTCGTCTATCTTATCTTCTAATAGTA---CCTCTAT  111

Query 112  AAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATACTACGGGCCTTCCCCAACATA  171
          ||| | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct 112  TAAG-----ATTG-AGTTGTTGACCCTACAAGCAAATCGTCGTCGCT-CCTCAAACITG  163

Query 172  AGGC-----AACTCATGCGGGGATACACAIGCGCCTCGGTCCGATATG--ATTGCCGCA  223
          ||| | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct 164  TCGTCTATCAACTCTAGAATGTTGTCTAAGCCGACA-GGACCGAACGGTCAATGTGGCG  222

Query 224  TTTTCACGGTT---GCCTCAT-CAAGCCCGCCAACGGTTAGTGGAACGAATATGAGGCA  279
          ||| | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct 223  TT--CACGATTACGGCATTATACAAG---GCCAA-----TCGTG---CGGAT--GCGGCA  267

Query 280  GA----CT-CTCAC--ATCGCTAT-CTG----T  300
          ||| | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct 268  GGGGCCCTTCTAACGAAGCGGGGTGCTGGATAT  300
  
```

click

# Example 1: Global alignment of two random 300bp nucleotide sequences

>Random seq 1

TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG  
 TTACGAGATTCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA  
 CTACGGGCCTTCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGCCTCGGTCCGATATGATTGCCGCATT  
 TTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCTGT

>Random seq 2

CCGTAACGTCGTACAGAAGGCGTGCGAGCACCAATGTCTATCATCGGTCACTTGTGTTAGGTGTACCAAAGCTG  
 AGAGTCGTCTATCTTATCTTCTAATAGTACCTCTATTAAGATTGAGTTGTTGACCCTACAAGCAAATCGTCGTCGCTC  
 CTAACACTTGTGCTCTATCAACTCTAGAATGTTGTCTAAGCCGACAGGACCGAACGGTCAATGTGGCGTTACG  
 ATTCAGGCATTATACAAGGCCAATCGTGCGGATGCGGCAGGGGCCCTTCTAACGAAGCGGGGTGCTGGATAT

**49% identity!**

**Total length > 300bp due to gaps**

**In this case,  
both strands  
are in the  
same  
orientation**

```

>lcl|58213
Length=300

NW Score = -243
Identities = 165/333 (49%), Gaps = 66/333 (19%)
Strand=Plus/Plus

Query 1   TCGAAGGCGCTCGGTAGAGTACGTGTCCCA--ACTGTTGCCTAAGCGCGGTACAGTAGG  58
          ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 1   CCGTAA-CG-TCGTACAGAAGGCGTGCGAGCACCAATGTCTATCATCG-GT-CACTTGT  56

Query 59   GCGAGGCACGCTACTGTTACGAGATTCTACCGAAG-----AAAAGTTAAGCCCCTCG  111
          ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 57   GTTAGGTGTACCAAAGCT--GAGAGTCGTCTATCTTATCTTCTAATAGTA---CCTCTAT  111

Query 112  AAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATACTACGGGCCTTCCCCAACATA  171
          ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 112  TAAG-----ATTG-AGTTGTTGACCCTACAAGCAAATCGTCGTCGCT-CCTCAAACITG  163

Query 172  AGGC-----AACTCATGCGGGGATACACAIGCGCCTCGGTCCGATATG--ATTGCCGCA  223
          ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 164  TCGCTCTATCAACTCTAGAATGTTGTCTAAGCCGACA-GGACCGAACGGTCAATGTGGCG  222

Query 224  TTTTCACGGTT---GCCTCAT-CAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCA  279
          || |||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 223  TT--CACGATTACGGCATTATACAAG---GCCAA-----TCGTG---CGGAT--GCGGCA  267

Query 280  GA----CT-CTCAC--ATCGCTAT-CTG----T  300
          | ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 268  GGGGCCCTTCTAACGAAGCGGGGTGCTGGATAT  300

```

# Example 2: Local alignment of a 300bp sequence and an internal **fragment** containing a single **insertion** and a single **mismatch** using BLAST

>Random seq 1

