Bioinformatics

Bi1X-2010

Part I:

Public databases

Arbel Tadmor

Overview

- Obtaining sequence data from the internet
- Aligning two sequences
- Using the biologist's google: BLAST

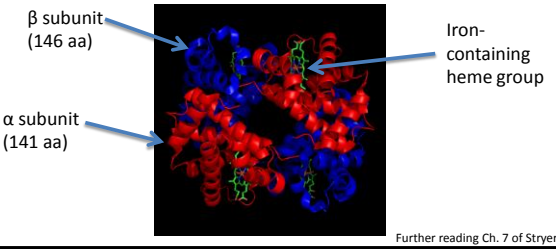
We'll use hemoglobin as a case study

Related reading: Ch. 6 of Stryer (Biochemistry, 7th edition)

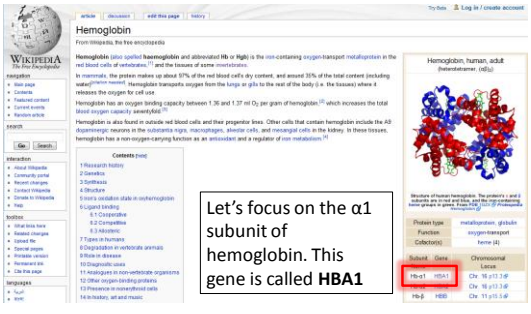
First stop: wiki

Structure of human hemoglobin

- In adults hemoglobin is a tetramer: $\alpha_2\beta_2$
- Each subunit contains a non-protein heme group (that holds an iron)
- The iron binds to an oxygen (shifting absorbance from blue to red)
- Multiple subunits give rise to cooperatively in oxygen binding and unbinding allowing this protein to release more oxygen in the tissues (making it a good transporter.



Further reading Ch. 7 of Stryer

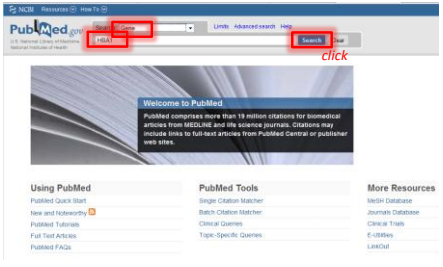


Let's focus on the α1 subunit of hemoglobin. This gene is called **HBA1**

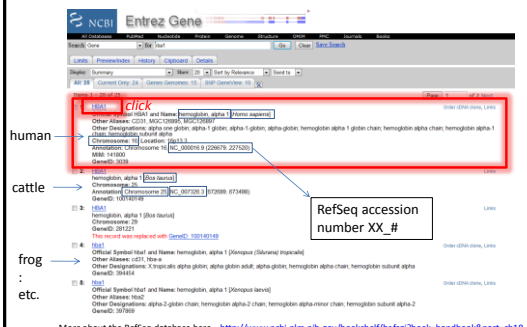
On which chromosome is this gene?

Let's look up this gene

- Go to the PubMed website: *google pubmed* (http://www.ncbi.nlm.nih.gov/PubMed/)
- Search for **HBA1** under **gene** category

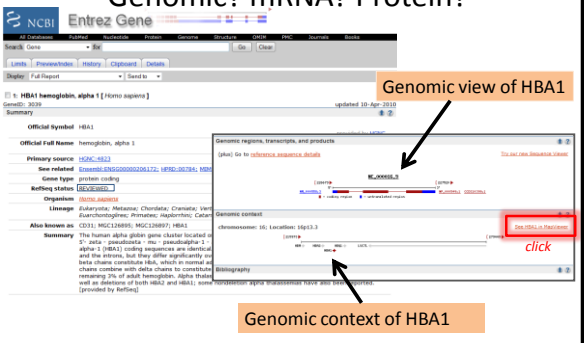


We'll start with human hemoglobin

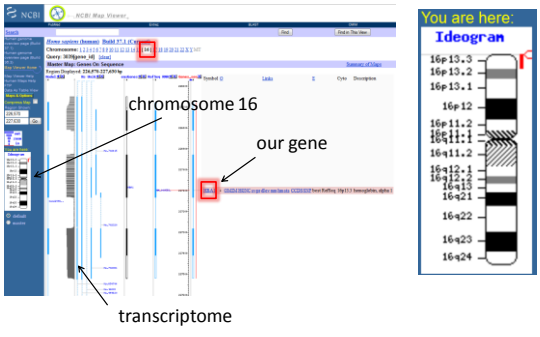


More about the RefSeq database here.. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook&part=ch18
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC539979/pdf/ekj025.pdf

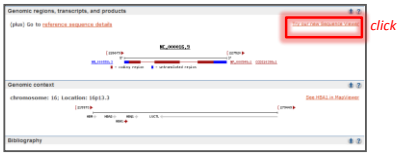
Which sequence should we pick? Genomic? mRNA? Protein?



