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 (match) = +2 or 4X4 nt substitution matrix

w (mismatch) = -1 or 4X4 nt substitution matrix

w (gap) = -3 or some other heuristic penalty

Find alignment with maximum score

Substitution matrix for amino acids
BLOSUM: **Blo**cks **Su**bstitution **M**atrix
(used to calculate an alignment score)

- Values in matrix are <u>empirical</u>. 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																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

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

Positive for chemically similar substitutions (more likely than chance)

Unlikely substitution compared with chance

Common aa have a low score

Rare aa have a high score

- *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- **B**asic **L**ocal **A**lignment **S**earch **T**ool

The Google of biology (*google blast*)

nt vs. **nt**

aa vs. **aa**

tras nt vs. **aa** database

aa vs. **trans nt** database

tras nt vs**. trans nt** database

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous 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 Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool

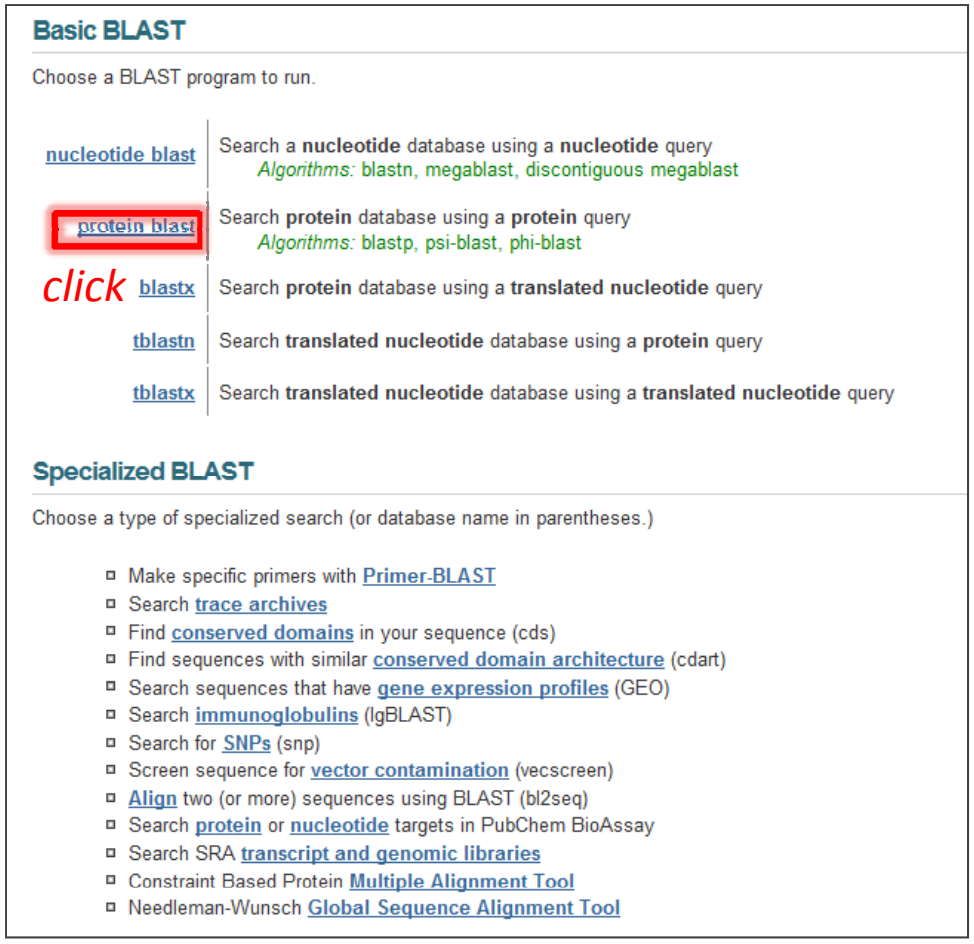
Global alignment →

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 λ parameter in a Poisson distribution $\lambda^n e^{-\lambda}/n!$ …)

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

Main BLAST window



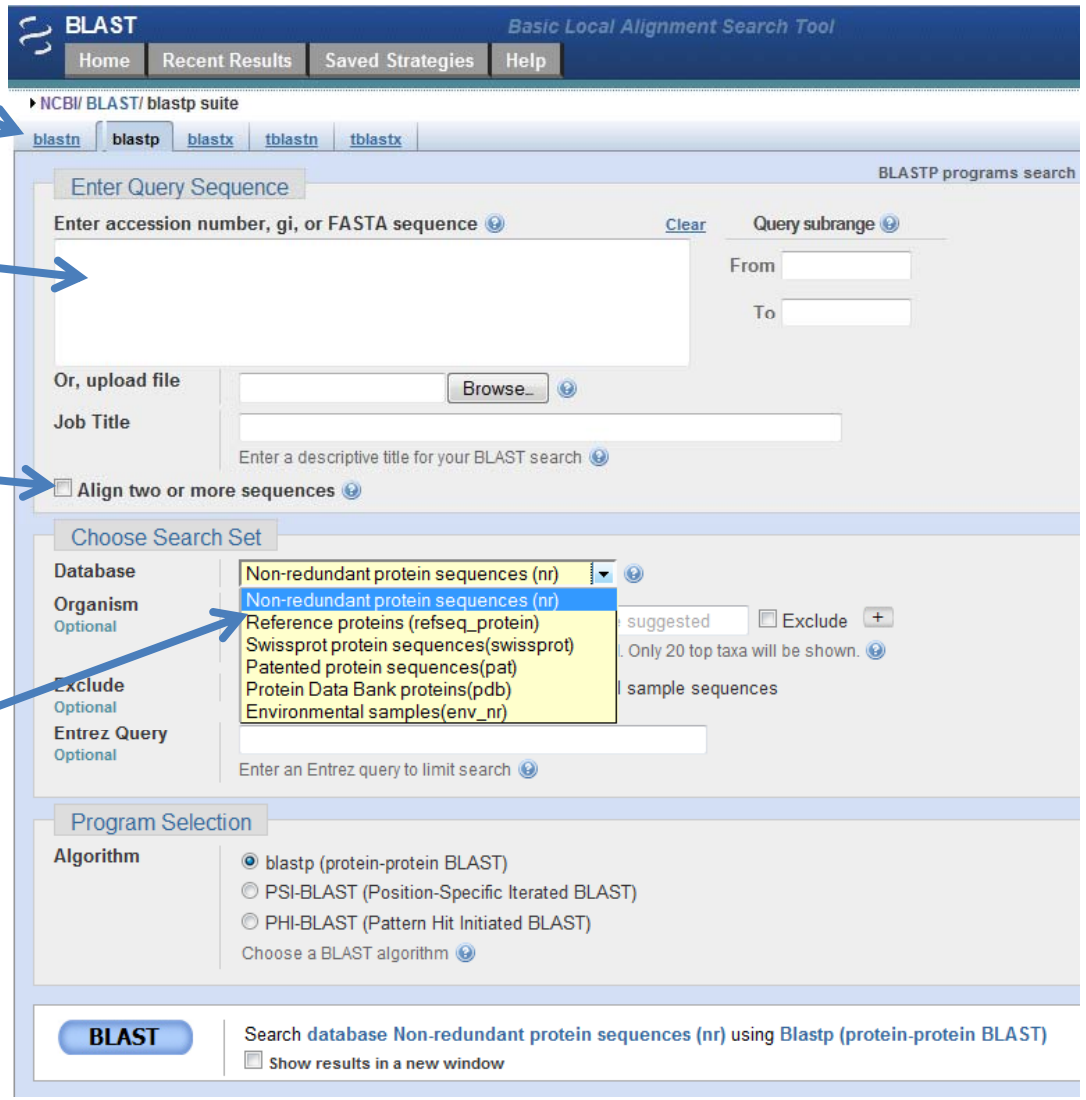
Main BLAST window

Choose BLAST tool

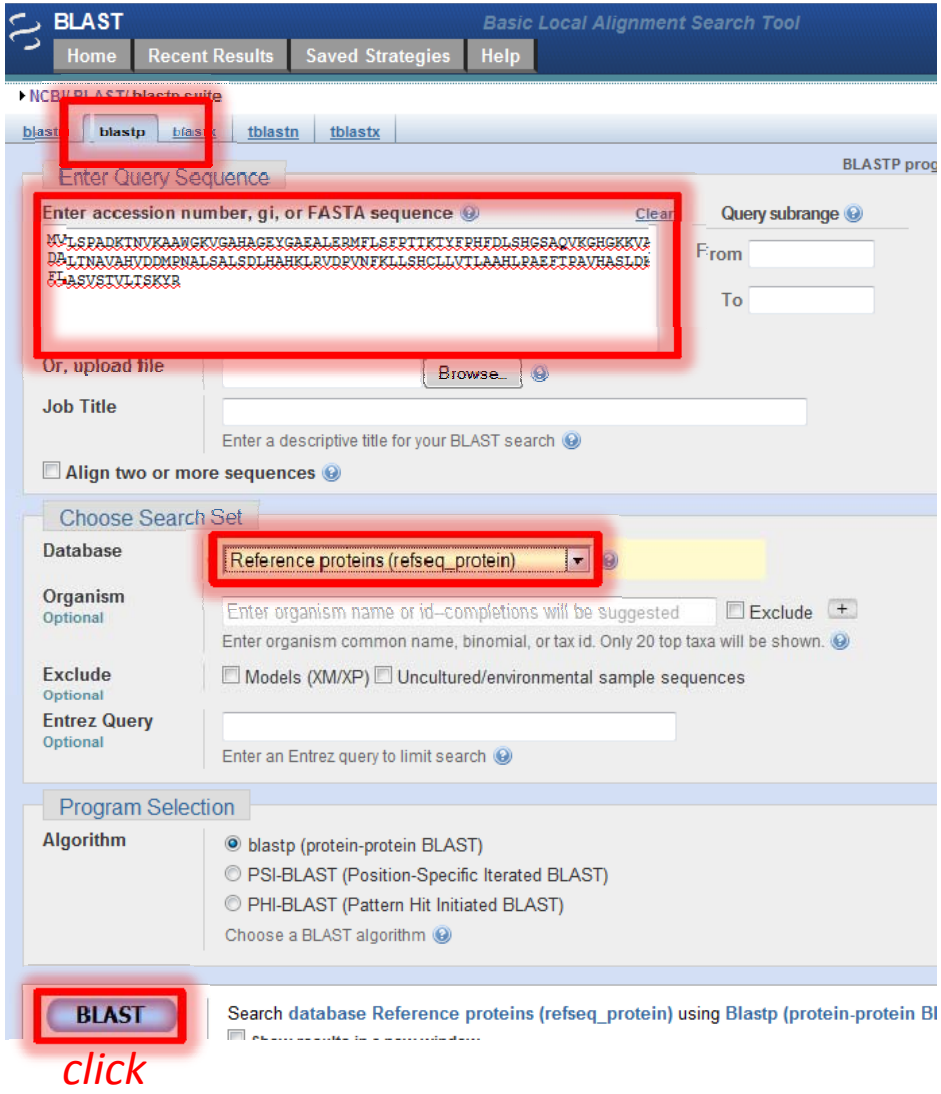
Paste nt or aa sequence (=**query**) here