```
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCCTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATTGCGGGGATACACATGCGCTCGGTCCGATATGATTGCCGCA
TTTTACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT
```

**Basic BLAST**  
Choose a BLAST program to run.

**nucleotide blast** Search a nucleotide database using a nucleotide query  
*Algorithms: blastn, megablast, blastn-short, megablast-short*

**protein blast** Search protein database using a protein query  
*Algorithms: blastp, psi-blast, phi-blast*

**blastx** Search protein database using a translated nucleotide query

**tblastn** Search translated nucleotide database using a protein query

**tblastx** Search translated nucleotide database using a translated nucleotide query

**Specialized BLAST**  
Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- Align two (or more) sequences** using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript and genomic libraries](#)
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch [Global Sequence Alignment Tool](#)



**BLAST** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastn suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number, gi, or FASTA sequence  Clear Query subrange

ICGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG  
TTACGAGATTCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCCTGATCTGGCATGAAATA  
CTACGGGCCTTCC**CCCAACATAAGGCAACTCAT****TGCGGGGATACACATGCG****CTCG**GTCCGATATGATTGCCGCA  
TTTTACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGAACGAATATGAGGCAGACTCTCACATCGCTATCT  
GT

Or, upload file  Browse...

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence  Clear Subject subrange

CCCAACATAAGGCAACTCAT**TGCGGGGATACACATGCG**ACTCG

Or, upload file  Browse...

Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

**Choose an algorithm (blastn)**

**BLAST** Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

Use blastn algorithm

click



## Example 2: Local alignment of a 300bp sequence and an internal **fragment** containing a single **insertion** and a single **mismatch** using BLAST

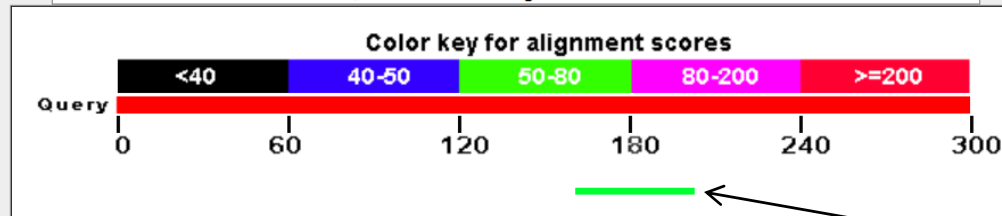
>Random seq 1

```
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCCTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGACTCGGTCCGATATGATTGCCGCA
TTTTACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT
```

### ▼ Graphic Summary

Distribution of 1 Blast Hits on the Query Sequence ⓘ

Mouse over to see the define, click to show alignments



Our aligned fragment

### ▼ Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer

#### Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
63129		66.2	66.2	14%	1e-16	95%	

## Example 2: Local alignment of a 300bp sequence and an internal **fragment** containing a single **insertion** and a single **mismatch** using BLAST

>Random seq 1

```
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCCTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATTGCGGGGATACACATGCGACTCGGTCCGATATGATTGCCGCA
TTTTACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT
```

▼ **Alignments**  
 Select All [Get selected sequences](#)