Where HBA1 falls on chromosome 16…



Genomic context



What is the length of …

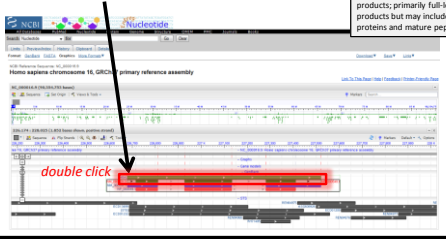
- The genomic sequence (NC_…)?
- The mRNA sequence (NM_…)?
- The protein sequence (NP_…)?

Hint: Double click on gene then Right click→ ***properties…***

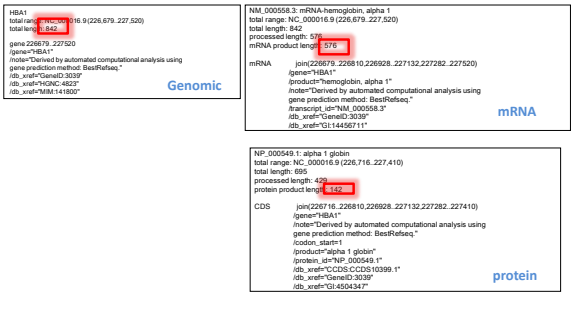
RefSeq codes:
NC_123456 Genomic Mixed Complete genomic molecules including genomes, chromosomes, organelles, plasmids.

NM_123456 mRNA Mixed Transcript products; mature messenger RNA (mRNA) transcripts. NP_123456

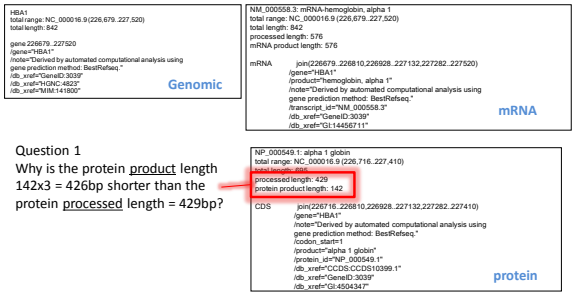
NP_123456789 Protein Mixed Protein products; primarily full-length precursor products but may include some partial proteins and mature peptide products.



Genomic vs. mRNA vs. protein sequences



Genomic vs. mRNA vs. protein sequences



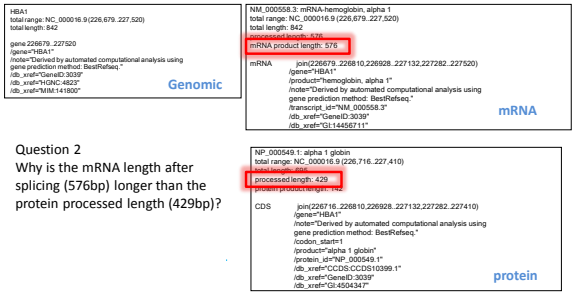
Question 1
Why is the protein product length 142x3 = 426bp shorter than the protein processed length = 429bp?

Genomic vs. mRNA vs. protein sequences

Question 1
Why is the protein product length 142x3 = 426bp shorter than the protein processed length = 429bp?

The stop codon was removed

Genomic vs. mRNA vs. protein sequences



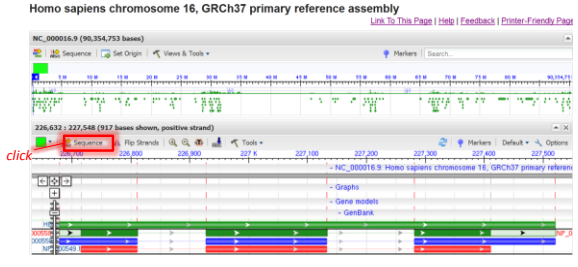
Question 2
Why is the mRNA length after splicing (576bp) longer than the protein processed length (429bp)?