Click here to align two sequences one against the other

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



Let’s search now for homologs of human HBA1 in the refseq_protein database

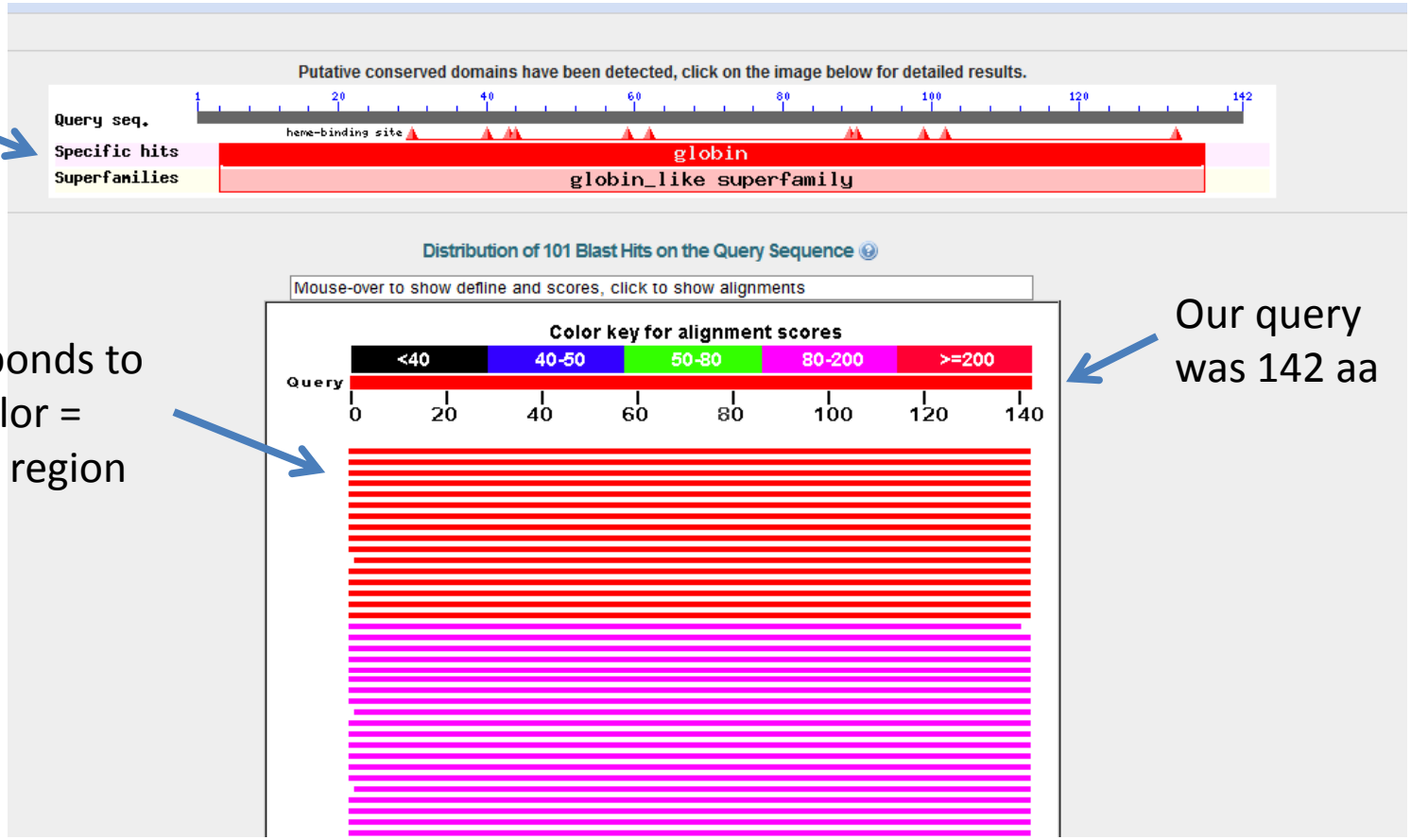


Results

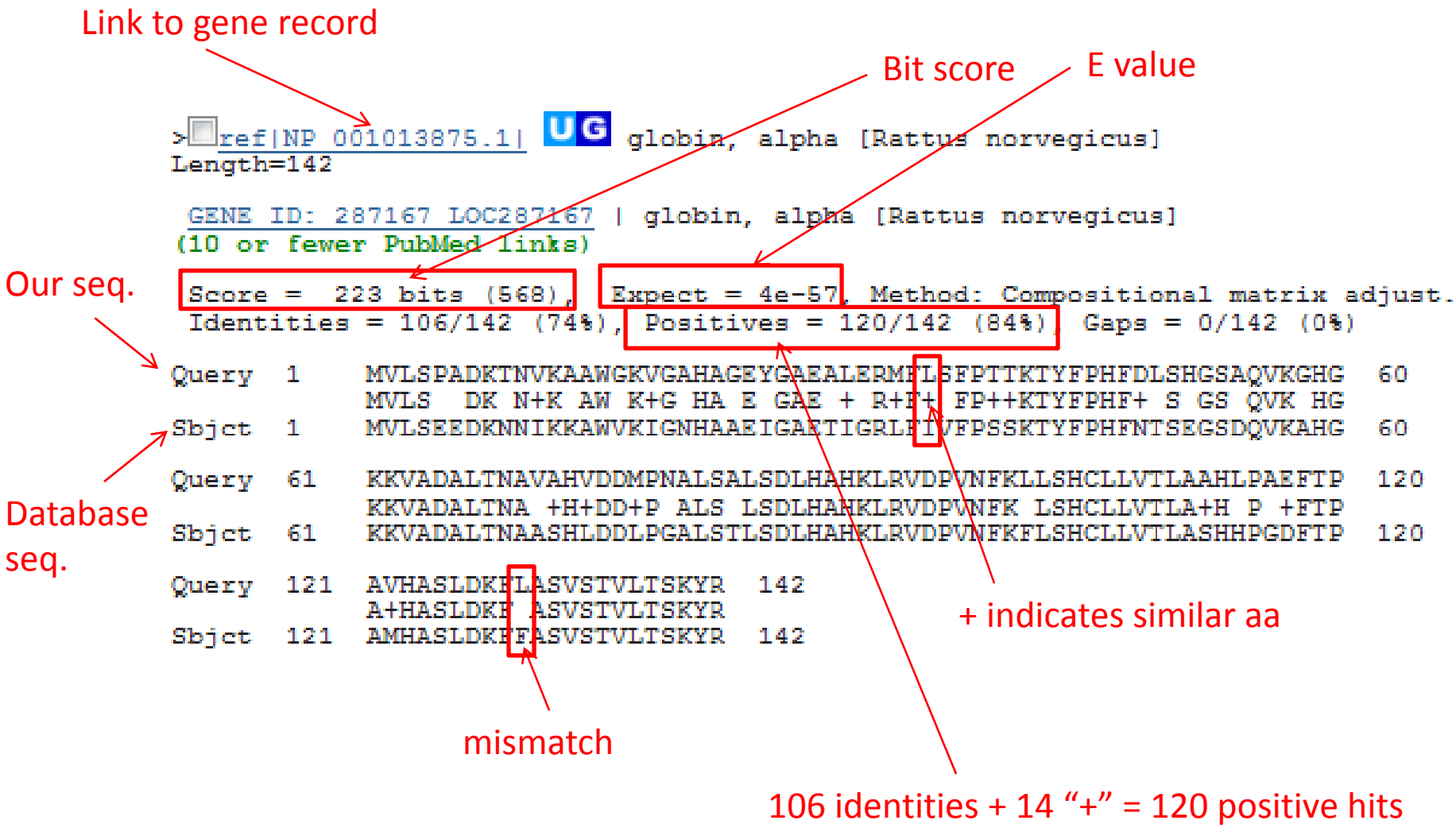
Conserved domains in the protein

Each line corresponds to a positive hit. Color = score for aligned region

Our query was 142 aa



Example of local alignment result

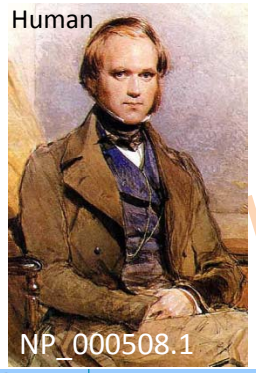


Results

Here's our sequence NP_000549.1

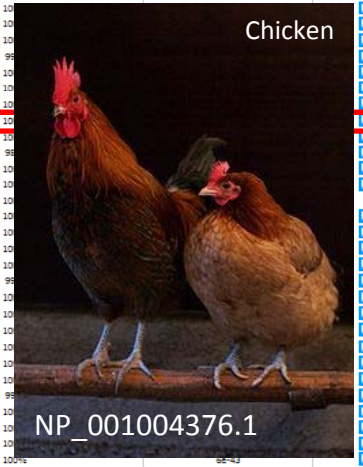
Accession	Description	Max score	Total score	Query coverage	E value	Links
NP_000549.1	hemoglobin subunit alpha [Homo sapiens] >ref|NP_000549.1| hemoglobin subunit alpha [Homo sapiens] >ref|NP_001036092.1| hemoglobin subunit alpha [Pan troglodytes] >ref|NP_001036091.1|	288	288	100%	3e-76	U G
XP_001094436.1	PREDICTED: similar to alpha 2 globin [Macaca mulatta]	276	276	100%	3e-73	U G
NP_001038189.1	theta 1 globin [Macaca mulatta]	276	276	100%	5e-73	U G
NP_001162287.1	hemoglobin, alpha 1 [Papio anubis]	262	262	100%	9e-69	U G
NP_001077424.1	hemoglobin alpha, adult chain 2 [Mus musculus] >ref|NP_032244.2| hemoglobin subunit alpha [Mus musculus]	255	255	100%	1e-66	U G
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	U G
NP_001078901.1	hemoglobin subunit alpha [Equus caballus]	251	251	100%	1e-65	U G
NP_001078900.1	alpha 2 globin [Equus caballus]	250	250	100%	2e-65	U G
NP_001075838.1	hemoglobin subunit alpha-1/2 [Oryctolagus cuniculus]	248	249	100%	7e-65	U G
NP_001013875.1	globin, alpha [Rattus norvegicus]	223	223	100%	4e-57	U G
XP_001517140.1	PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha chain) (Alpha-globin) [Ornithorhynchus anatinus]	216	216	99%	5e-55	U G
NP_001007723.1	hemoglobin alpha 2 chain [Rattus norvegicus]	211	211	100%	2e-53	U G
NP_001028158.1	hemoglobin subunit alpha [Monodelphis domestica]	210	210	100%	2e-53	U G
NP_037228.1	hemoglobin alpha, adult chain 2 [Rattus norvegicus]	209	209	100%	8e-53	U G
NP_001004376.1	hemoglobin subunit alpha-A [Gallus gallus]	207	207	100%	2e-52	U G
XP_002196132.1	PREDICTED: putative hemoglobin alpha [Taeniopygia guttata]	206	206	100%	5e-52	U G
XP_356935.3	PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha chain) (Alpha-globin) [Mus musculus] >ref|XP_920048.2| PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha	196	196	98%	7e-49	U G
XP_002196164.1	PREDICTED: hemoglobin, zeta [Taeniopygia guttata]	191	191	100%	2e-47	U G
NP_001079746.1	hemoglobin, alpha 2 [Xenopus laevis]	185	185	100%	9e-46	U G
XP_002724712.1	PREDICTED: mCG1039741-like [Rattus norvegicus] >ref|XP_002727804.1| PREDICTED: mCG1039741-like [Rattus norvegicus]	184	184	100%	2e-45	
NP_001029153.1	hemoglobin, theta T1 [Mus musculus]	184	184	100%	2e-45	U G
NP_001162288.1	hemoglobin, theta 1 [Papio anubis]	184	184	100%	2e-45	U G
NP_001164885.1	hemoglobin subunit zeta [Oryctolagus cuniculus]	184	184	100%	3e-45	U G
XP_001510395.1	PREDICTED: similar to zeta-1-globin [Ornithorhynchus anatinus]	182	182	100%	6e-45	U G
NP_001004375.1	hemoglobin subunit alpha-D [Gallus gallus]	182	182	99%	8e-45	U G
NP_001164886.1	zeta globin [Oryctolagus cuniculus]	182	182	100%	8e-45	U G
NP_001004374.1	hemoglobin subunit pi [Gallus gallus]	182	182	100%	8e-45	U G
NP_005322.1	hemoglobin subunit theta-1 [Homo sapiens]	182	182	100%	1e-44	U G
XP_001915477.1	PREDICTED: similar to Hemoglobin subunit theta-1 (Hemoglobin theta-1 chain) (Theta-1-globin) [Equus caballus]	181	181	100%	1e-44	U G
NP_001164873.1	hemoglobin, zeta [Oryctolagus cuniculus]	181	181	100%	2e-44	U G
NP_001079797.1	hemoglobin alpha 3 subunit [Xenopus laevis]	178	178	47%	1e-43	U G
XP_002196147.1	PREDICTED: putative hemoglobin alpha-D chain [Taeniopygia guttata]	177	177	99%	2e-43	U G
NP_001079749.1	hemoglobin subunit alpha-4 [Xenopus laevis]	177	177	100%	2e-43	U G
NP_001164887.1	theta 1 globin [Oryctolagus cuniculus] >ref|NP_001164972.1| hemoglobin subunit theta-1 [Oryctolagus cuniculus]	176	176	100%	4e-43	U G
NP_001133726.1	hemoglobin, alpha 2 [Xenopus (Silurana) tropicalis]	176	176	100%	5e-43	U G
NP_001108014.1	hemoglobin subunit zeta [Equus caballus]	176	176	100%	6e-43	U G
NP_005323.1	hemoglobin subunit zeta [Homo sapiens]	176	176	100%	6e-43	U G
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	U G
XP_001372225.1	PREDICTED: similar to pol polyprotein [Monodelphis domestica]	176	348	100%	2e-42	U G
NP_034535.1	hemoglobin subunit zeta [Mus musculus]	173	173	100%	4e-42	U G
NP_001135556.1	hypothetical protein LOC100216102 [Xenopus (Silurana) tropicalis]	172	172	100%	8e-42	U G
NP_001015904.1	hemoglobin alpha 3 subunit [Xenopus (Silurana) tropicalis]	172	172	100%	1e-41	U G
XP_001788743.1	PREDICTED: similar to Hemoglobin subunit zeta (Hemoglobin zeta chain) (Zeta-globin) [Bos taurus]	171	171	100%	1e-41	U G
NP_001073802.1	hemoglobin, theta 1 [Bos taurus]	171	171	96%	2e-41	U G
NP_001081493.1	hemoglobin, alpha 1 [Xenopus laevis]	169	169	100%	5e-41	U G
NP_988860.1	hemoglobin, alpha 1 [Xenopus (Silurana) tropicalis]	166	166	100%	4e-40	U G
XP_580707.3	PREDICTED: similar to Hemoglobin subunit zeta (Hemoglobin zeta chain) (Zeta-globin) [Bos taurus]	165	165	98%	1e-39	U G
NP_001165373.1	alpha globin larval-3 [Xenopus (Silurana) tropicalis]	165	165	100%	1e-39	G
XP_001061675.2	PREDICTED: hemoglobin, theta 1 [Rattus norvegicus] >ref|XP_347267.4| PREDICTED: hemoglobin, theta 1 [Rattus norvegicus]	164	164	100%	2e-39	
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	U G
NP_001165374.1	alpha globin larval-4 [Xenopus (Silurana) tropicalis]	163	163	100%	4e-39	G
NP_001016009.1	alpha globin larval-5 [Xenopus (Silurana) tropicalis]	163	163	100%	4e-39	U G
NP_001005092.1	hemoglobin, zeta [Xenopus (Silurana) tropicalis]	162	162	100%	1e-38	U G
NP_001118023.1	alpha-globin IV [Oncorhynchus mykiss]	158	158	100%	2e-37	U G
NP_001107321.1	alpha globin larval-2 [Xenopus (Silurana) tropicalis]	155	155	100%	1e-36	U G
NP_001089702.1	hypothetical protein LOC734765 [Xenopus laevis]	154	154	100%	3e-36	U G
XP_002723955.1	PREDICTED: zeta globin-like [Oryctolagus cuniculus]	153	153	98%	5e-36	G
NP_571332.2	hemoglobin alpha adult-1 [Danio rerio]	153	153	100%	5e-36	U G
NP_001013479.1	hypothetical protein LOC497166 [Danio rerio]	152	152	100%	8e-36	U G
NP_001028265.1	alpha globin-like [Danio rerio]	152	152	100%	8e-36	U G
XP_002723956.1	PREDICTED: zeta globin-like [Oryctolagus cuniculus]	151	151	98%	2e-35	G
NP_778165.1	hemoglobin, theta T2 [Mus musculus]	151	151	88%	2e-35	U G
NP_001117134.1	hemoglobin subunit alpha [Salmo salar]	150	150	100%	3e-35	U G
NP_001085638.1	hemoglobin, zeta [Xenopus laevis]	150	150	100%	4e-35	U G

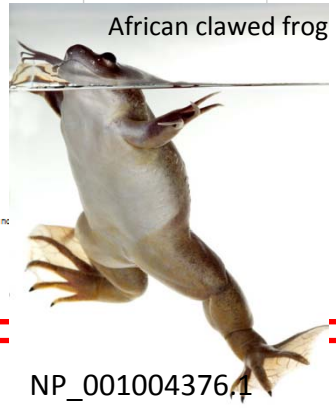
Results



Here's our sequence
NP_000549.1

Accession	Description	Max score	Total score	Query coverage	E value	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|	288	288	100%	3e-76	UG
XP_001094436.1	PREDICTED: similar to alpha 2 globin [Macaca mulatta]	276	276	100%	3e-73	UG
NP_001038189.1	theta 1 globin [Macaca mulatta]	276	276	100%	5e-73	UG
NP_001162287.1	hemoglobin, alpha 1 [Papio anubis]	262	262	100%	9e-69	UG
NP_001077426.1	hemoglobin alpha, adult chain 2 [Mus musculus] >ref|NP_032244.2| hemoglobin subunit alpha [Mus musculus]	255	255	100%	1e-66	UG
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	UG
NP_001078901.1	hemoglobin subunit alpha [Equus caballus]	251	251	100%	1e-65	UG
NP_001078900.1	alpha 2 globin [Equus caballus]	250	250	100%	2e-65	UG
NP_001075838.1	hemoglobin subunit alpha-1/2 [Oryctolagus cuniculus]	249	249	100%	3e-65	UG
NP_001013875.1	globin, alpha [Rattus norvegicus]	223	223	100%	3e-56	UG
XP_001517140.1	PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha chain) (Alpha-globin) [Ornithorhynchus anatinus]	216	216	99%	4e-54	UG
NP_001007723.1	hemoglobin alpha 2 chain [Rattus norvegicus]	211	211	100%	1e-52	UG
NP_001028158.1	hemoglobin subunit alpha [Monodelphis domestica]	210	210	100%	2e-52	UG
NP_037228.1	hemoglobin alpha, adult chain 2 [Rattus norvegicus]	209	209	100%	4e-52	UG
NP_001004376.1	hemoglobin subunit alpha-A [Gallus gallus]	207	207	100%	1e-51	UG
XP_001196132.1	PREDICTED: putative hemoglobin alpha [Taeniopygia guttata]	206	206	100%	2e-51	UG
XP_356935.3	PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha chain) (Alpha-globin) [Mus musculus] >ref|XP_920048.2| PREDICTED: similar to Hemoglobin subunit alpha (Hemoglobin alpha	196	196	98%	3e-48	UG
XP_001196164.1	PREDICTED: hemoglobin, zeta [Taeniopygia guttata]	191	191	100%	1e-46	UG
NP_001079746.1	hemoglobin, alpha 2 [Xenopus laevis]	185	185	100%	6e-45	UG
XP_002724712.1	PREDICTED: mCG1039741-like [Rattus norvegicus] >ref|XP_002727804.1| PREDICTED: mCG1039741-like [Rattus norvegicus]	184	184	100%	1e-44	UG
NP_001029153.1	hemoglobin, theta T1 [Mus musculus]	184	184	100%	1e-44	UG
NP_001162288.1	hemoglobin, theta 1 [Papio anubis]	184	184	100%	1e-44	UG
NP_001164885.1	hemoglobin subunit zeta [Oryctolagus cuniculus]	184	184	100%	1e-44	UG
XP_001510395.1	PREDICTED: similar to zeta-1-globin [Ornithorhynchus anatinus]	182	182	100%	4e-44	UG
NP_001004375.1	hemoglobin subunit alpha-D [Gallus gallus]	182	182	99%	4e-44	UG
NP_001164886.1	zeta globin [Oryctolagus cuniculus]	182	182	100%	4e-44	UG
NP_001004374.1	hemoglobin subunit pi [Gallus gallus]	182	182	100%	4e-44	UG
NP_005322.1	hemoglobin subunit theta-1 [Homo sapiens]	182	182	100%	4e-44	UG
XP_001915477.1	PREDICTED: similar to Hemoglobin subunit theta-1 (Hemoglobin theta-1 chain) (Theta-1-globin) [Equus caballus]	181	181	100%	5e-44	UG
NP_001164873.1	hemoglobin, zeta [Oryctolagus cuniculus]	181	181	100%	5e-44	UG
NP_001079797.1	hemoglobin alpha 3 subunit [Xenopus laevis]	178	178	100%	6e-43	UG
XP_002196147.1	PREDICTED: putative hemoglobin alpha-D chain [Taeniopygia guttata]	177	177	99%	1e-42	UG
NP_001079749.1	hemoglobin subunit alpha-4 [Xenopus laevis]	177	177	100%	1e-42	UG
NP_001164887.1	theta 1 globin [Oryctolagus cuniculus] >ref|NP_001164972.1| hemoglobin subunit theta-1 [Oryctolagus cuniculus]	176	176	100%	2e-42	UG
NP_001139726.1	hemoglobin, alpha 2 [Xenopus (Silurana) tropicalis]	176	176	100%	2e-42	UG
NP_001108014.1	hemoglobin subunit zeta [Equus caballus]	176	176	100%	6e-43	UG
NP_005323.1	hemoglobin subunit zeta [Homo sapiens]	176	176	100%	6e-43	UG
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%	7e-43	UG
XP_001372225.1	PREDICTED: similar to pol polyprotein [Monodelphis domestica]	348	348	100%	2e-42	UG
NP_034535.1	hemoglobin subunit zeta [Mus musculus]	173	173	100%	4e-42	UG
NP_001135556.1	hypothetical protein LOC100216102 [Xenopus (Silurana) tropicalis]	172	172	100%	8e-42	UG
NP_001015904.1	hemoglobin alpha 3 subunit [Xenopus (Silurana) tropicalis]	172	172	100%	1e-41	UG
XP_001788743.1	PREDICTED: similar to Hemoglobin subunit zeta (Hemoglobin zeta chain) (Zeta-globin) [Bos taurus]	171	171	100%	1e-41	UG
NP_001073802.1	hemoglobin, theta 1 [Bos taurus]	171	171	96%	2e-41	UG
NP_001081493.1	hemoglobin, alpha 1 [Xenopus laevis]	169	169	100%	5e-41	UG
NP_988860.1	hemoglobin, alpha 1 [Xenopus (Silurana) tropicalis]	168	168	100%	4e-40	UG
XP_580707.3	PREDICTED: similar to Hemoglobin subunit zeta (Hemoglobin zeta chain)	165	165	98%	1e-39	UG
NP_001165373.1	alpha globin larval-3 [Xenopus (Silurana) tropicalis]	165	165	100%	1e-39	G
XP_001061675.2	PREDICTED: hemoglobin, theta 1 [Rattus norvegicus] >ref|XP_347267.4|	164	164	100%	2e-39	UG
XP_001151282.1	PREDICTED: similar to zeta-1-globin isoform 1 [Pan troglodytes] >ref|XP_	163	163	100%	3e-39	UG
NP_001165374.1	alpha globin larval-4 [Xenopus (Silurana) tropicalis]	163	163	100%	3e-39	UG
NP_001016009.1	alpha globin larval-5 [Xenopus (Silurana) tropicalis]	163	163	100%	3e-39	UG
NP_001005092.1	hemoglobin, zeta [Xenopus (Silurana) tropicalis]	162	162	100%	5e-39	UG
NP_001118023.1	alpha-globin IV [Oncorhynchus mykiss]	158	158	100%	6e-38	UG
NP_001107321.1	alpha globin larval-2 [Xenopus (Silurana) tropicalis]	155	155	100%	5e-37	UG
NP_001089702.1	hypothetical protein LOC734765 [Xenopus laevis]	154	154	100%	1e-36	UG
XP_002723955.1	PREDICTED: zeta globin-like [Oryctolagus cuniculus]	153	153	100%	2e-36	UG
NP_571332.2	hemoglobin alpha adult-1 [Danio rerio]	153	153	100%	2e-36	UG
XP_001334294.3	hypothetical protein LOC497166 [Danio rerio]	152	152	100%	3e-36	UG
NP_001028265.1	alpha globin-like [Danio rerio]	152	152	100%	3e-36	UG
XP_002723956.1	PREDICTED: zeta globin-like [Oryctolagus cuniculus]	151	151	100%	3e-36	UG
NP_778165.1	hemoglobin, theta T2 [Mus musculus]	151	151	88%	2e-35	UG
NP_001117134.1	hemoglobin subunit alpha [Salmo salar]	150	150	100%	3e-35	UG





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?