```
>lcl|63129
Length=43

Score = 66.2 bits (72), Expect = 1e-16
Identities = 41/43 (95%), Gaps = 1/43 (2%)
Strand=Plus/Plus
```

Query	163	CCCAACATAAGGCAACTCAT-GCGGGGATACACATGCGCCTCG	204
Sbjct	1	CCCAACATAAGGCAACTCATTGCGGGGATACACATGCGACTCG	43

**Indel** (points to the gap in the query sequence)

**mismatch** (points to the difference between the query and subject sequences at the end)

# Example 3: Local alignment of two random 300bp nucleotide sequences using BLAST

**Basic BLAST**

Choose a BLAST program to run.

**nucleotide blast** Search a nucleotide database using a nucleotide query  
Algorithms: blastn, tblastn, tblastx

**protein blast** Search protein database using a protein query  
Algorithms: blastp, psi-blast, phi-blast

**blastx** Search protein database using a translated nucleotide query

**tblastn** Search translated nucleotide database using a protein query

**tblastx** Search translated nucleotide database using a translated nucleotide query

**Specialized BLAST**

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript and genomic libraries](#)
- Constraint Based Protein [Multiple Alignment Tool](#)
- Needleman-Wunsch [Global Sequence Alignment Tool](#)

click



**BLAST** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastn suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Or, upload file

Job Title

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence

Or, upload file

Program Selection

Optimize for

Highly similar sequences (megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm (blast)

**BLAST** Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)

click

Use blastn algorithm

**BLAST** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Reis]

NCBI/BLAST/blastn suite-2sequences/ Formatting Results - WYK7V7X9114

Edit and Resubmit Save Search Strategies Formatting options Download

**Blast 2 sequences**

**Nucleotide Sequence (300 letters)**

Query ID	Id 28033	Subject ID	28035
Description	None	Description	None
Molecule type	nucleic acid	Molecule type	nucleic acid
Query Length	300	Subject Length	300
		Program	BLASTN 2.2.23+ Citation

No significant similarity found. For reasons why, click here

Other reports: Search Summary

number of matches of length L

$4^{-L} \times 300^2 \times 2$

# Example 3: Local alignment of two random 300bp nucleotide sequences using BLAST

Algorithm parameters Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences: 100  
Select the maximum number of aligned sequences to display

Short queries:  Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 7 Reduce word size from 11 to 7

Scoring Parameters

Match/Mismatch Scores: 2,-3

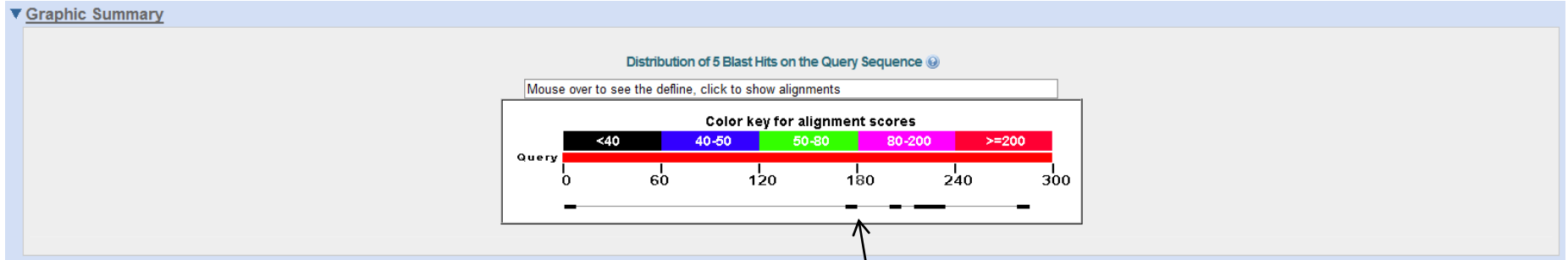
Gap Costs: Existence: 5 Extension: 2

Filters and Masking

Filter:  Low complexity regions  
 Species-specific repeats for: Human

Mask:  Mask for lookup table only  
 Mask lower case letters

# Example 3: Local alignment of two random 300bp nucleotide sequences using BLAST



Tiny fragments were aligned

▼ **Alignments**

Select All [Get selected sequences](#)

```
>cl145871
Length=300

Score = 17.5 bits (18), Expect = 0.45
Identities = 15/19 (78%), Gaps = 0/19 (0%)
Strand=Plus/Minus
Query 217 TGCCGCATTTTCACGGTTG 235
Sbjct 267 TGCCGCATCCGCACGATTG 249

Score = 15.7 bits (16), Expect = 1.6
Identities = 8/8 (100%), Gaps = 0/8 (0%)
Strand=Plus/Minus
Query 280 GACTCTCA 287
Sbjct 81 GACTCTCA 74

Score = 13.9 bits (14), Expect = 5.4
Identities = 7/7 (100%), Gaps = 0/7 (0%)
Strand=Plus/Plus
Query 3 GAAGGCG 9
Sbjct 17 GAAGGCG 23

Score = 13.9 bits (14), Expect = 5.4
Identities = 7/7 (100%), Gaps = 0/7 (0%)
Strand=Plus/Minus
Query 175 CAACTCA 101
Sbjct 124 CAACTCA 118

Score = 13.9 bits (14), Expect = 5.4
Identities = 7/7 (100%), Gaps = 0/7 (0%)
Strand=Plus/Minus
Query 202 TCGGTCC 208
Sbjct 207 TCGGTCC 201
```

Score = 17.5 bits (18), Expect = 0.45  
Identities = 15/19 (78%), Gaps = 0/19 (0%)  
Strand=Plus/Minus

Query 217 TGCCGCATTTTCACGGTTG 235  
Sbjct 267 TGCCGCATCCGCACGATTG 249

## Example 3: Global alignment of two random 300 residue amino acid sequences

We will generate random sequences of 300 aain Matlab:

```
>> rand_int = floor(20*rand([1,300])+1);  
>> rand_aa = int2aa(rand_int)
```

```
rand_aa =  
EMQSSVHIKTADYYITYFGHHFIVGEWLPNIRFPYFFWIITTDARNDM  
AIFNCQDETQSKKPSYNSDANNNYQYMWGCDLQEAKVTAMGNLHLWNH  
RGPRFQKDHACQLCEPHRGITETKRQKIDCSMNPHIPHARKHYRGLNY  
MYMAENMRFIELPTEEEFWSWWDWVSWRMEMWGS DLMPEQYMRMDSWE  
NSEQCRKHSIGRCLHYRHLNLWDDRFAQVSCFNMWWEIFPIGQRHDGY  
LYRVRESMIQNQDENTVCPAMFAANWQLLKEHHVIRGSKEYREWFFINV  
WHTEGSRVAQAH
```

# Example 3: Global alignment of two random 300 residue amino acid sequences

>Random seq 1

```
EMQSSVHIKTADYYITYFGHHFIVGEWLPNIRFPYFFWIITTDARNDMAIFNCQDETQSKKPSYNSDANNNYQYMW
GCDLQEAKVTAMGNLHLWNHRGPRFQKDHACQLCEPHRGITETKRQKIDCSMNPHIPHARKHYRGLNYMYMAE
NMRFIELPTEEEFWSWWDWVSWRMEMWGS DLMPEQYMRMDSWENSEQCRKHSIGRCLHYRHLNLWDDRFA
QVSCFNMWWEIFPIGQRHDGYLYRVRESMIQNQDENTVCPAMFAANWQLLKEHHVRSKEYREWFFINVWHT
GSRVAQAH
```

>Random seq 2

```
SYLTKSAIQEVCLCKVNVNDMNRFAVLGPYGFMSKGWCSVQPFHIYVPGAKKGMQRQCETDDLMDMTSQE
DHEQYGGKRCPKCTHLLDRIAHKMAPDKMSRWGGGEKQGE MFVYGDYQKYRQHKWCVHSLEHPYWNNWFALW
GQCCGKQMTNPMIYRKCAKTKTCMDQAPVPSLQCQVCLCHNGSTYLTPANCCDCQVEQHESGNMGRWIRYQ
MFCVFLWKAITPKAPHFSTASKRQNKLRVQEEALQHYYNKGPLQIWPDDGWF MNRWHIILQCWYMGKFWRLH
MKCNARESEMVML
```

```
>lcl|31039 unnamed protein product
Length=300
```

**28% similarity**

**only 12% identity**

```
NW Score = -118
Identities = 46/354 (12%) Positives = 100/354 (28%), Gaps = 108/354 (30%)
```

```
Query 1 EMQSS-----VHIKTADYYITYFGHHFIVGEWLPNIRFPYFF-----WIITDA 44
      + V++ + + + F+ W E + W++
Sbjct 1 SYLTKSAIQEVCLCKVNVNDMNRFAVLGPYGFMSKGWCSVQPFHIYVPGAKKGMQRQC 60

Query 45 RND--MAIFNCQDETQ---SKKPSYNSDANNNYQYM-----WGCDLQEAKVTAMGNLH 92
      D M + + + D Q + P + M WG + Q ++ G+
Sbjct 61 ETDDLMDMTSQEDHEQYGGKRCPKCTHLLDRIAHKMAPDKMSRWGGGEKQ-GEMFVYGDYQ 119

Query 93 -----LWNHRGPRFQKDHACQLCEP--HRGITETKRQKIDCSMNPHIPHA 135
      WN+ + + Q+ P +R +TK C +P
Sbjct 120 KYRQHKWCVHSLEHPYWNNWFALWGQCCGKQMTNPMIYRKCAKTKT----CMDQAPVPSL 175

Query 136 RKH----YRGLNYMYMAENMRFIELPTEEEFWSWWDWVSWRME---MWGSDLMPE--QYM 186
      + + G Y+ A N ++ E W+ ++M +W + + P+ +
Sbjct 176 QCQVCLCHNGSTYLTPA-NCCDCQVEQHESGNMGRWIRYQMFCVFLWKA-ITPKAPHFS 233