Genomic vs. mRNA vs. protein sequences

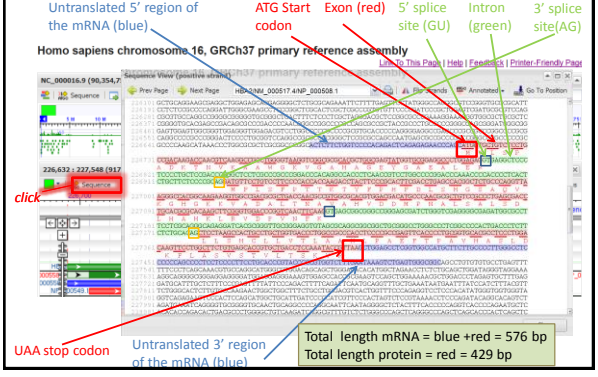
Question 2
Why is the mRNA length after splicing (576bp) longer than the protein processed length (429bp)?

The protein sequence is the sequence between the start and stop codons. The mRNA includes an additional untransalted 5' region and an untransalted 3' regions

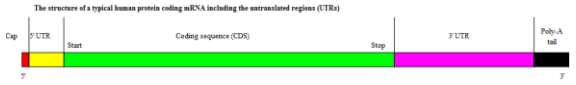
Let's look at the sequence



Let's look at the sequence



Typical Eukaryotic mRNA



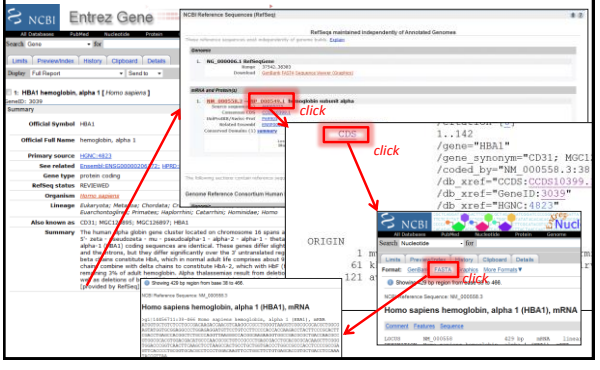
We would like to find similar proteins in nature

Which sequence should we use for the search?

- Genomic?
- mRNA ?
- Protein?

OK let's grab the aa sequence of the protein

and the nt sequence of the protein



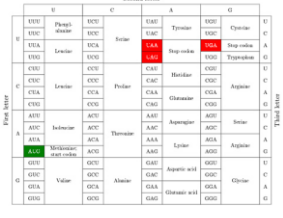
Nucleotide versus amino acid sequences

> NP_000549.1 Homo sapiens nt (429)

ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGG
CCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCA
CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC
GCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAA

> NP_000549.1 Homo sapiens aa (142)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHA
HKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR



Nucleotide versus amino acid sequences

> NP_000549.1 Homo sapiens nt (429)

ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGG
CCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCA
CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC
GCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAA

> NP_000549.1 Homo sapiens aa (142)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHA
HKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

- Which sequence should we use to search with, the amino acid sequence or the nucleotide sequence?

Nucleotide versus amino acid sequences

> NP_000549.1 Homo sapiens nt (429)

ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGG
CCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACGAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCA
CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC
GCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAA

> NP_000549.1 Homo sapiens aa (142)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHA
HKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

- Which sequence should we use to search with, the amino acid sequence or the nucleotide sequence?

It depends on your goal, but generally to find homologs, aa sequences is the way to go:

- Selection pressure on amino acid sequence is much stronger than on nt sequence
- Two random nt sequences share 25% (50% with gaps) identical characters whereas amino acids sequences share only 5% (10-15% with gaps), improving the signal to noise considerably <we'll see this later…>

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 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***

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

	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	-3	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	-3	-2	-3	-2	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
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