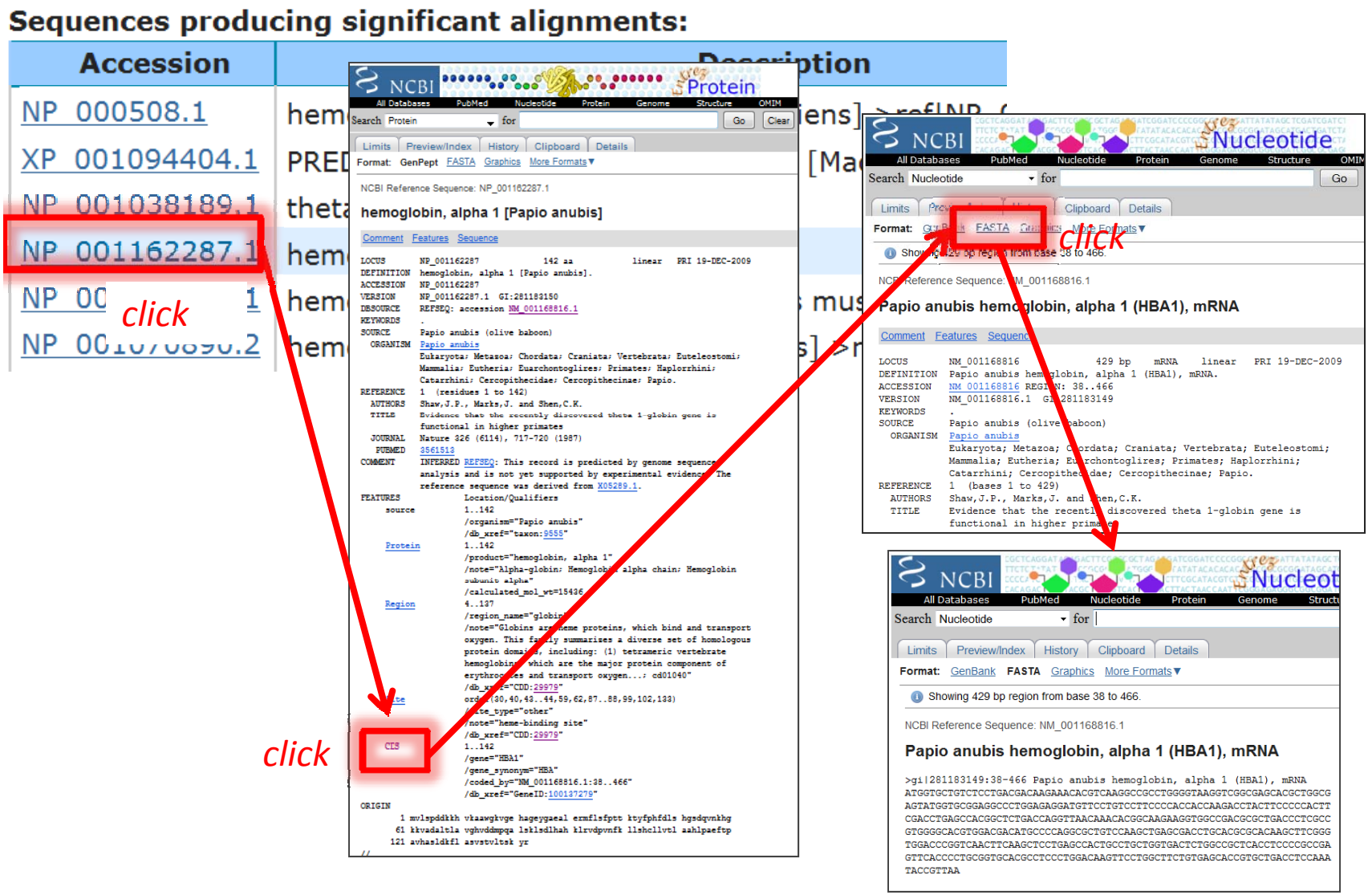
  Amino acid

  – 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)

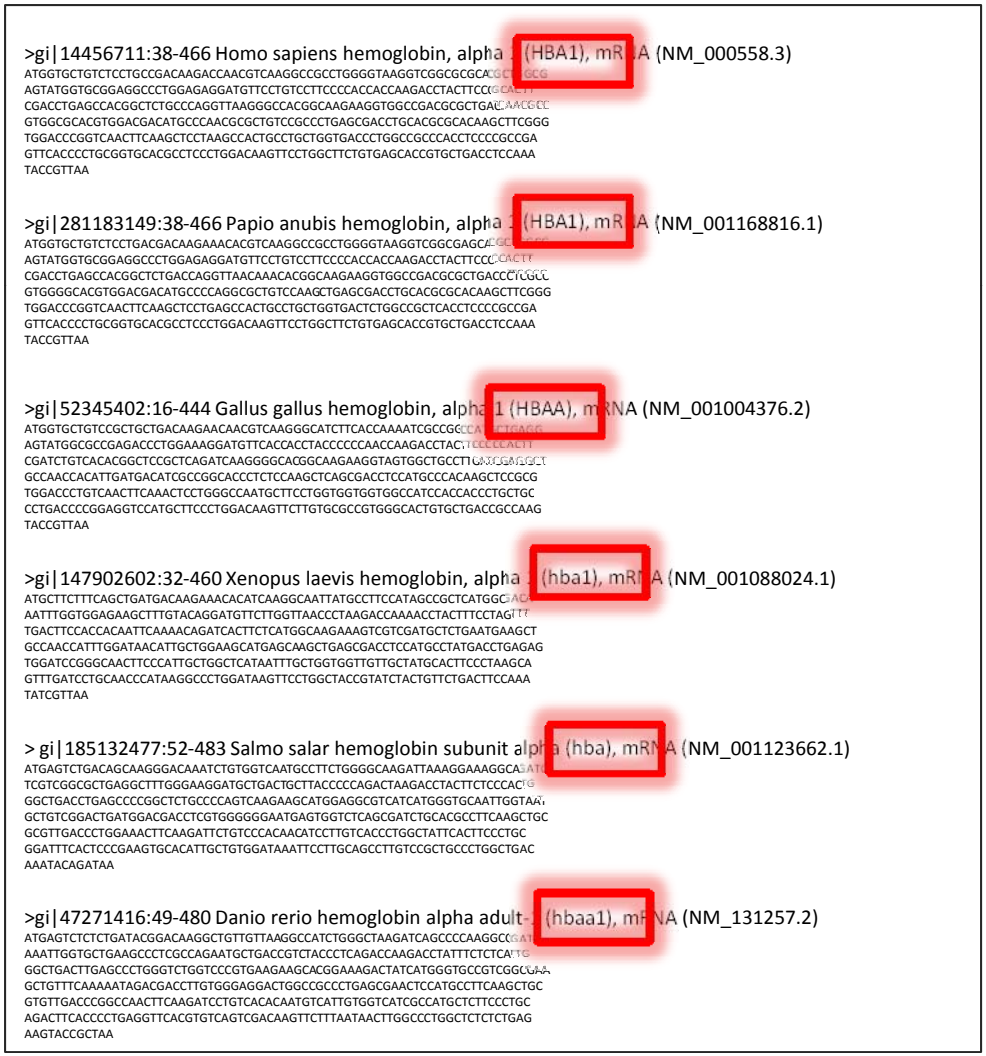
How to obtain nt sequence?



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)
ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCA
AGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCC
CGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGAC
GTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGG
TGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGA
GTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAA
TACCGTTAA

>gi|281183149:38-466 Papio anubis hemoglobin, alpha 1 (HBA1), mRNA (NM_001168816.1)
ATGGTGCTGTCTCCTGACGACAAGAAACACGTCAAGGCCGCCTGGGGTAAGGTCGGCGAGCA
AGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCC
CGACCTGAGCCACGGCTCTGACCAGGTTAACAAACACGGCAAGAAGGTGGCCGACGCGCTGACCC
GTGGGGCACGTGGACGACATGCCCCAGGCGCTGTCCAAGCTGAGCGACCTGCACGCGCACAAGCTTCGGG
TGGACCCGGTCAACTTCAAGCTCCTGAGCCACTGCCTGCTGGTGACTCTGGCCGCTCACCTCCCCGCCGA
GTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAA
TACCGTTAA

>gi|52345402:16-444 Gallus gallus hemoglobin, alpha 1 (HBAA), mRNA (NM_001004376.2)
ATGGTGCTGTCCGCTGCTGACAAGAACAACGTCAAGGGCATCTTCACCAAAATCGCCGGCCA
AGTATGGCGCCGAGACCCTGGAAAGGATGTTCACCACCTACCCCCAACCAAGACCTAC
CGATCTGTCACACGGCTCCGCTCAGATCAAGGGGCACGGCAAGAAGGTAGTGGCTGCCTT
GCCAACCACATTGATGACATCGCCGGCACCCTCTCCAAGCTCAGCGACCTCCATGCCCACAAGCTCCGCG
TGGACCCTGTCAACTTCAAACTCCTGGGCCAATGCTTCCTGGTGGTGGTGGCCATCCACCACCCTGCTGC
CCTGACCCCGGAGGTCCATGCTTCCCTGGACAAGTTCTTGTGCGCCGTGGGCACTGTGCTGACCGCCAAG
TACCGTTAA

>gi|147902602:32-460 Xenopus laevis hemoglobin, alpha 1 (hba1), mRNA (NM_001088024.1)
ATGCTTCTTTCAGCTGATGACAAGAAACACATCAAGGCAATTATGCCTTCCATAGCCGCTCATGGC
AATTTGGTGGAGAAGCTTTGTACAGGATGTTCTTGGTTAACCCTAAGACCAAAACCTACTTTCCTAG
TGACTTCCACCACAATTCAAAACAGATCACTTCTCATGGCAAGAAAGTCGTCGATGCTCTGAATGAAGCT
GCCAACCATTTGGATAACATTGCTGGAAGCATGAGCAAGCTGAGCGACCTCCATGCCTATGACCTGAGAG
TGGATCCGGGCAACTTCCCATTGCTGGCTCATAATTTGCTGGTGGTTGTTGCTATGCACTTCCCTAAGCA
GTTTGATCCTGCAACCCATAAGGCCCTGGATAAGTTCCTGGCTACCGTATCTACTGTTCTGACTTCCAAA
TATCGTTAA

> gi|185132477:52-483 Salmo salar hemoglobin subunit alpha (hba), mRNA (NM_001123662.1)
ATGAGTCTGACAGCAAGGGACAAATCTGTGGTCAATGCCTTCTGGGGCAAGATTAAAGGAAAGGCA
TCGTCGGCGCTGAGGCTTTGGGAAGGATGCTGACTGCTTACCCCCAGACTAAGACCTACTTCTCCCAC
GGCTGACCTGAGCCCCGGCTCTGCCCCAGTCAAGAAGCATGGAGGCGTCATCATGGGTGCAATTGGTAA
GCTGTCGGACTGATGGACGACCTCGTGGGGGGAATGAGTGGTCTCAGCGATCTGCACGCCTTCAAGCTGC
GCGTTGACCCTGGAAACTTCAAGATTCTGTCCCACAACATCCTTGTCACCCTGGCTATTCACTTCCCTGC
GGATTTCACTCCCGAAGTGCACATTGCTGTGGATAAATTCCTTGCAGCCTTGTCCGCTGCCCTGGCTGAC
AAATACAGATAA

>gi|47271416:49-480 Danio rerio hemoglobin alpha adult-1 (hbaa1), mRNA (NM_131257.2)
ATGAGTCTCTCTGATACGGACAAGGCTGTTGTTAAGGCCATCTGGGCTAAGATCAGCCCCAAGGCC
AAATTGGTGCTGAAGCCCTCGCCAGAATGCTGACCGTCTACCCTCAGACCAAGACCTATTTCTCTCA
GGCTGACTTGAGCCCTGGGTCTGGTCCCGTGAAGAAGCACGGAAAGACTATCATGGGTGCCGTCGGCGAA
GCTGTTTCAAAAATAGACGACCTTGTGGGAGGACTGGCCGCCCTGAGCGAACTCCATGCCTTCAAGCTGC
GTGTTGACCCGGCCAACTTCAAGATCCTGTCACACAATGTCATTGTGGTCATCGCCATGCTCTTCCCTGC
AGACTTCACCCCTGAGGTTCACGTGTCAGTCGACAAGTTCTTTAATAACTTGGCCCTGGCTCTCTCTGAG
AAGTACCGCTAA

Bioinformatics

Bi1X-2010

Part III: Multiple alignment and phylogenetic analysis

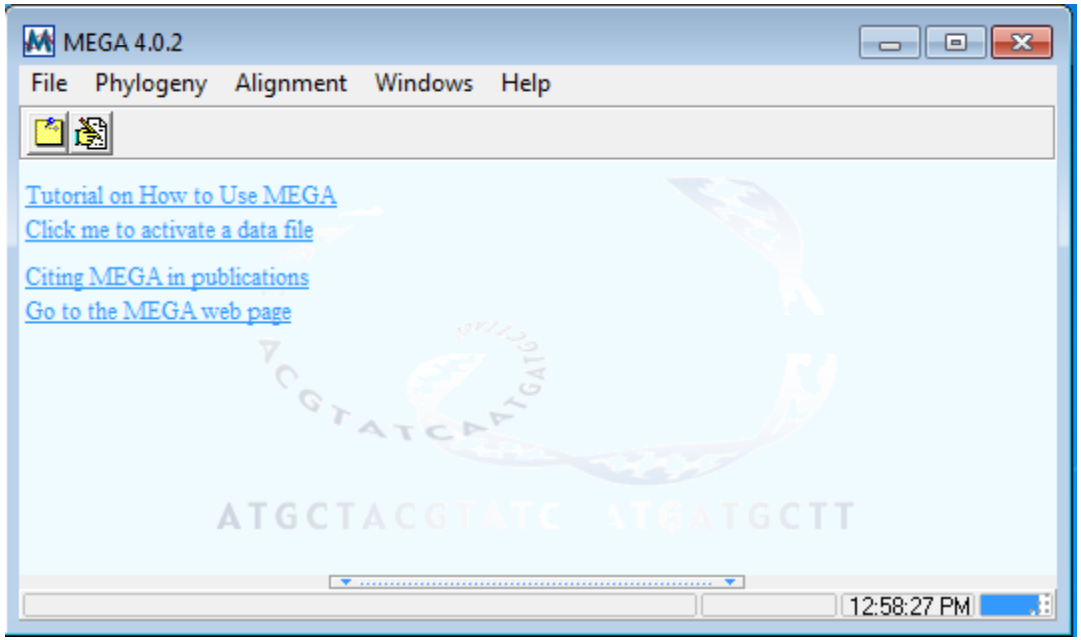
Arbel Tadmor

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

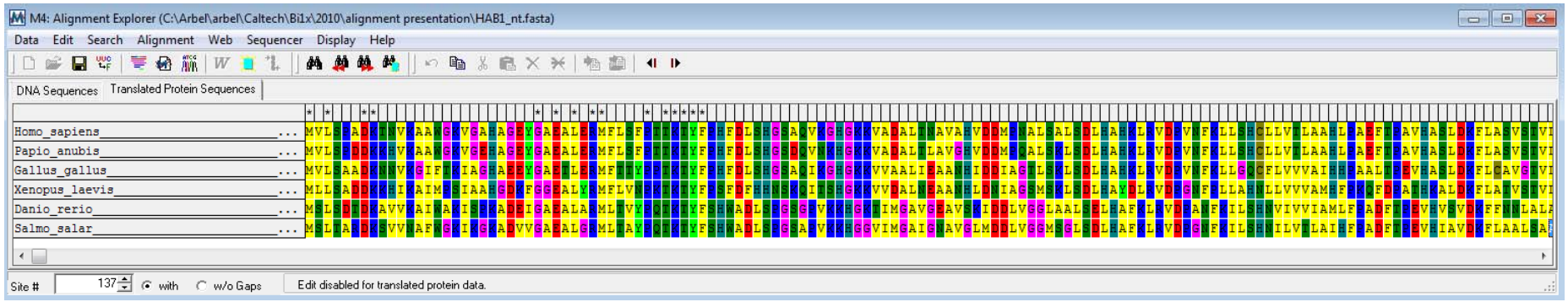
- Google MEGA4 and install
- Launch MEGA4
- Drag the FASTA file into MEGA4



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

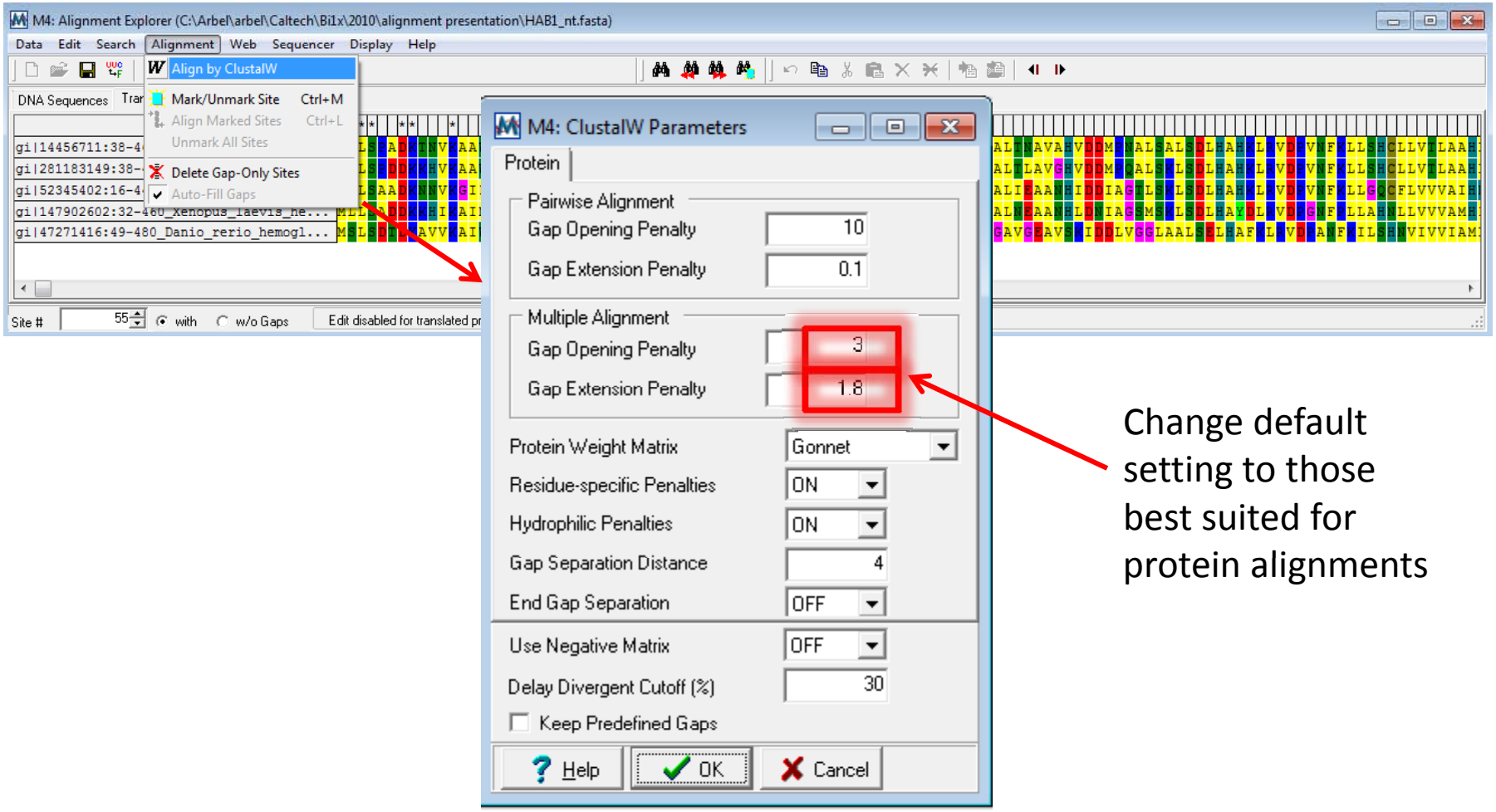
Read more at http://www.ebi.ac.uk/2can/tutorials/nucleotide/clustalw.html