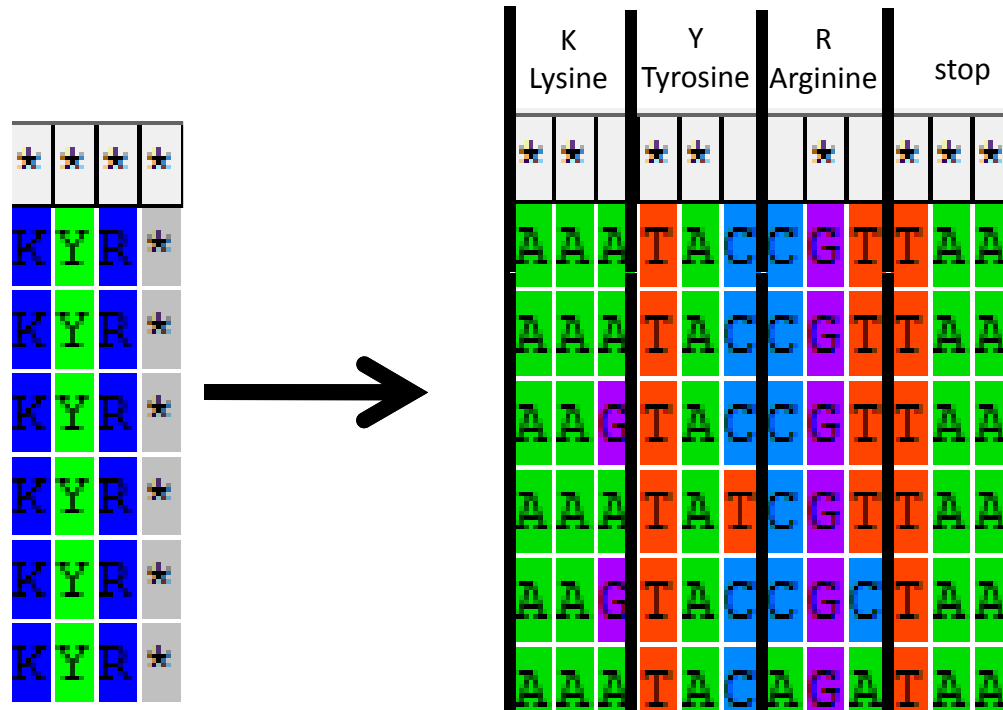
Query 187 RMDSWENSEQCRKHSIGRCLHYRHLNLWDDRFAQVSCFNMWWEIFPIGQRHDGYLYRVRE 246
      +N + ++ ++ + L+W D DG+
Sbjct 234 TASKRQNKLRVQEEALQHYYNKGPLQIWPD-----DGW----- 266