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

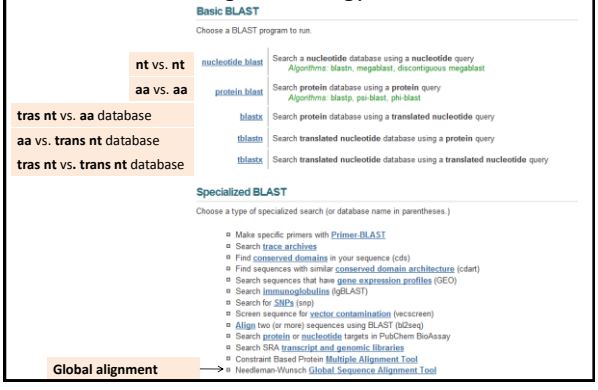
- **Global alignment –** attempts to align every residue in every sequence

```
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*)

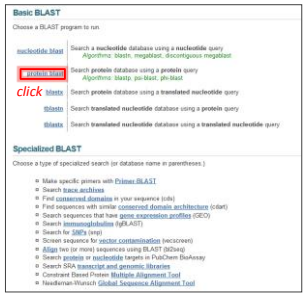


BLAST Scores and Statistics

- **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

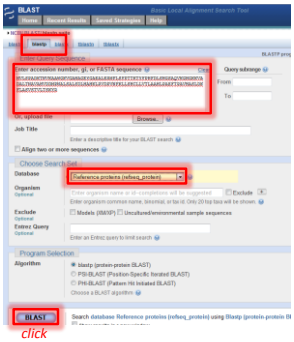
Choose BLAST tool

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

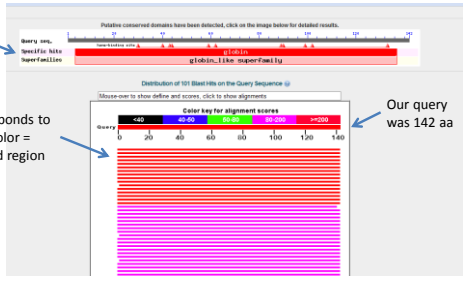


Results

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

Link to gene record

Bit score

E value

Our seq.

Database seq.

+ indicates similar aa

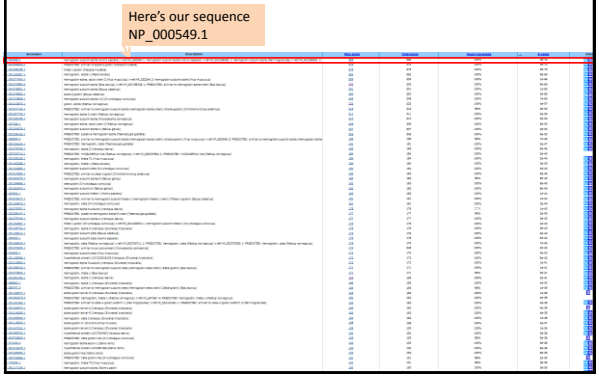
mismatch

106 identities + 14 "+" = 120 positive hits

```
>ref|NP_001013875.1| globin, alpha [Rattus norvegicus]
Length=142

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

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

Query  1    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG  60
            MVLS  DK N+K AW K+G HA E GAE + R+F  FP++KTYFPHF+ S GS QVK HG
Sbjct  1    MVLSEEDKNNIKKAWVKIGNHAAEIGAETIGRLFVFPSSKTYFPHFNTSEGSDQVKAHG  60

Query  61   KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP  120
            KKVADALTNA +H+DD+P  ALS LSDLHAHKLRVDPVNFK LSHCLLVTLA+H P +FTP
Sbjct  61   KKVADALTNAASHLDDLPGALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLASHHPGDFTP  120