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



Change default setting to those best suited for protein alignments

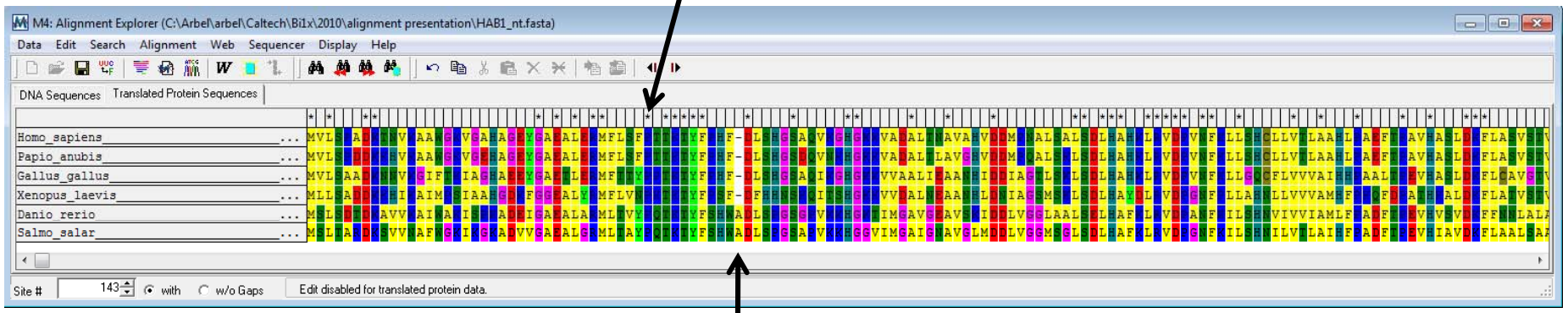
Manual inspection of alignment

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

Alignment using ClustalW

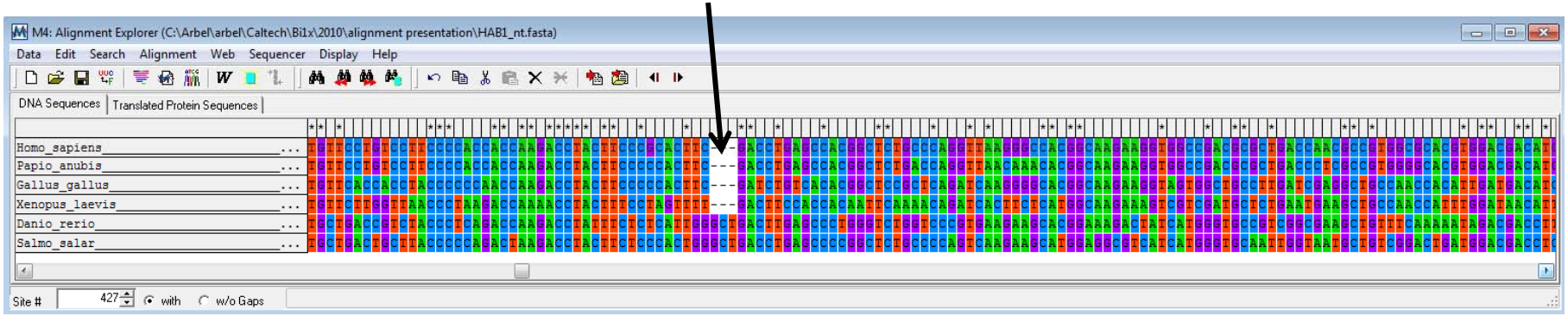
aa view

Star indicates conserved character

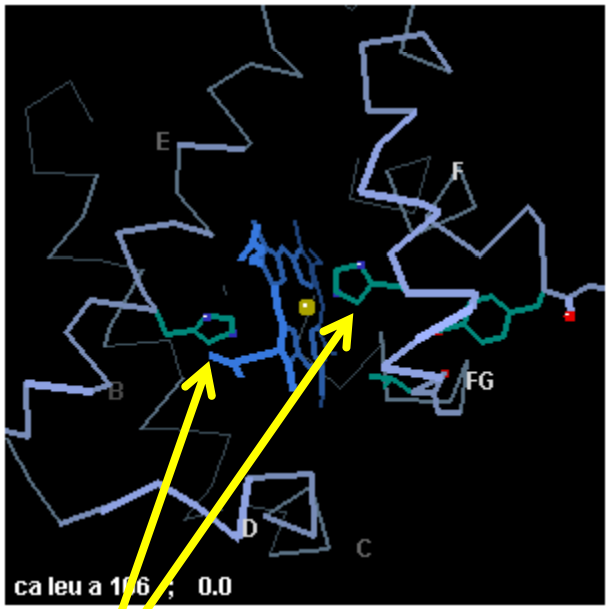


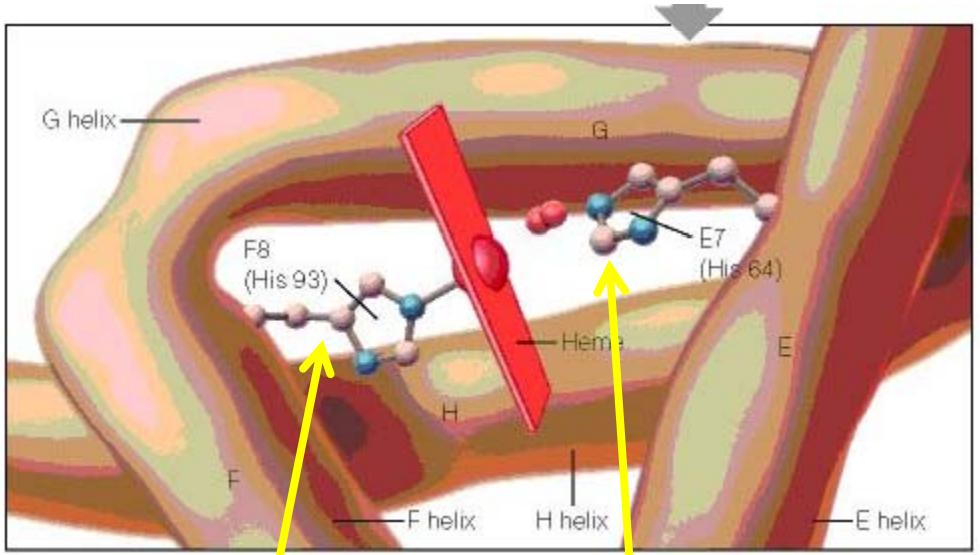
nt view

Deletion/insertion event



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



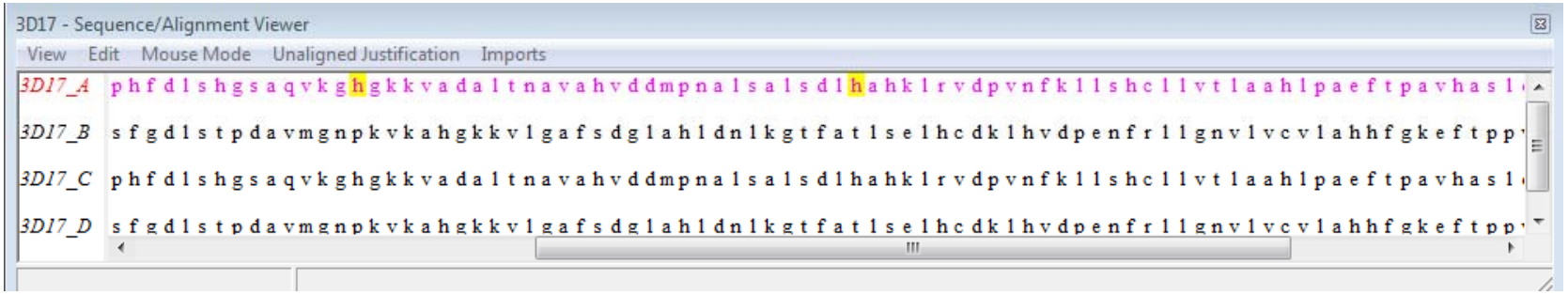


Mathews et al. 2000. Biochemistry 3 rd edition

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 Fe2+) and superoxide (bound to Fe3+). Oxygen must be released in the former because the later 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.

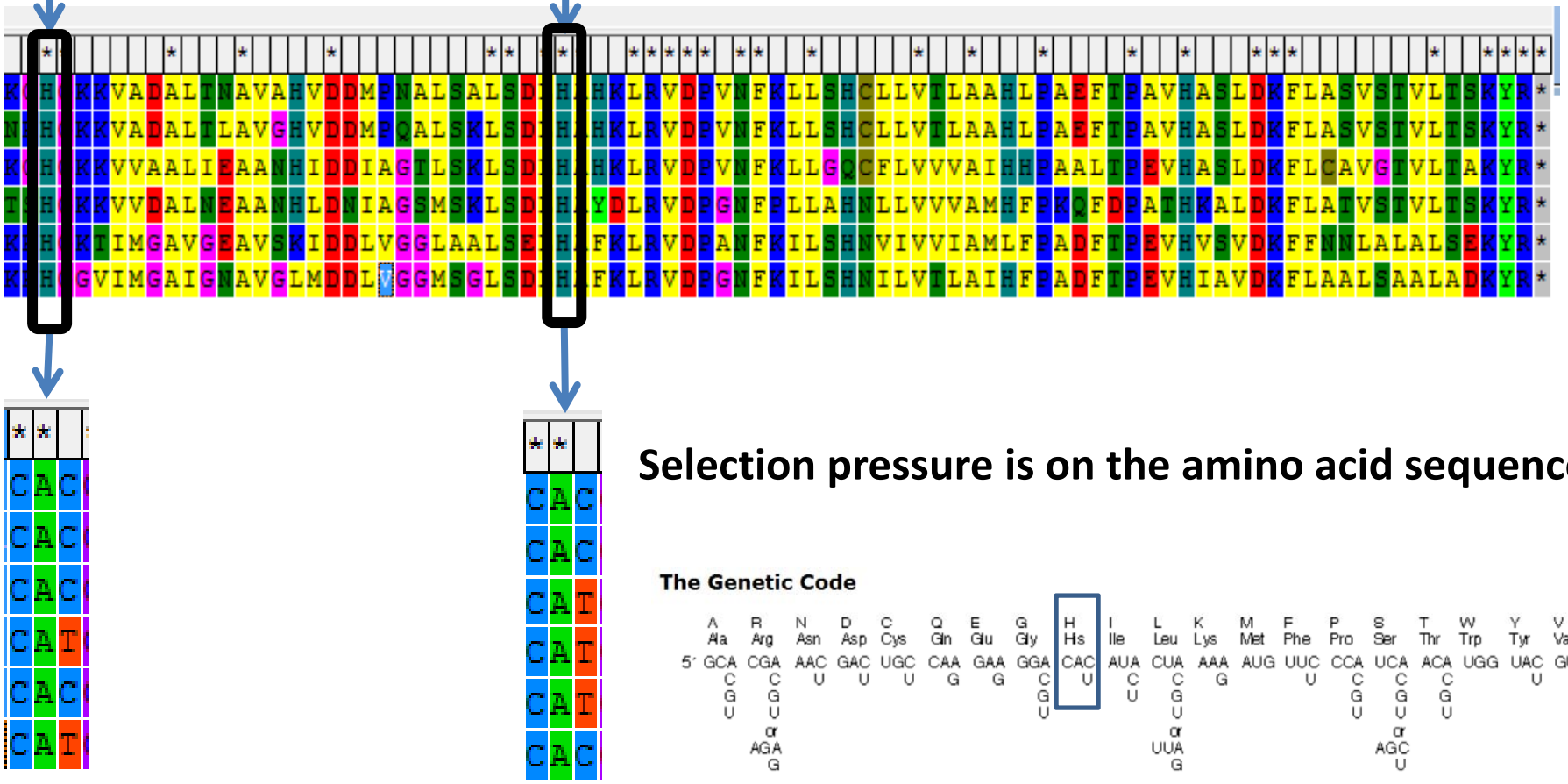
Cn3D demo of Homo sapiens hemoglobin subunit α



Histidines are conserved

Conserved His58

Conserved His87



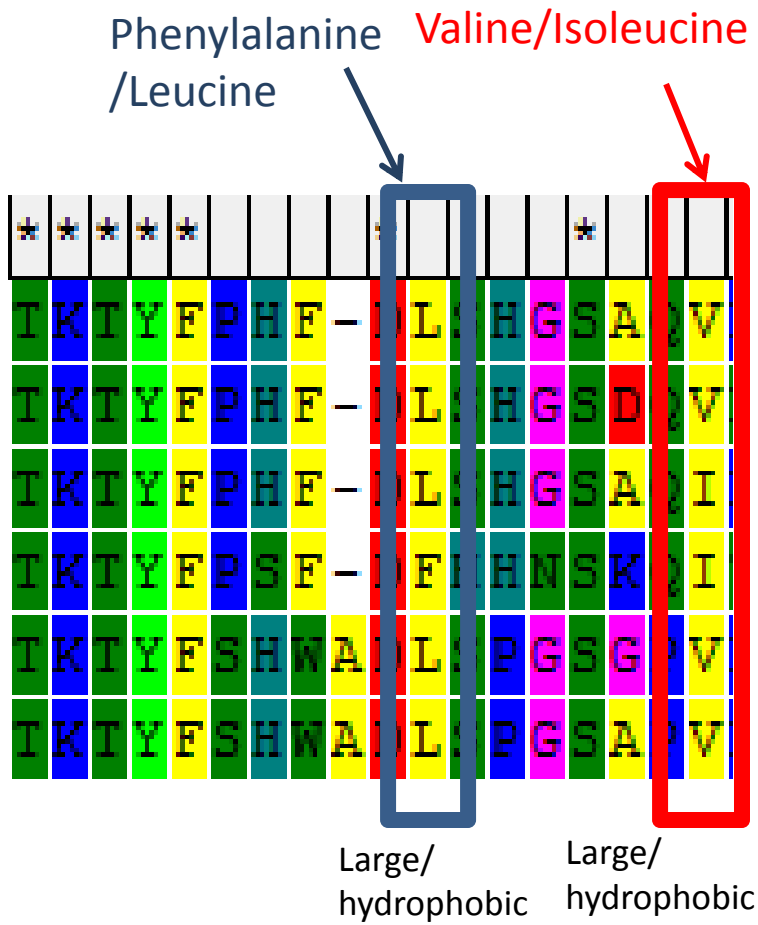
Selection pressure is on the amino acid sequence

Some residues mutate to chemically similar residues

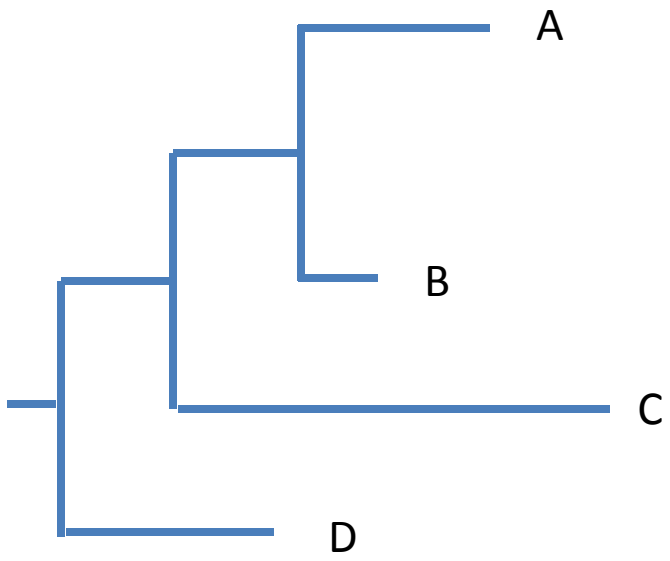
nonpolar polar basic acidic (stop codon)

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

1st base	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
	CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine
A	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
	AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
	AUG[A] (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
G	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine
	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



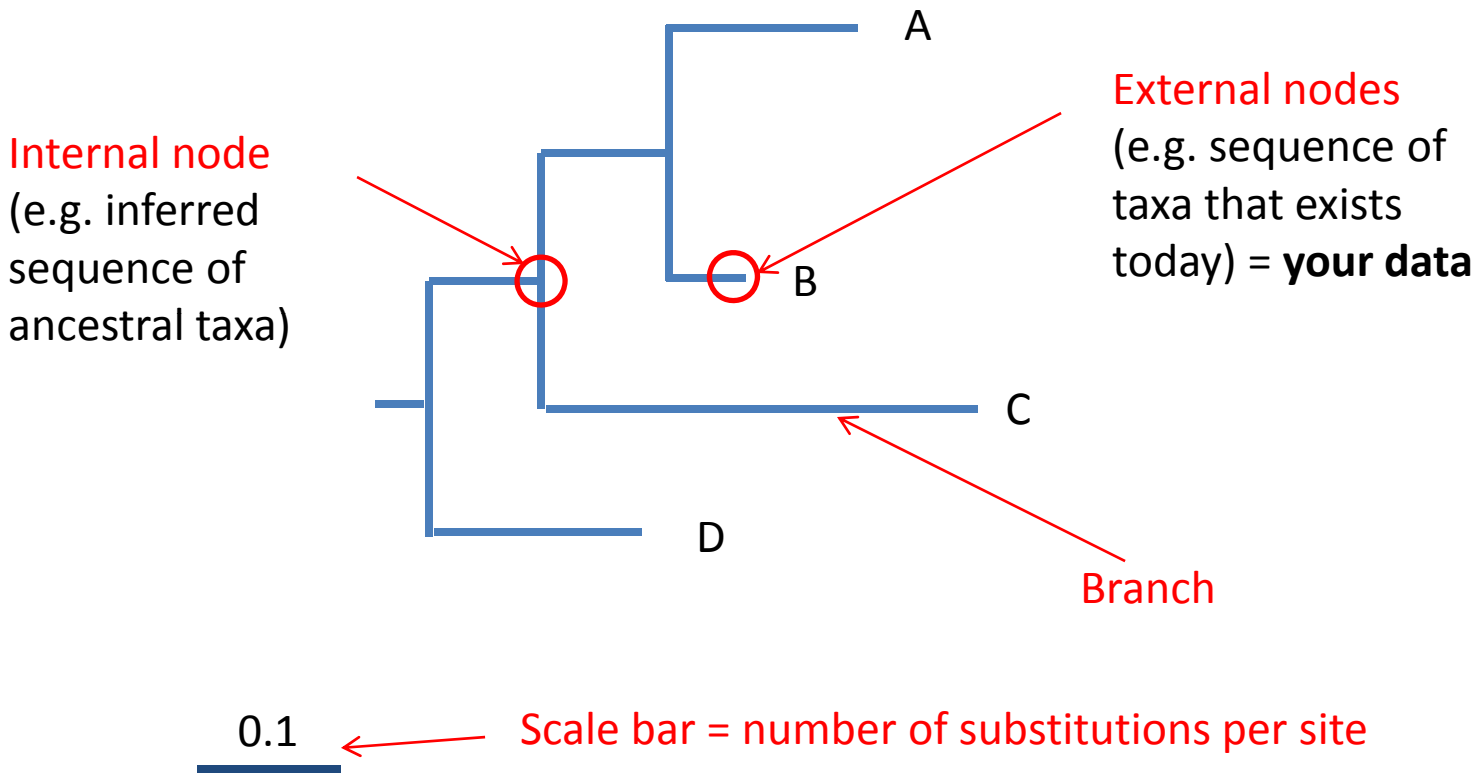
Phylogenetic analysis



What are trees good for?