Query 247 SMIQNQDENTVCPAMFAANWQLLKEHHVRSKEYREWFFINVWHTGSRVAQAH 300
      F W ++ + G K +R N +E +
Sbjct 267 -----FMNRWHIILQCWYMG-KFWRLHMKCNARESEMVML 300
```

For >100 aa, >25% identity is required to say with almost certainty that an alignment is not the result of chance

# Selection pressure is primarily on the level of the amino acids

Let's look at the three residues at the C terminus

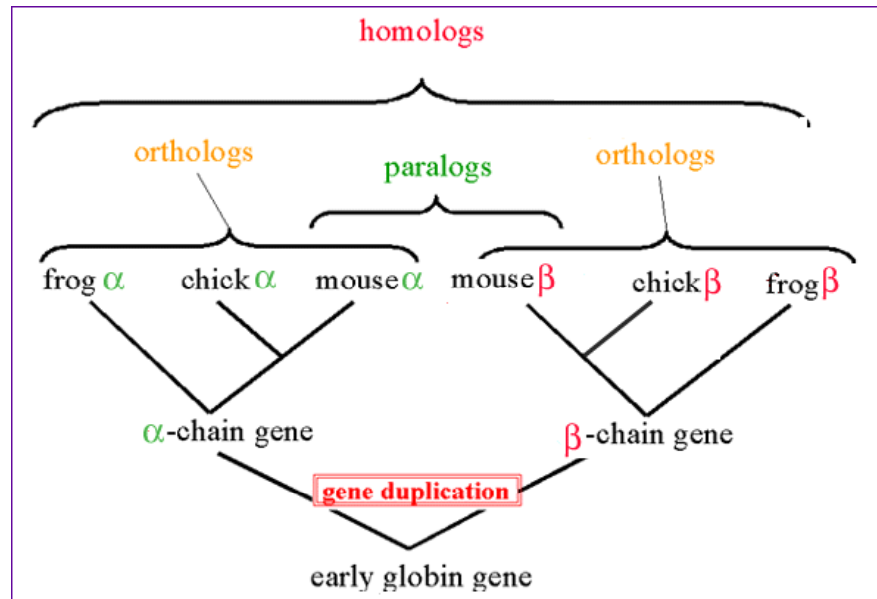


## The Genetic Code

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	
5' GCA	CGA	AAC	GAC	UGC	CAA	GAA	GGA	CAC	AUA	CUA	AAA	AUG	UUC	CCA	UCA	ACA	UGG	UAC	GUA	3'
C	C	U	U	U	G	G	C	U	C	C	G		U	C	C	C		U	C	
G	G						G		U	G				G	G	G			G	
U	U						U			U				U	U	U			U	
	or									or					or					
	AGA									UUA					AGC					
	G									G				U	U					



# Homologs, orthologs, paralogs

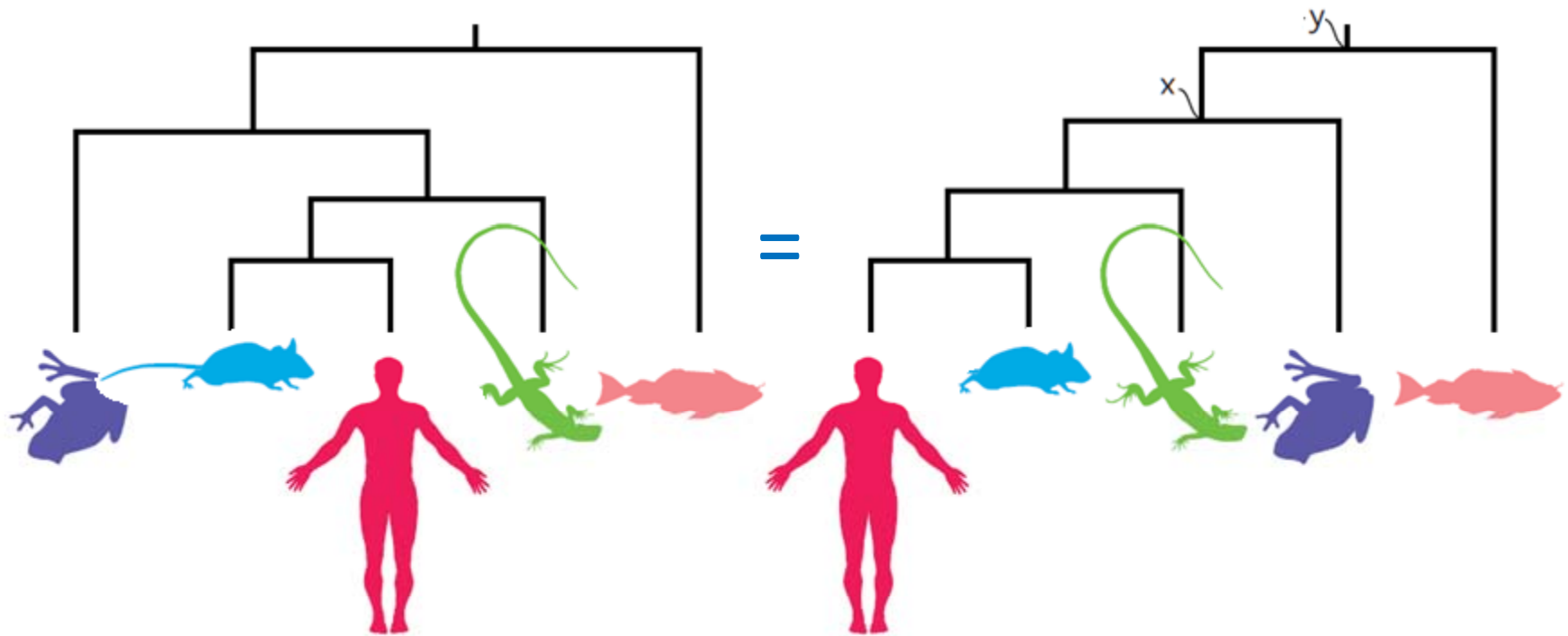


**Homologs:** similar sequences in two different organisms derived from a common ancestor sequence.

**Orthologs:** Similar sequences in two different organisms that have arisen due to a speciation event. Functionality has been retained.

**Paralogs:** Similar sequences within a single organism that have arisen due to a gene duplication event. Functionality has diverged.

# Tree challenge



Is the **frog** more closely related to the **fish** or the **human**?

Suggested reading: **The Tree-Thinking Challenge**, Baum et al. Science 2005