Query  121  AVHASLDKFLASVSTVLTSKYR  142
            A+HASLDK LASVSTVLTSKYR
Sbjct  121  AMHASLDKFLASVSTVLTSKYR  142
```

Results

Here's our sequence NP_000549.1



Results

Here's our sequence NP_000549.1

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?

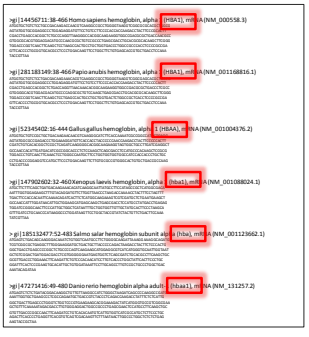
Before we align…

- 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)

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'



Bioinformatics

Bi1X-2010

Part III: Multiple alignment and phylogenetic analysis

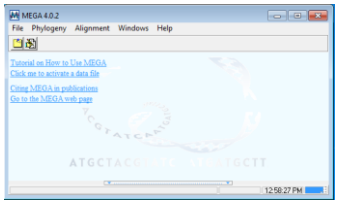
Arbel Tadmor

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

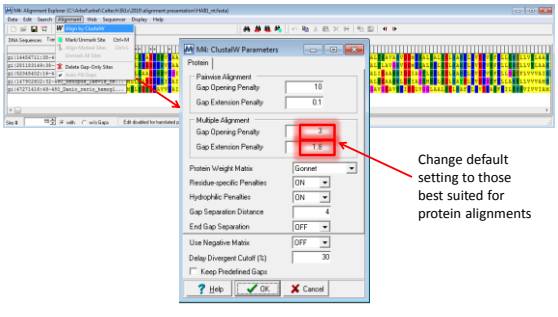
Read more at http://www.ebi.ac.uk/2can/tutorials/nucleotide/clustalw.html

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

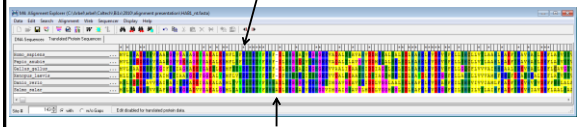
Manual inspection of alignment

- 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

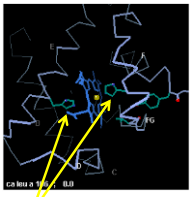
Star indicates conserved character

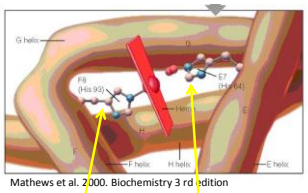


nt view

Deletion/insertion event

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





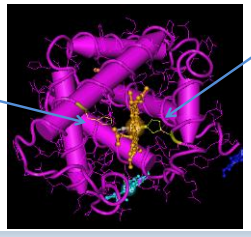
Mathews et al. 2000. Biochemistry 3 rd edition

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

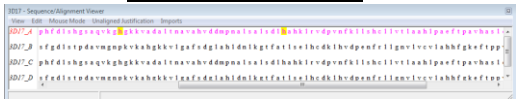
Distal His: The bound oxygen can be in two states, dioxygen (bound to Fe2+) and superoxide (bound to Fe3+). Oxygen must be released in the former because the later is both harmful and leaves the iron in a state that cannot bind oxygen. The distal histidine binds more strongly to superoxide and the oxygen is therefore less likely to be released.

Cn3D demo of Homo sapiens hemoglobin subunit α

His58
Distal His



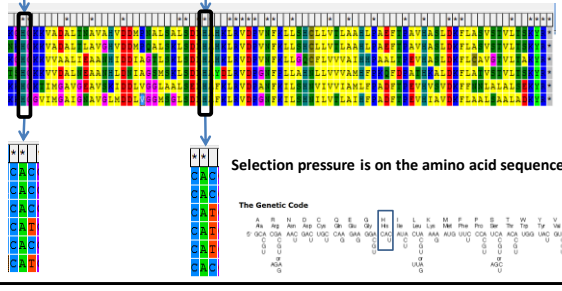
His87
Proximal His



Histidines are conserved

Conserved His58

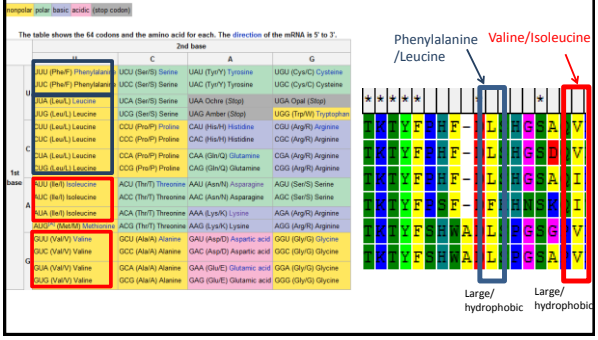
Conserved His87



Selection pressure is on the amino acid sequence

The Genetic Code

Some residues mutate to chemically similar residues



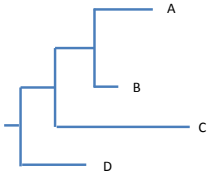
Phenylalanine /Leucine

Valine/Isoleucine

Large/ hydrophobic

Large/ hydrophobic

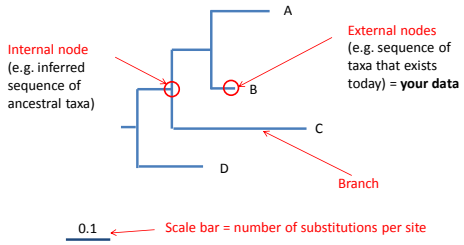
Phylogenetic analysis



What are trees good for?

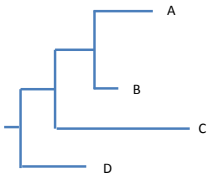
- Identify close relatives
- Determine evolutionary relationship between sequences