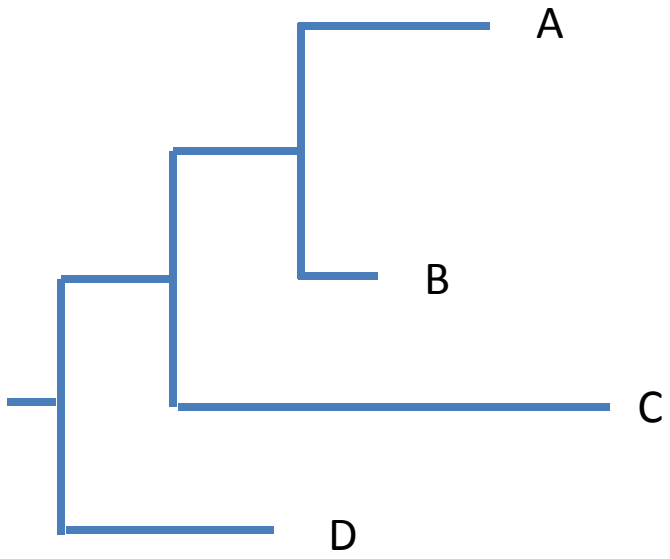
- Identify close relatives
- Determine evolutionary relationship between sequences

Some tree terminology



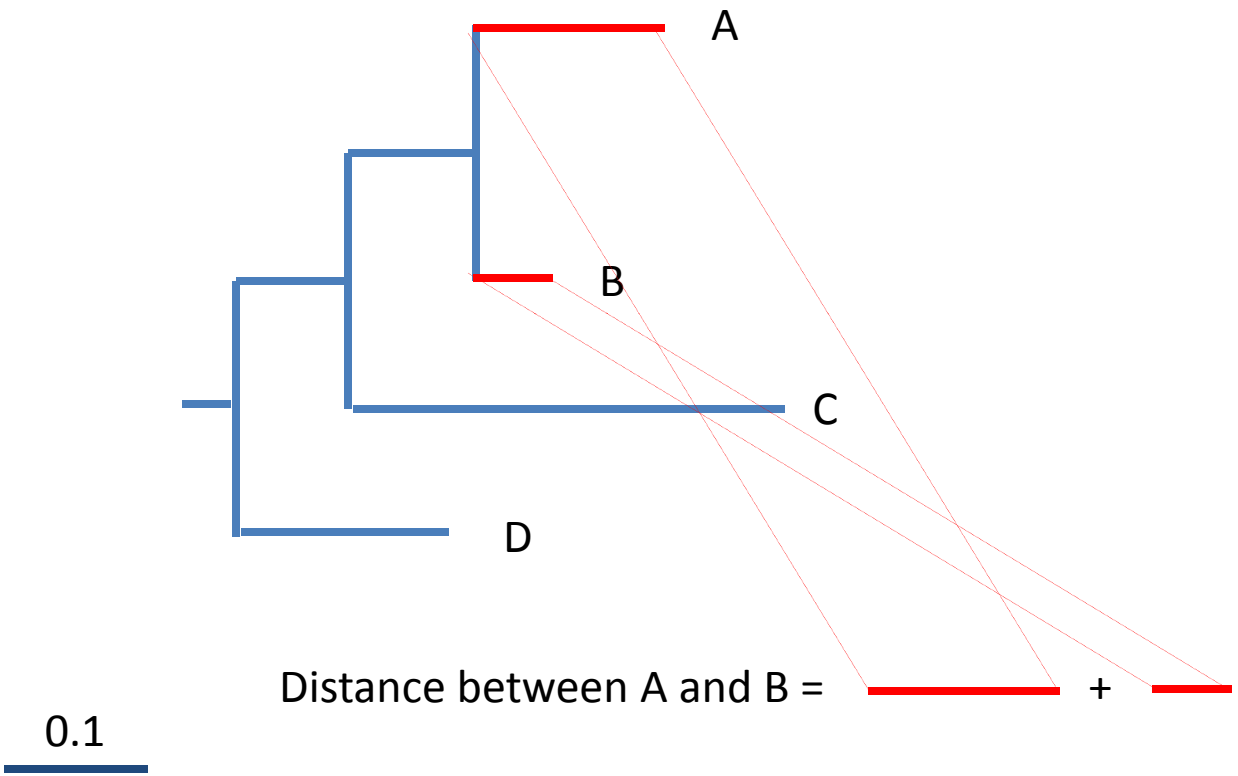
We will discuss only **bifurcating** trees: each node has only two immediate descendant lineages, i.e. we assume evolutionary speciation is a binary process

Trees have two elements: branch lengths and branching order



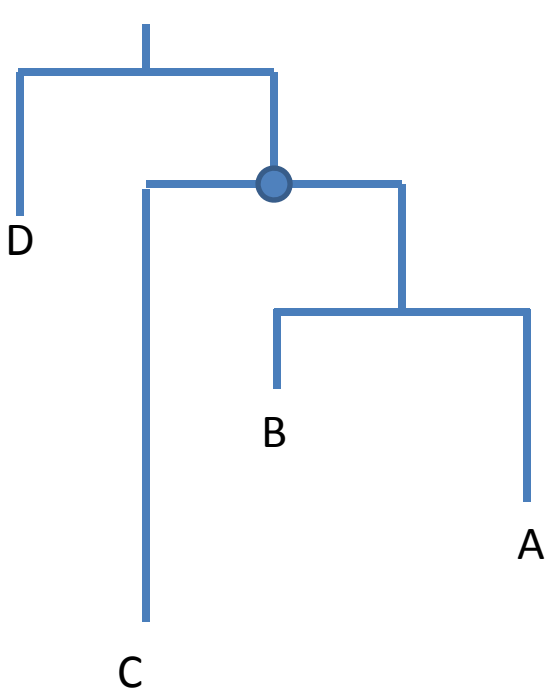
- **Topology** = branching order
- To estimate a tree you need to estimate
  - Branch lengths (simple problem)
  - Branching order (difficult for many seq)

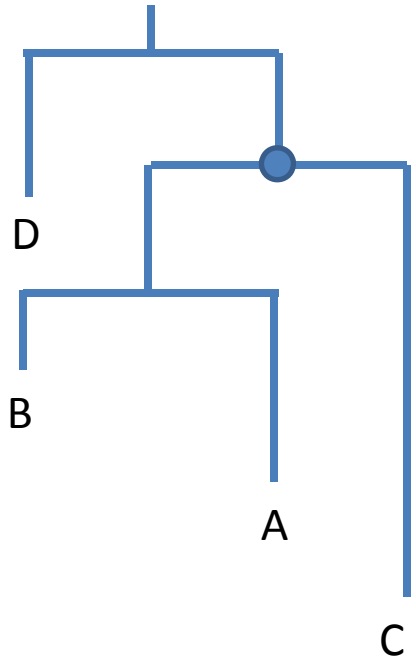
Branch lengths



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

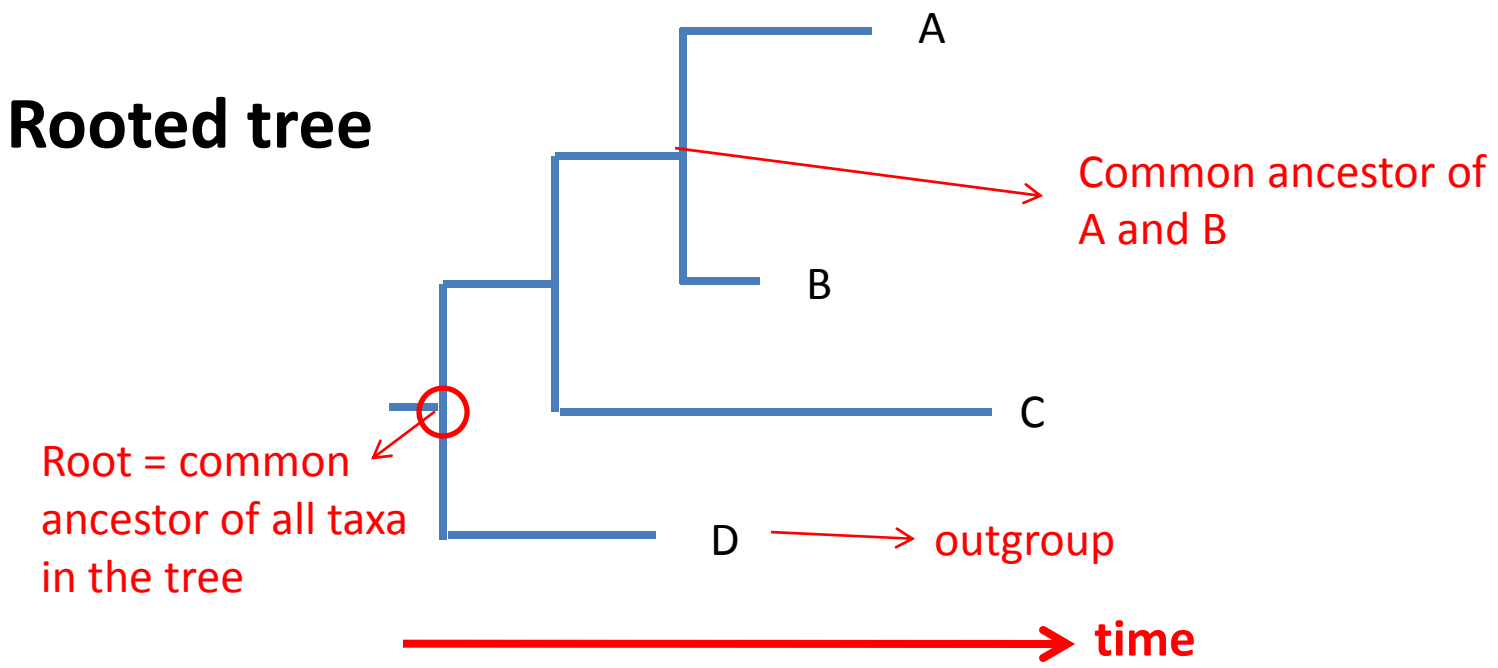
Branch order:
note that trees are like mobiles…





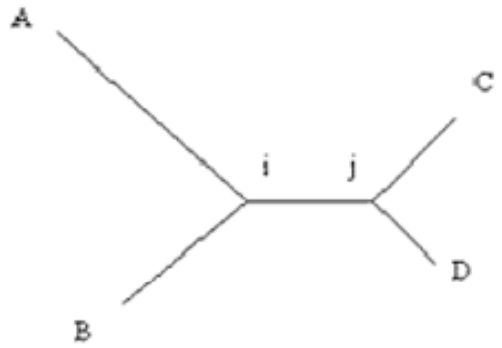
These two trees are equivalent in every way

Rooted trees

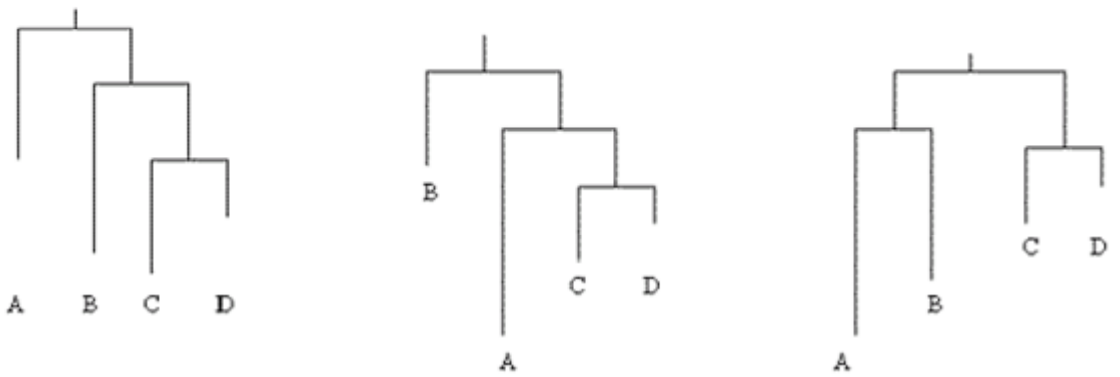


- **If** a tree it rooted, root = left most internal node
- By selecting a node to be a **root** you set a time arrow
- **The more recently species share a common ancestor, the more closely related they are** (e.g. A and C are more closely related than A and D)
- Which node is the root? You "break the symmetry" by adding additional information: e.g. you know D is more is 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
1. Multiple alignment 2. Calculate **distance matrix** = matrix of evolutionary distances between all pairs of aligned sequences 3. Calculate tree topology	1. Multiple alignment 2. Compare characters at each column in the alignment and give each topology a score 3. Choose the topology with the best score
Example: Neighbor Joining (many others)	Examples: • Maximum Likelihood methods • Maximum Parsimony methods • Bayesian methods

Measuring evolutionary distance between two sequences

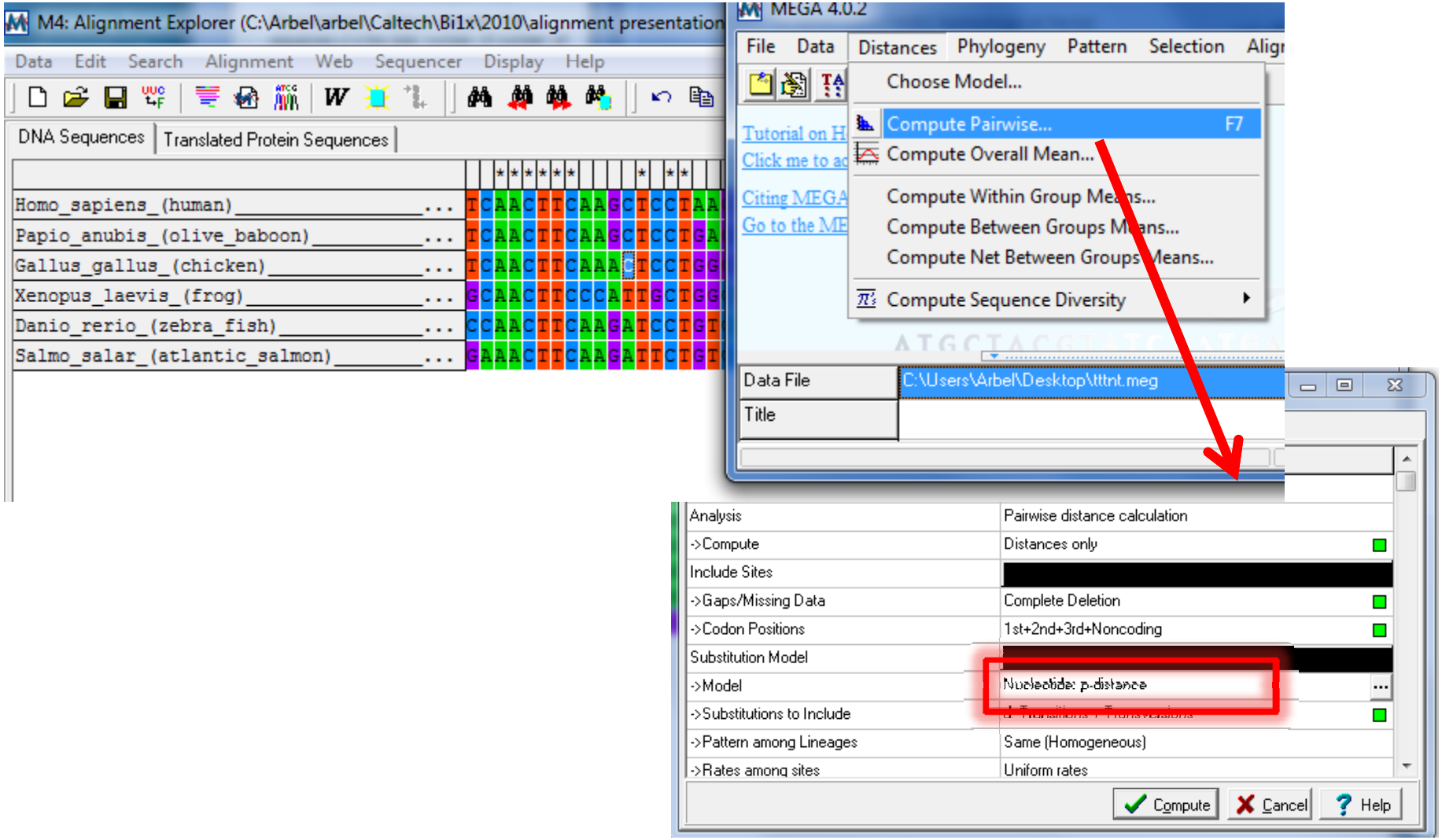
- The evolutionary distance between two sequences *d* is the total number of number of aa/nt substitutions per site between the two sequences
- ***d=2rt***

  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



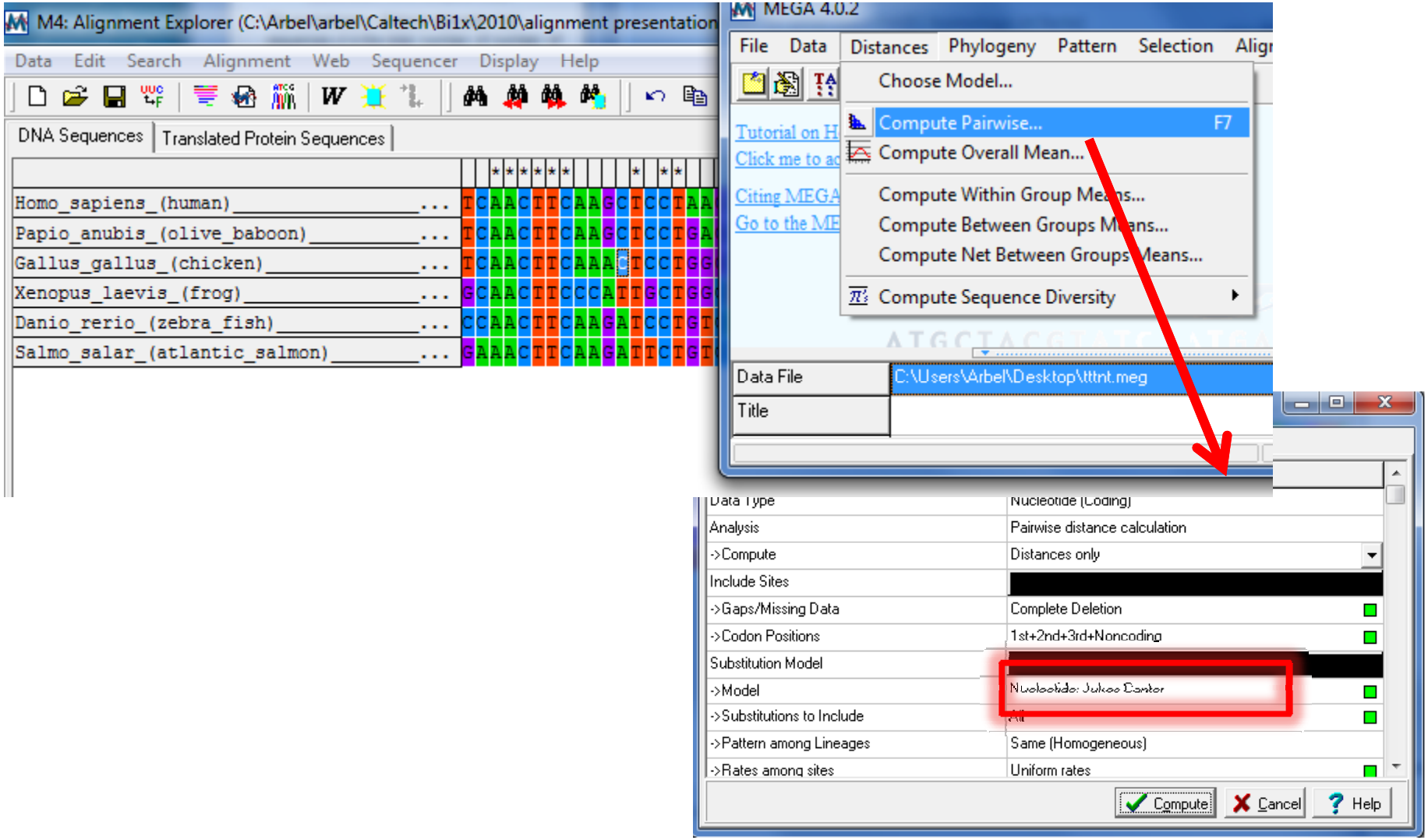
Calculating distances in MEGA4: **p distance**