Some tree terminology



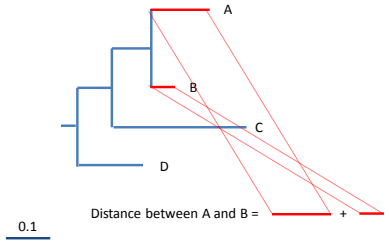
We will discuss only **bifurcating** trees: each node has only two immediate descendant lineages, i.e. we assume evolutionary speciation is a binary process

Trees have two elements: branch lengths and branching order



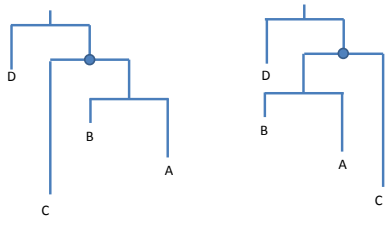
- **Topology** = branching pattern
- To estimate a tree you need to estimate
  - Branch lengths (simple problem)
  - Branching order (difficult for many seq)

Branch lengths



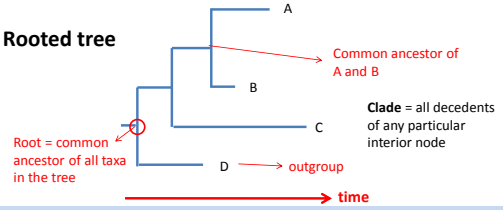
- **Evolutionary distance** = accumulated horizontal distance between two external nodes = estimated number of substitution per site that differ between the two sequences = ***d***

Branch order: note that trees are like mobiles…



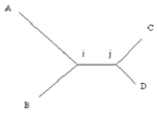
These two trees are equivalent in every way

Rooted trees



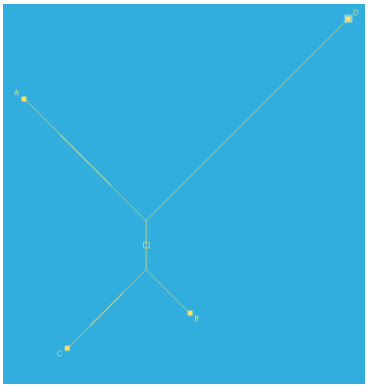
- If a tree it rooted, root = left most internal node
- By selecting a node to be a **root** you set a time arrow
- **The more recently species share a common ancestor, the more closely related they are** (e.g. A and C are more closely related than A and D)
- Which node is the root? You "break the symmetry" 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

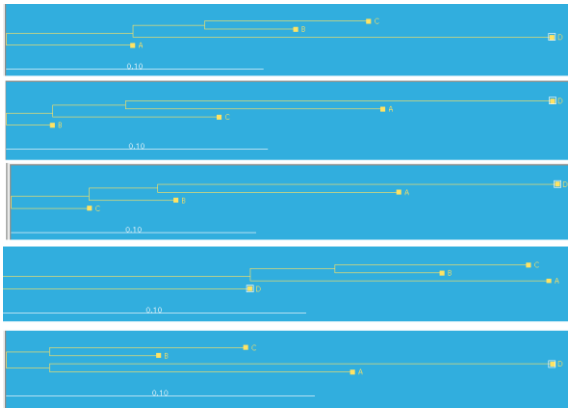
Unrooted trees



5 ways to root this tree

- 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

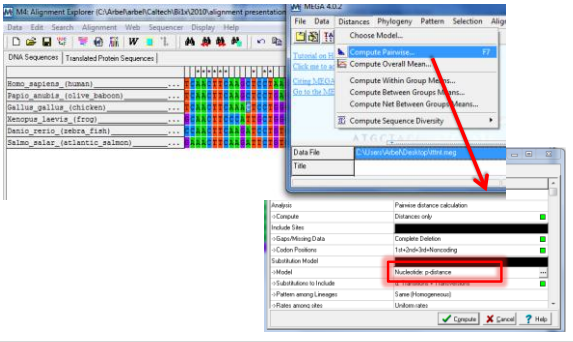
- The evolutionary distance between two sequences *d* is the total number of number of aa/nt substitutions per site between the two sequences
- ***d=2rt***

  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**



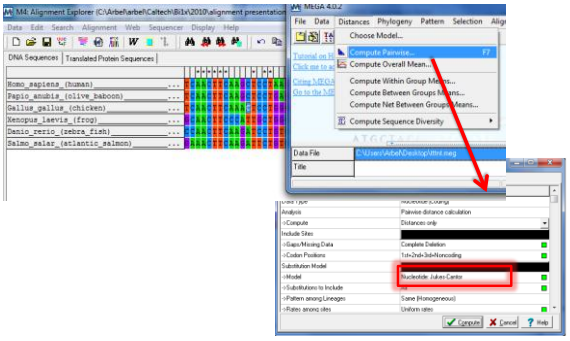
Calculating distances in MEGA4: **p distance**

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

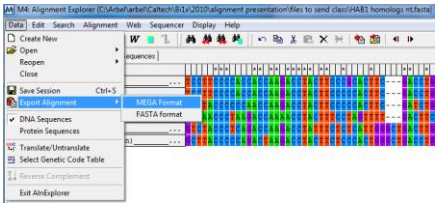
	1	2	3	4	5	6
1. Homo sapiens (human) ...						
2. Papio anubis (olive baboon) ...	0.0531					
3. Gallus gallus (chicken) ...	0.3173	0.3317				
4. Xenopus laevis (frog) ...	0.5837	0.5441	0.5020			
5. Danio rerio (zebra fish) ...	0.5393	0.5736	0.5251	0.6472		
6. Salmo salar (atlantic salmon) ...	0.5637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

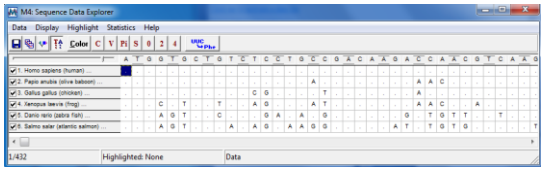
Building a Neighbor joining tree

Export **nt** alignment in MEGA format



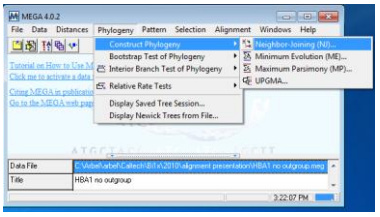