p distance = 100 - 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



Calculating distances in MEGA4: Evolutionary distance

JC method

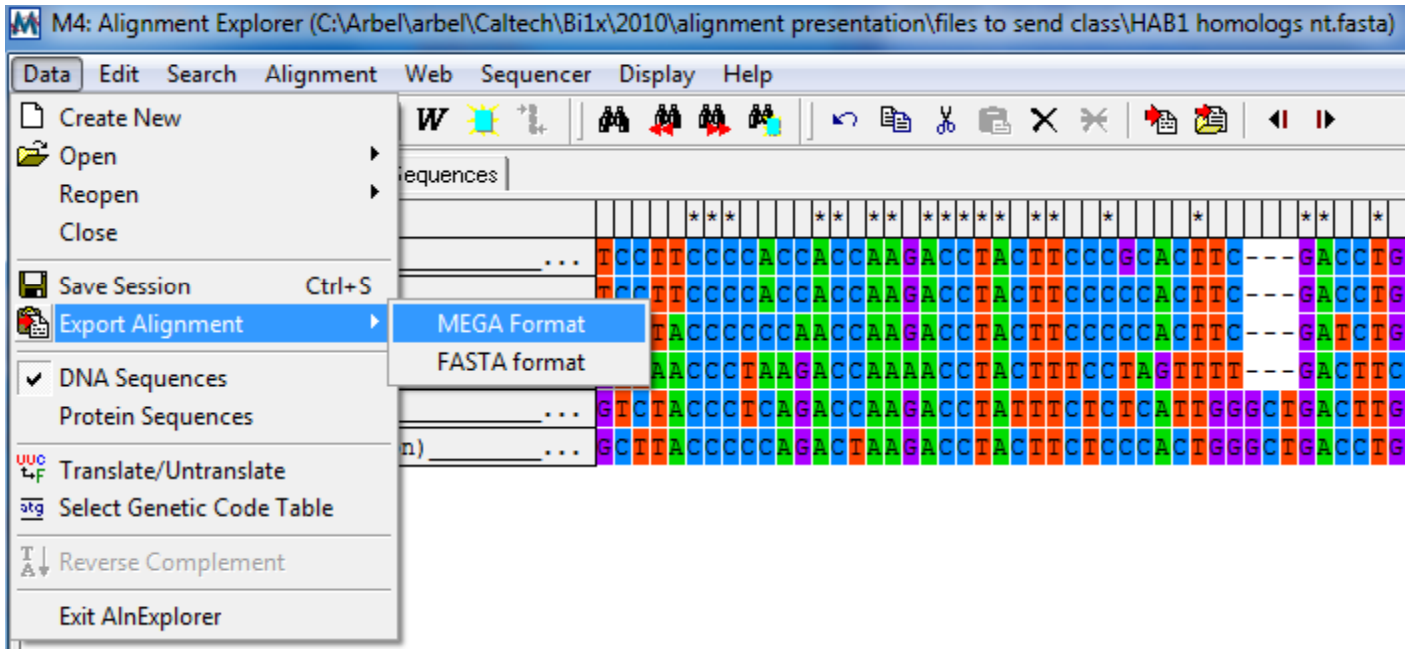
	1	2	3	4	5	6
1. Homo sapiens (human) ...						
2. Papio anubis (olive baboon) ...	0.0531					
3. Gallus gallus (chicken) ...	0.3173	0.3317				
4. Xenopus laevis (frog) ...	0.5837	0.5441	0.5020			
5. Danio rerio (zebra fish) ...	0.5393	0.5736	0.5251	0.6472		
6. Salmo salar (atlantic salmon) ...	0.5637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

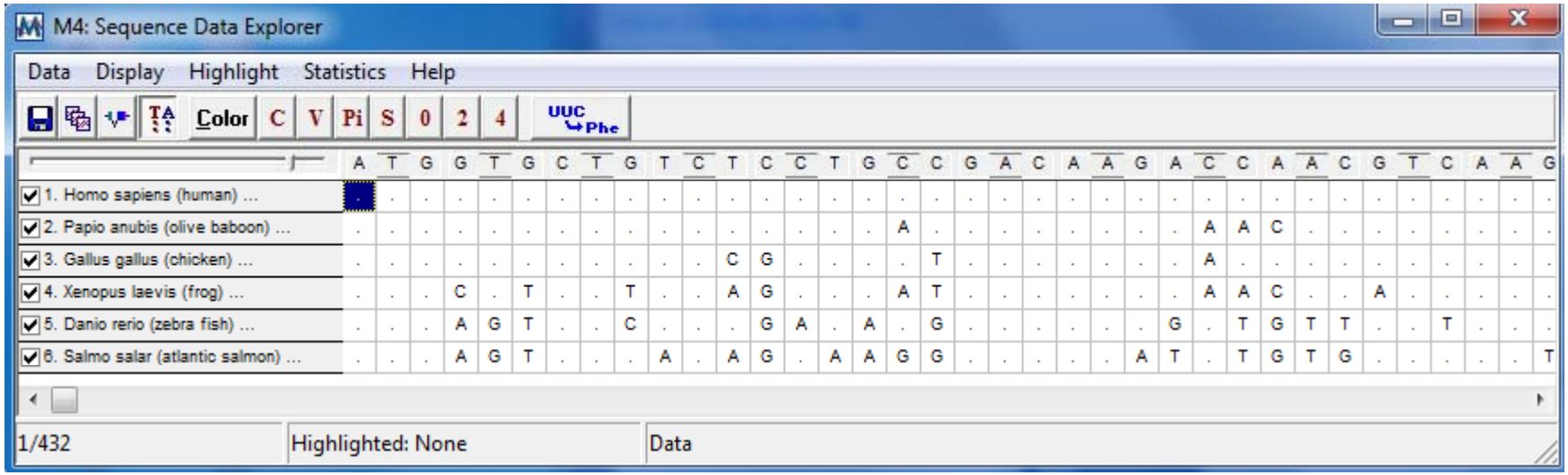
Building a Neighbor joining tree

Export **nt** alignment in MEGA format



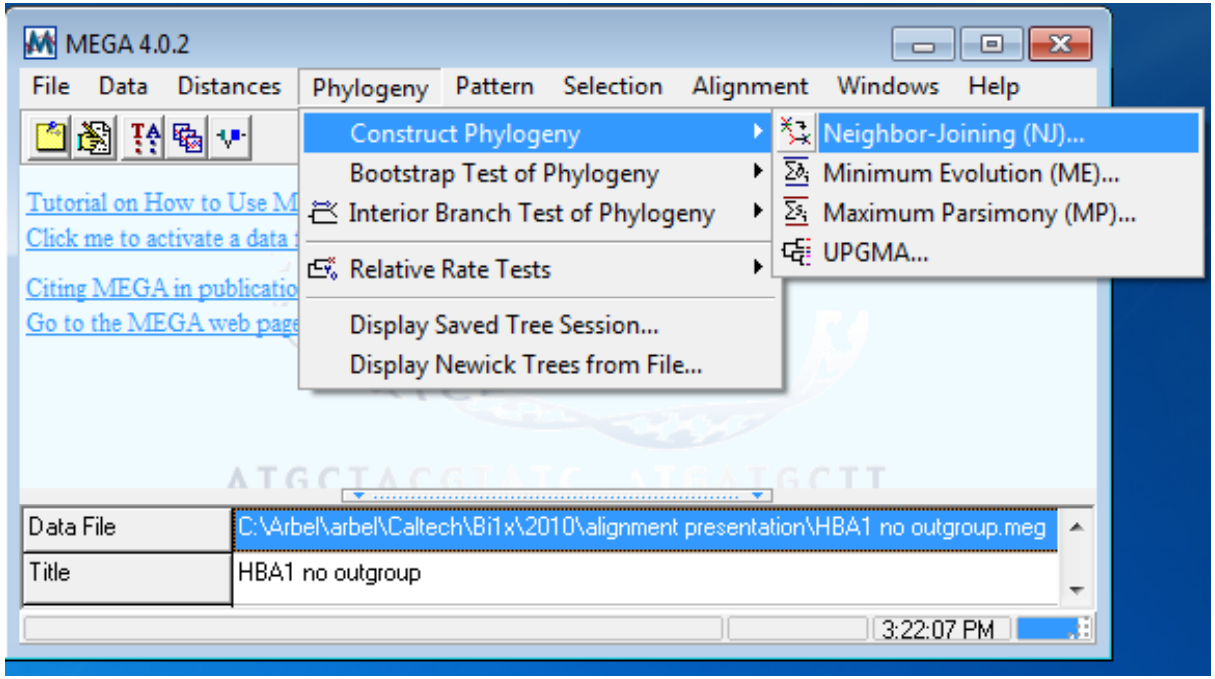
Building a Neighbor joining tree

Close MEGA4 and open your MEGA file.
You should get this window with a **nt** sequence:



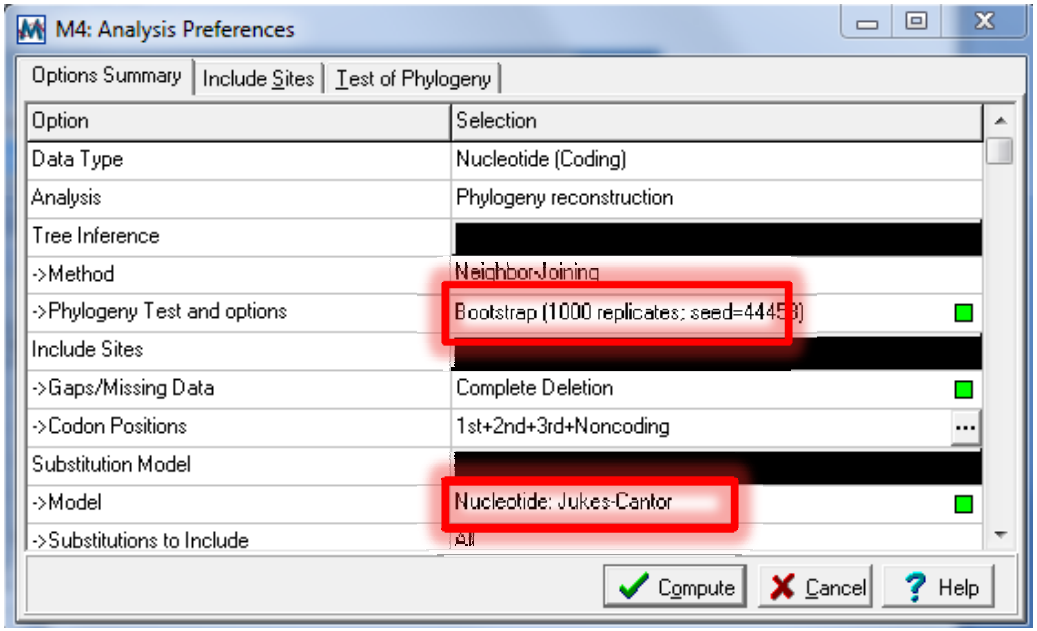
Building a Neighbor joining tree

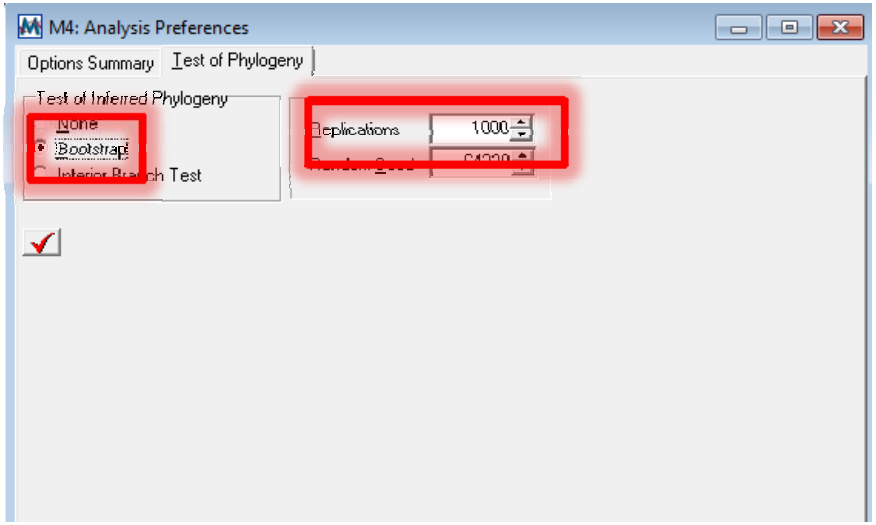
Now let's build a NJ tree



Building a Neighbor joining tree

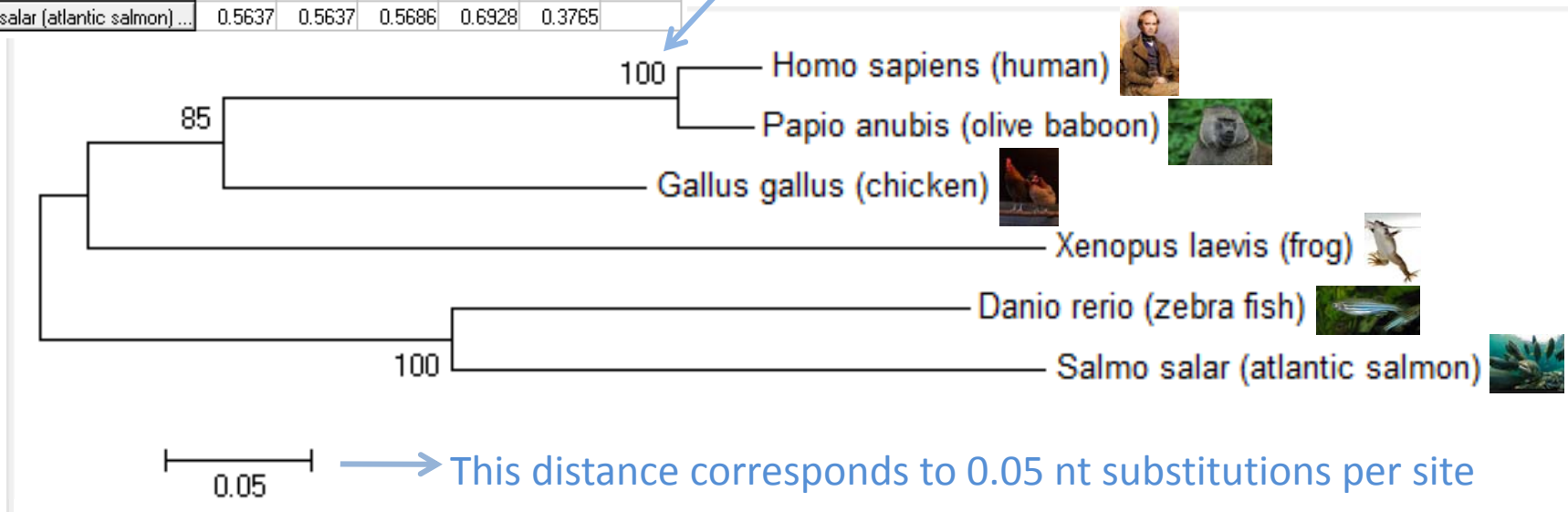
- Use JC nt model
- Calculate bootstrap support with 1000 replications





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	



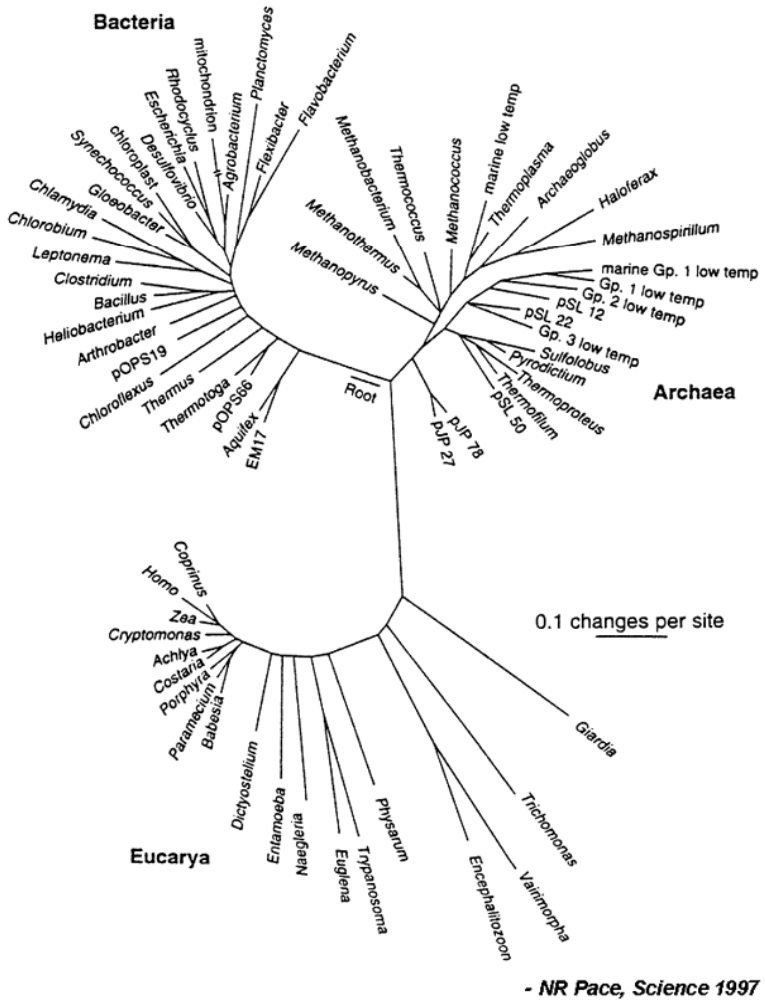
- Human and primates group together
- Fish group together