Building a Neighbor joining tree

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



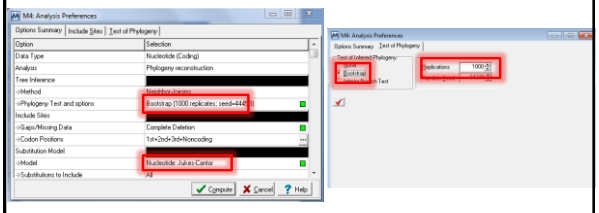
Building a Neighbor joining tree

Now let's build a NJ tree

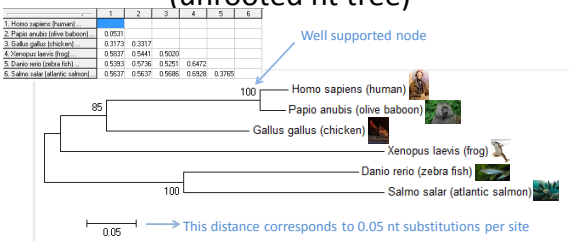


Building a Neighbor joining tree

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



Building a Neighbor joining tree (unrooted nt tree)



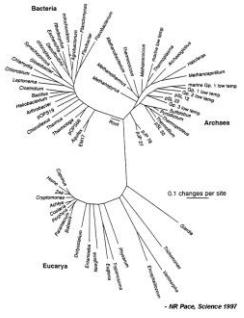
- Human and primates group together
- Fish group together

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



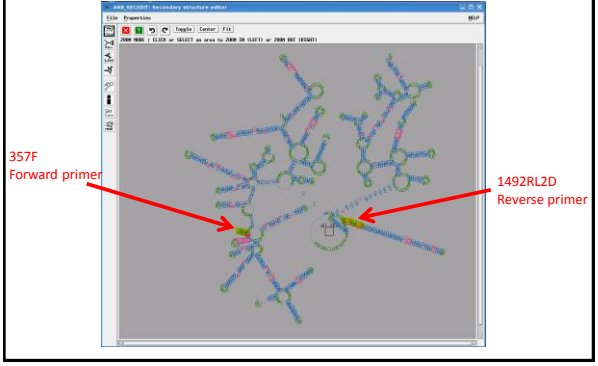
How are SSU rRNA sequences different?

- 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



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

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

>Random seq 1
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGCCTCGGTCCGATATGATTGCCGCATT
TTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCTGT

>Random seq 2
CCGTAACGTCGTCACAGAAGGCGTGCGAGCACCAATGTCTATCATCGGTCACTTGTGTTAGGTGTACCAAAGCTG
AGAGTCGTCTATCTTATCTTCTAATAGTACCTCTATTAAGATTGAGTTGTTGACCCTACAAGCAAATCGTCGTCGCTC
CTCAAACTTGTCGCTCTATCAACTCTAGAATGTTGTCTAAGCCGACAGGACCGAACGGTCAATGTGGCGTTCACG
ATTCAGGCATTATACAAGGCCAATCGTGCGGATGCGGCAGGGGCCCTTCTAACGAAGCGGGGTGCTGGATAT

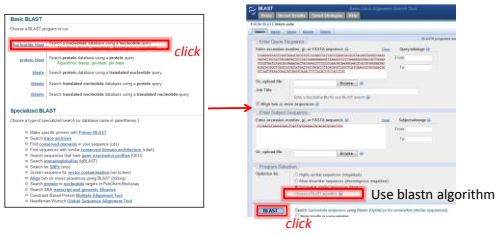
49% identity!

In this case, both strands are in the same orientation

Total length > 300bp due to gaps

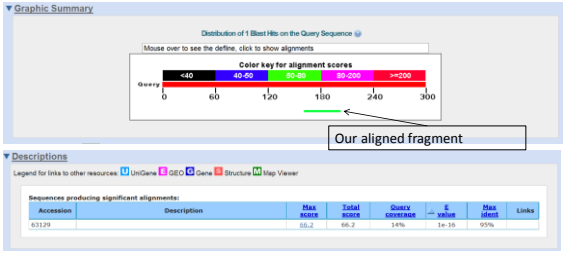
Example 2: Local alignment of a 300bp sequence and an internal fragment containing a single insertion and a single mismatch using BLAST

>Random seq 1
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATTGCGGGGATACACATGCGACTCGGTCCGATATGATTGCCGCA
TTTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT



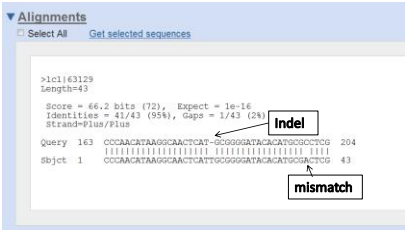
Example 2: Local alignment of a 300bp sequence and an internal fragment containing a single insertion and a single mismatch using BLAST

>Random seq 1
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATTGCGGGGATACACATGCGACTCGGTCCGATATGATTGCCGCA
TTTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT



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
CTACGGGCCTTCCCCCAACATAAGGCAACTCATTGCGGGGATACACATGCGACTCGGTCCGATATGATTGCCGCA
TTTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCT
GT



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
```