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



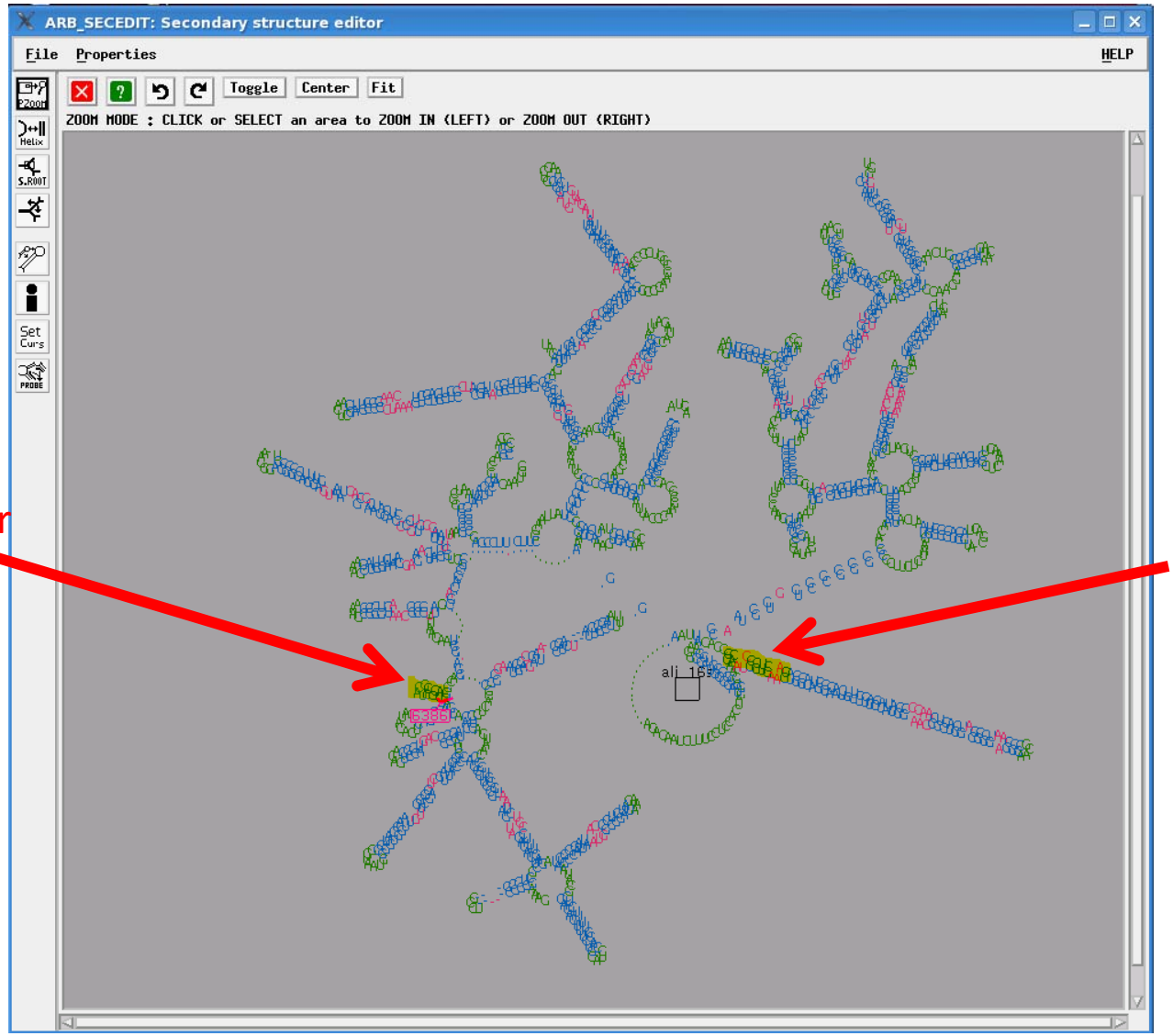
How are SSU rRNA sequences different?

- 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:

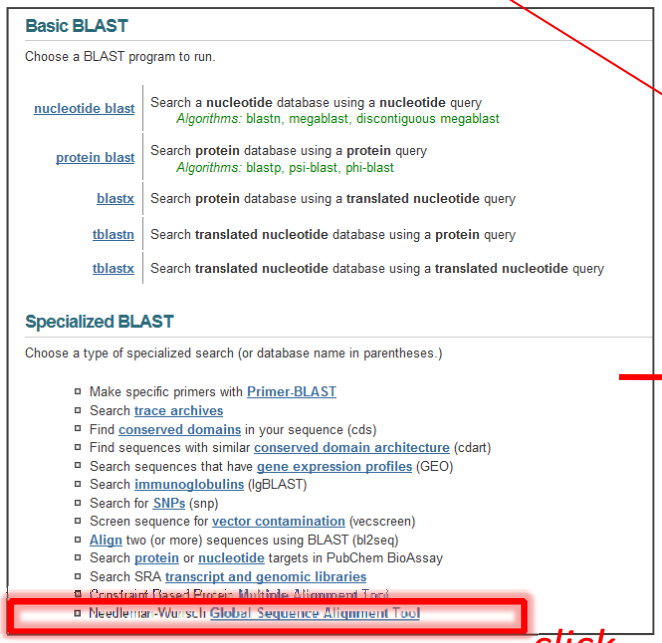
```
>> rand_int = floor(4*rand([1,300])+1);
>> rand_nt  = int2nt(rand_int)

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

rand~U(0,1)

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

>Random seq 1
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGCCTCGGTCCGATATGATTGCCGCATT
TTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCTGT

>Random seq 2
CCGTAACGTCGTCACAGAAGGCGTGCGAGCACCAATGTCTATCATCGGTCACTTGTGTTAGGTGTACCAAAGCTG
AGAGTCGTCTATCTTATCTTCTAATAGTACCTCTATTAAGATTGAGTTGTTGACCCTACAAGCAAATCGTCGTCGCTC
CTCAAACTTGTCGCTCTATCAACTCTAGAATGTTGTCTAAGCCGACAGGACCGAACGGTCAATGTGGCGTTCACG
ATTCAGGCATTATACAAGGCCAATCGTGCGGATGCGGCAGGGGCCCTTCTAACGAAGCGGGGTGCTGGATAT



click

```
>lcl|58213
Length=300

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

Query  1    TCGAAGGCGCTCGGTAGAGTACGTGTCCCA--ACTGTTGCCTAAGCGCGCGTACAGTAGG  58
            ||| |  || |||  | || | |  || | |   ||   || |||    || | || | |
Sbjct  1    CCGTAA-CG-TCGTCACAGAAGGCGTGCGAGCACCAATGTCTATCATCG-GT-CACTTGT  56

Query  59   GCGAGGCACGCTACTGTTACGAGATTC--CTACCGAAG-----AAAAGTTAAGCCCCTCG  111
            |  |||     | |  | |  |||| ||  ||| |  |       | |||    || ||
Sbjct  57   GTTAGGTGTACCAAAGCT--GAGAGTCGTCTATCTTATCTTCTAATAGTA---CCTCTAT  111

Query  112  AAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATACTACGGGCCTTCCCCCAACATA  171
             |||       || | ||    ||| |   || | ||  |  ||    || || | ||| |
Sbjct  112  TAAG------ATTG-AGTTGTTGACCCTACAAGCAAATCGTCGTCGCT-CCTCAAACTTG  163

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

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

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

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

>Random seq 1
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGCCTCGGTCCGATATGATTGCCGCATT
TTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCTGT

>Random seq 2
CCGTAACGTCGTCACAGAAGGCGTGCGAGCACCAATGTCTATCATCGGTCACTTGTGTTAGGTGTACCAAAGCTG
AGAGTCGTCTATCTTATCTTCTAATAGTACCTCTATTAAGATTGAGTTGTTGACCCTACAAGCAAATCGTCGTCGCTC
CTCAAACTTGTCGCTCTATCAACTCTAGAATGTTGTCTAAGCCGACAGGACCGAACGGTCAATGTGGCGTTCACG
ATTCAGGCATTATACAAGGCCAATCGTGCGGATGCGGCAGGGGCCCTTCTAACGAAGCGGGGTGCTGGATAT

49% identity!

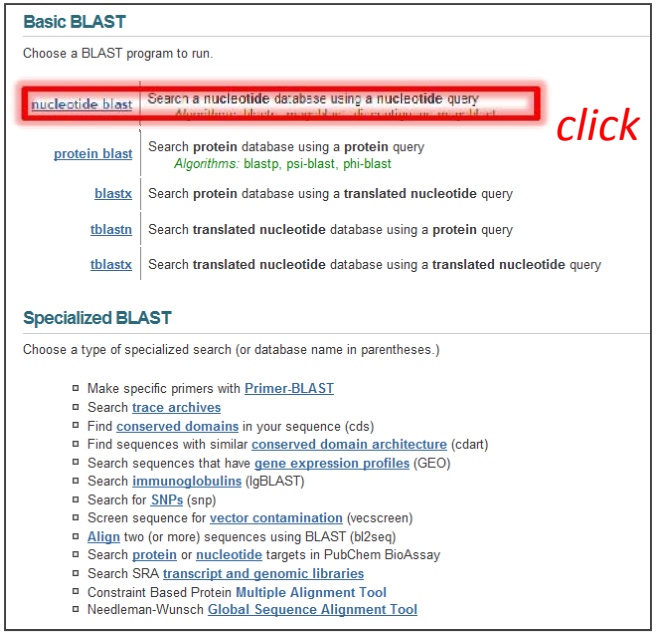
In this case, both strands are in the same orientation

Total length > 300bp due to gaps

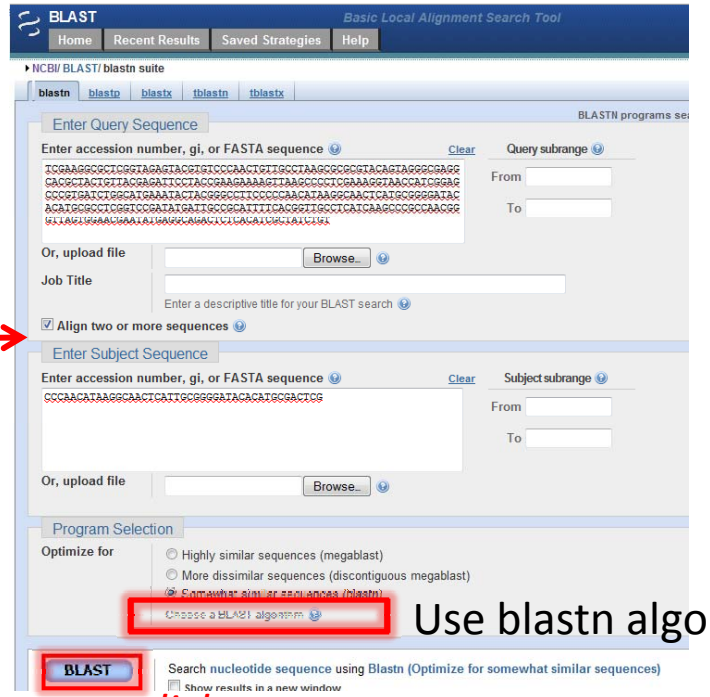
```
>lcl|58213
Length=300

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

Query  1    TCGAAGGCGCTCGGTAGAGTACGTGTCCCA--ACTGTTGCCTAAGCGCGCGTACAGTAGG  58
            || |  || |||  | || | | || | |  ||   || |||    || || || | |
Sbjct  1    CCGTAA-CG-TCGTCACAGAAGGCGTGCGAGCACCAATGTCTATCATCG-GT-CACTTGT  56

Query  59   GCGAGGCACGCTACTGTTACGAGATTC--CTACCGAAG-----AAAAGTTAAGCCCCTCG  111
            |  |||    | |  | |  |||| ||  ||| |  |       || |||    || ||
Sbjct  57   GTTAGGTGTACCAAAGCT--GAGAGTCGTCTATCTTATCTTCTAATAGTA---CCTCTAT  111

Query  112  AAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATACTACGGGCCTTCCCCCAACATA  171
             |||      || | ||    ||| |   || | ||  |  ||   || || | ||| |
Sbjct  112  TAAG------ATTG-AGTTGTTGACCCTACAAGCAAATCGTCGTCGCT-CCTCAAACTTG  163

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

Query  224  TTTTCACGGTT---GCCTCAT-CAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCA  279
            ||  |||| ||   || | || ||||   |||||     | |||   || ||  | ||||
Sbjct  223  TT--CACGATTCAGGCATTATACAAG---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
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCC**CCCAACATAAGGCAACTCATTGCGGGGATACACATGCGACTCG**GTCCGATATGATTGCCGCA
TTTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT



click

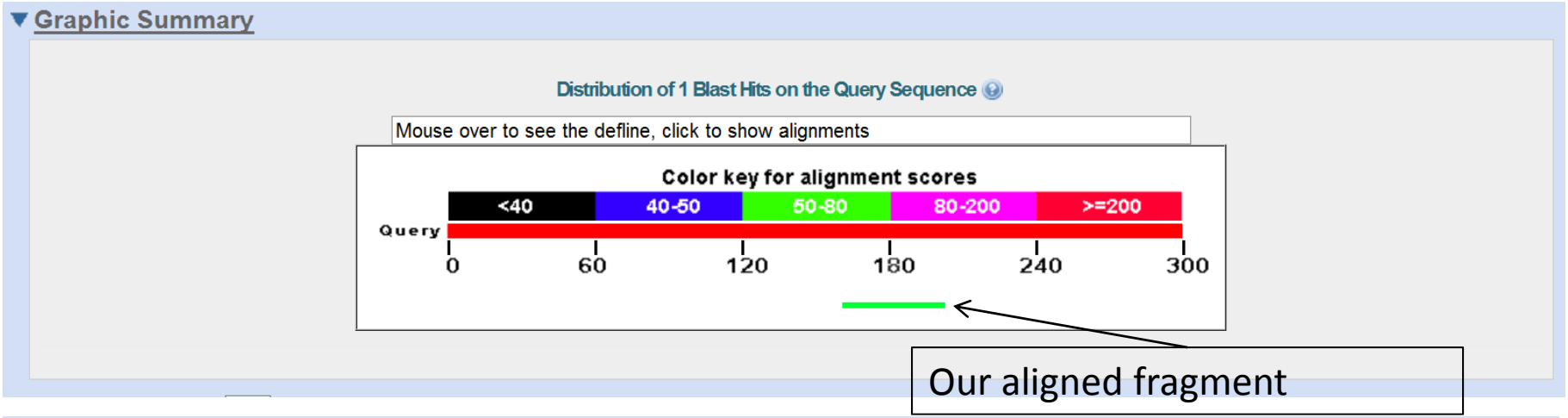


Use blastn algorithm

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
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCC**CCCAACATAAGGCAACTCATTGCGGGGATACACATGCGACTCG**GTCCGATATGATTGCCGCA
TTTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT



Descriptions

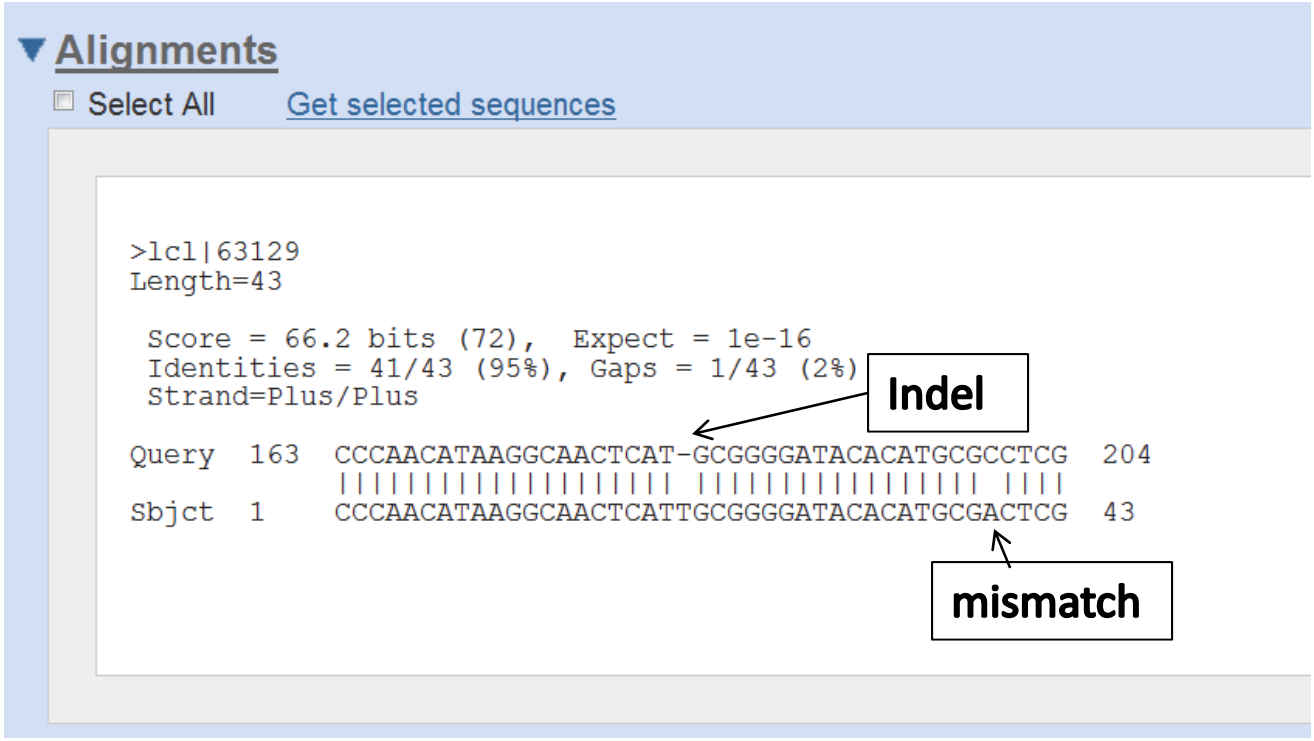
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:

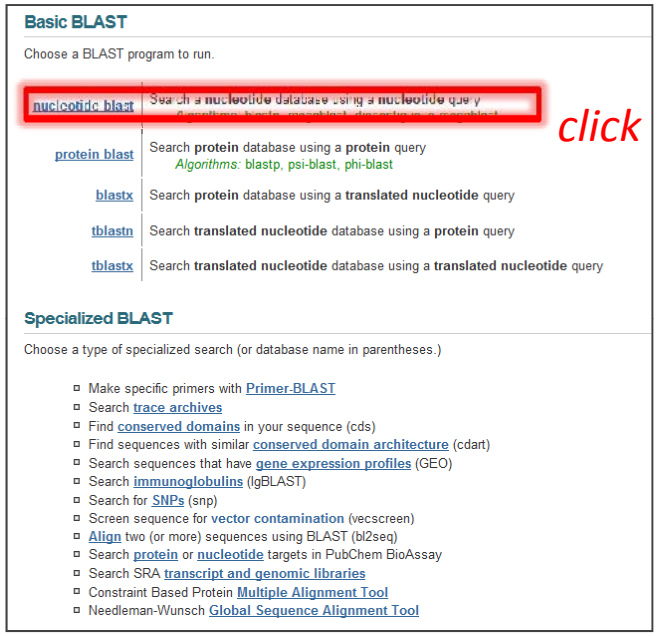
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
63129		66.2	66.2	14%	1e-16	95%	

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

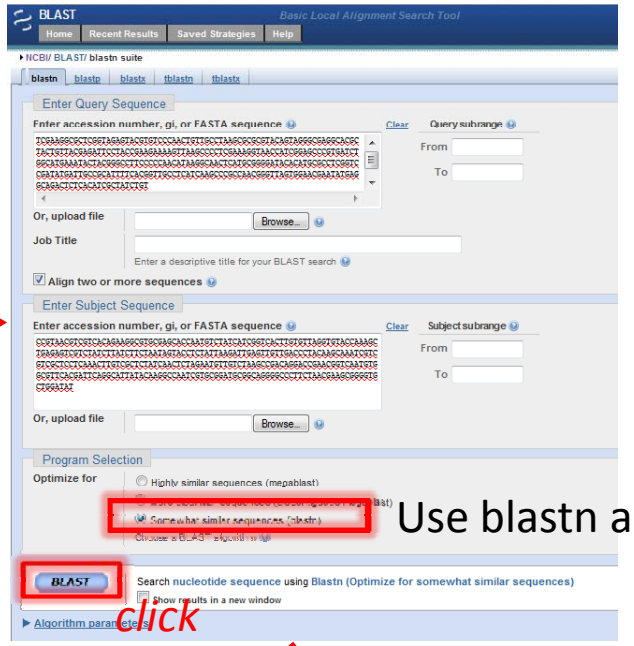
>Random seq 1
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCC**CCCAACATAAGGCAACTCATTGCGGGGATACACATGCGACTCG**GTCCGATATGATTGCCGCA
TTTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT



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

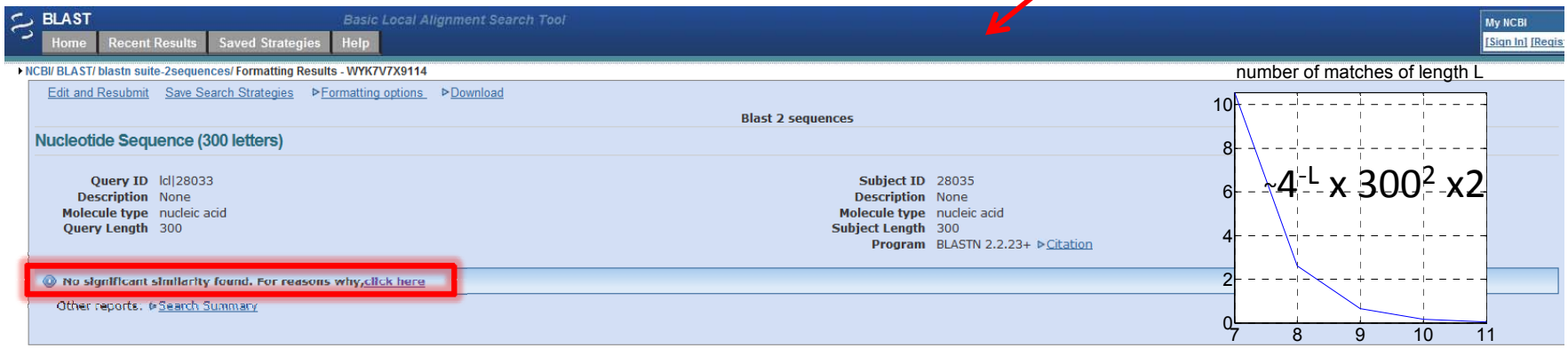


click

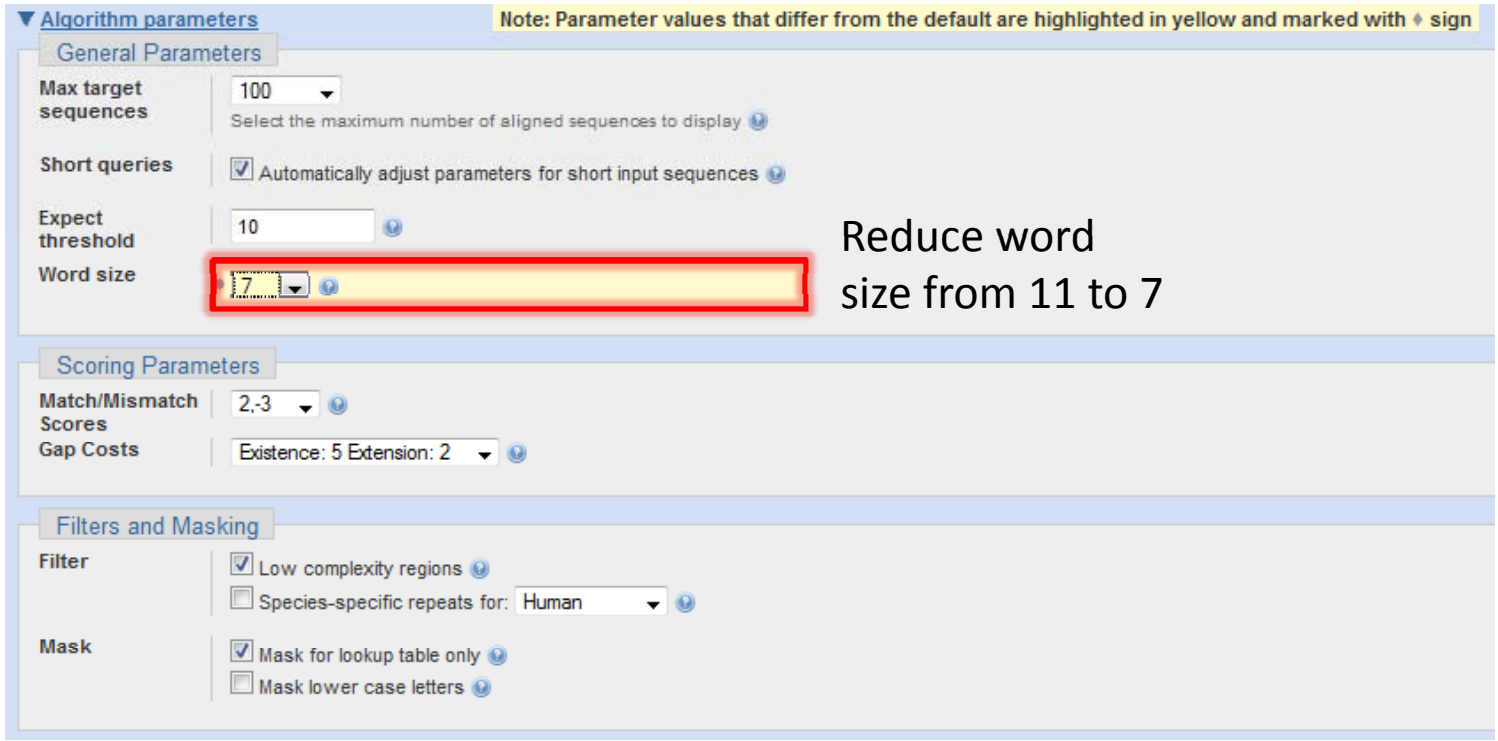


Use blastn algorithm

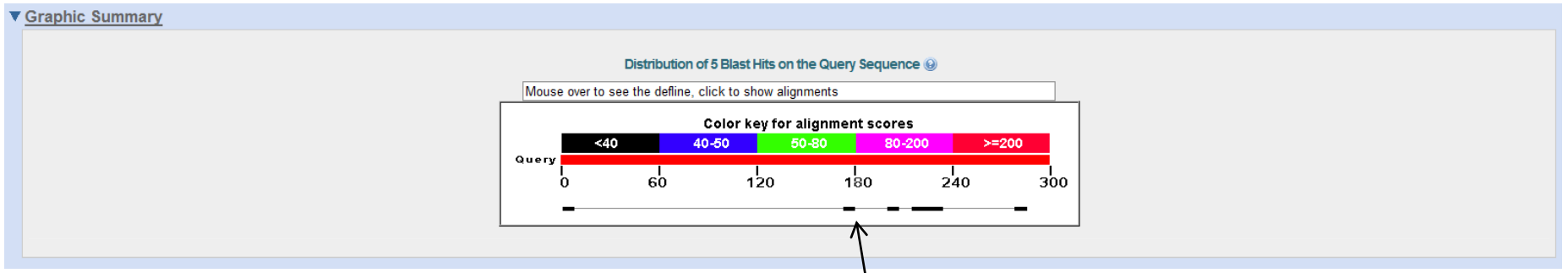
click



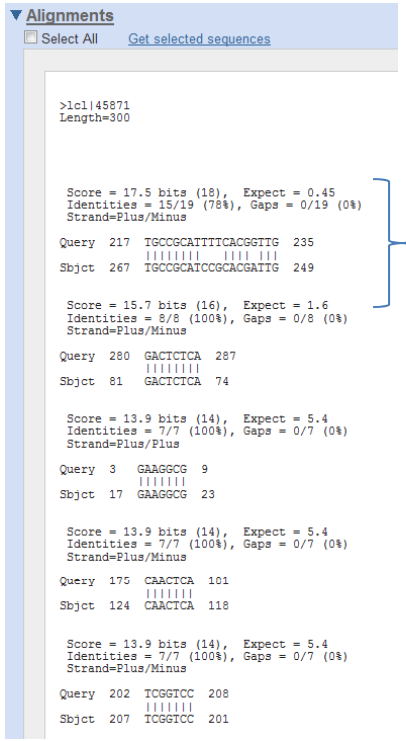
Example 3: Local alignment of two random 300bp nucleotide sequences using BLAST



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



Tiny fragments were aligned