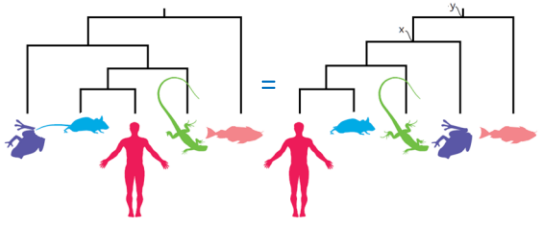
Example 3: Global alignment of two random 300 residue amino acid sequences

>Random seq 1
EMQSSVHIKTADYYITYFGHHFIVGEWLPNIRFPYFFWIITTDARNDMAIFNCQDETQSKKPSYNSDANNNYQYMW
GCDLQEAKVTAMGNLHLWNHRGPRFQKDHACQLCEPHRGITETKRQKIDCSMNPHIPHARKHYRGLNYMYMAE
NMRFIELPTEEEFWSWWDWVSWRMEMWGSDLMPEQYMRMDSWENSEQCRKHSIGRCLHYRHLNLWDDRFA
QVSCFNMWWEIFPIGQRHDGYLYRVRESMIQNQDENTVCPAMFAANWQLLKEHHVRGSKEYREWFFINVWHTE
GSRVAQAH

>Random seq 2
SYLTKSAIQEVCLCKVNVNDMNRFAVLGPYGFMSKGWCSVQPFHIYVPGAKKGWMMQRQCETDDLMDMTSQE
DHEQYGGKRCPKCTHLLDRIAHKMAPDKMSRWGGEKQGEMFVYGDYQKYRQHKWCVHSLEHPYWNNWFALW
GQCCGKQMTNPMIYRKCAKTKTCMDQAPVPSLQCQVCLCHNGSTYLTPANCCDCQVEQHESGNMGGRWIRYQ
MFCVFLWKAITPKAPHFSTASKRQNKLRVQEEALQHYYNKGPLQIWPDDGWFMNRWHIILQCWYMGKFWRLH
MKCNARESEMVML

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

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