```
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
MYMAENMRFIELPTEEEFWSWWDWVSWRMEMWGSDLMPEQYMRMDSWE
NSEQCRKHSIGRCLHYRHLNLWDDRFAQVSCFNMWWEIFPIGQRHDGY
LYRVRESMIQNQDENTVCPAMFAANWQLLKEHHVRGSKEYREWFFINV
WHTEGSRVAQAH
```

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

>Random seq 1
EMQSSVHIKTADYYITYFGHHFIVGEWLPNIRFPYFFWIITTDARNDMAIFNCQDETQSKKPSYNSDANNNYQYMW
GCDLQEAKVTAMGNLHLWNHRGPRFQKDHACQLCEPHRGITETKRQKIDCSMNPHIPHARKHYRGLNYMYMAE
NMRFIELPTEEEFWSWWDWVSWRMEMWGSDLMPEQYMRMDSWENSEQCRKHSIGRCLHYRHLNLWDDRFA
QVSCFNMWWEIFPIGQRHDGYLYRVRESMIQNQDENTVCPAMFAANWQLLKEHHVRGSKEYREWFFINVWHTE
GSRVAQAH

>Random seq 2
SYLTKSAIQEVCLCKVNVNDMNRFAVLGPYGFMSKGWCSVQPFHIYVPGAKKGWMMQRQCETDDLMDMTSQE
DHEQYGGKRCPKCTHLLDRIAHKMAPDKMSRWGGEKQGEMFVYGDYQKYRQHKWCVHSLEHPYWNNWFALW
GQCCGKQMTNPMIYRKCAKTKTCMDQAPVPSLQCQVCLCHNGSTYLTPANCCDCQVEQHESGNMGGRWIRYQ
MFCVFLWKAITPKAPHFSTASKRQNKLRVQEEALQHYYNKGPLQIWPDDGWFMNRWHIILQCWYMGKFWRLH
MKCNARESEMVML

only 12% identity

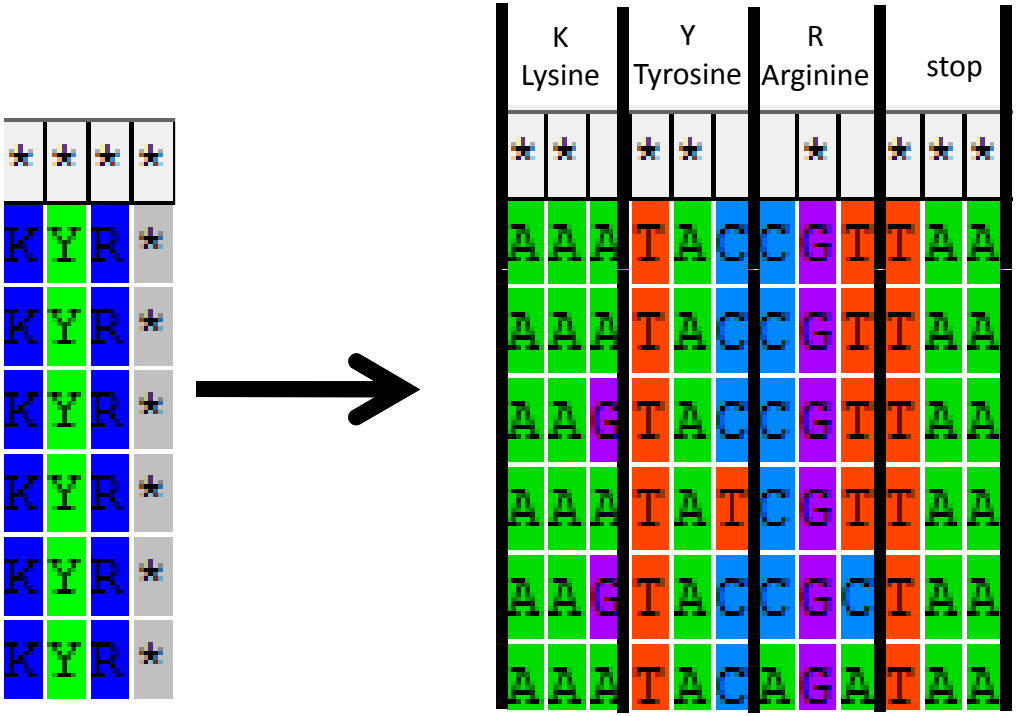
28% similarity

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

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

Query  1    EMQSS----------VHIKTADYYITYFGHHFIVGEWLPNIRFPYFF------WIITTDA  44
             +           V++   + +     + F+   W     F  +       W++
Sbjct  1    SYLTKSAIQEVCLCKVNVNDMNRFAVLGPYGFMSKGWCSVQPFHIYVPGAKKGWMMQRQC  60

Query  45   RND--MAIFNCQDETQ---SKKPSYNSDANNNYQYM-------WGCDLQEAKVTAMGNLH  92
               D  M + + +D  Q    + P       +      M       WG + Q  ++   G+
Sbjct  61   ETDDLMDMTSQEDHEQYGGKRCPKCTHLLDRIAHKMAPDKMSRWGGEKQ-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-NCCDCQVEQHESGNMGGRWIRYQMFCVFLWKA-ITPKAPHFS  233

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

Query  247  SMIQNQDENTVCPAMFAANWQLLKEHHVRGSKEYREWFFINVWHTEGSRVAQAH  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

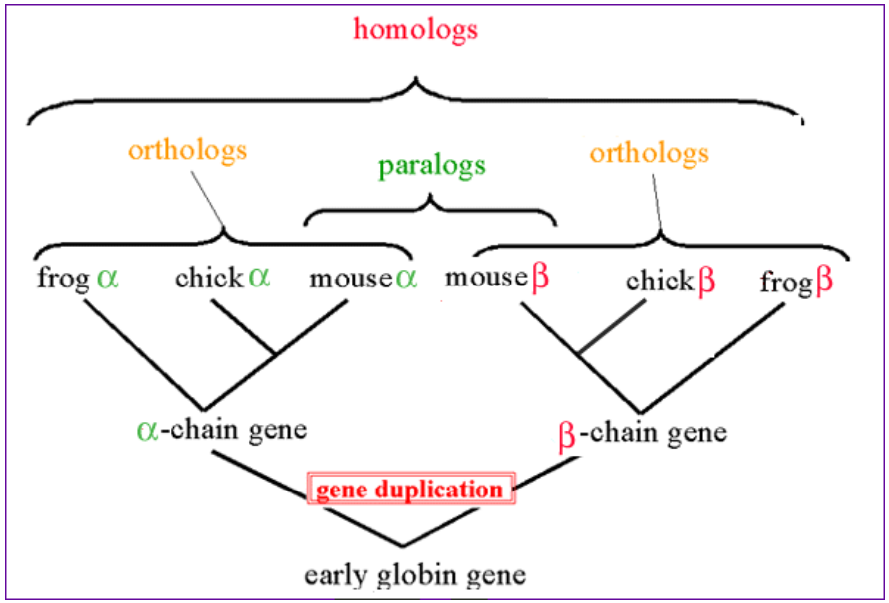
Let's look at the three residues at the C terminus



The Genetic Code

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val	
5′	GCA C G U	CGA C G U or AGA G	AAC U	GAC U	UGC U	CAA G	GAA G	GGA C G U	CAC U	AUA C U	CUA C G U or UUA G	AAA G	AUG	UUC U	CCA C G U	UCA C G U or AGC U	ACA C G U	UGG	UAC U	GUA C G U	3′

Homologs, orthologs, paralogs

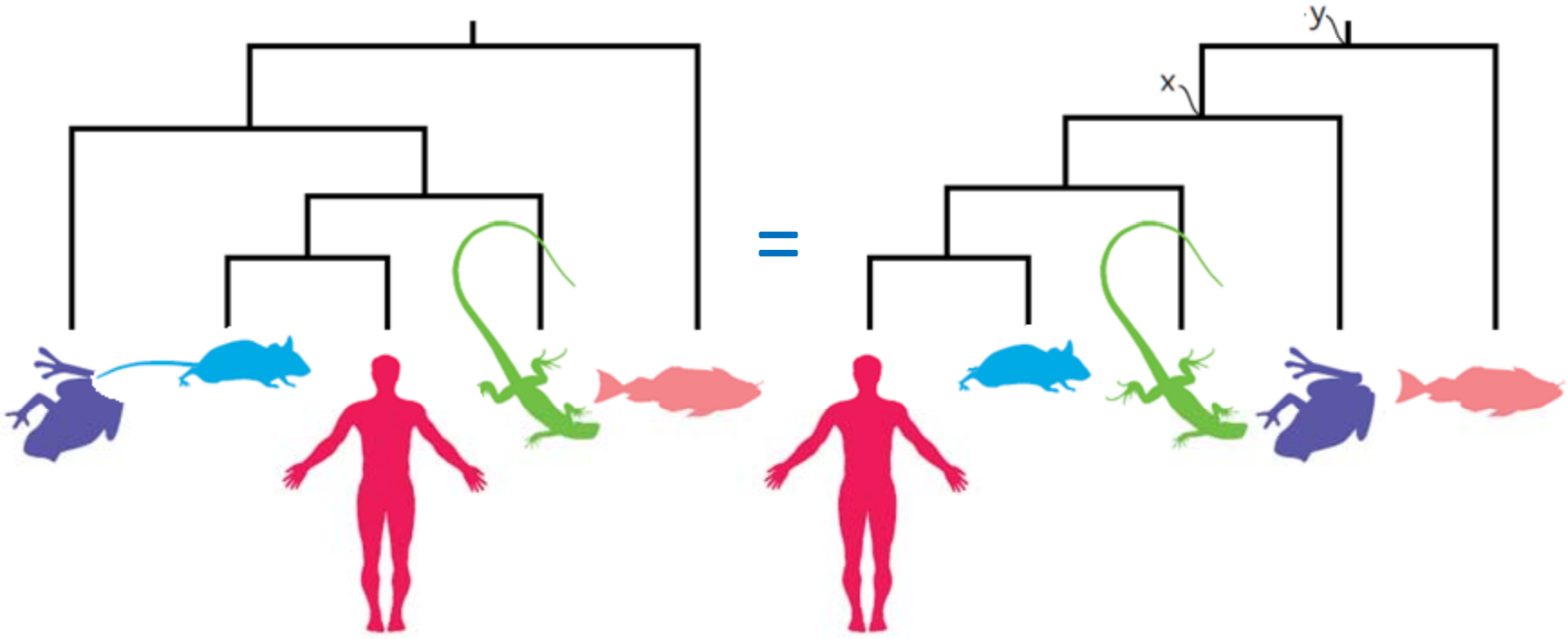


Homologs: similar sequences in two different organisms derived from a common ancestor sequence.

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