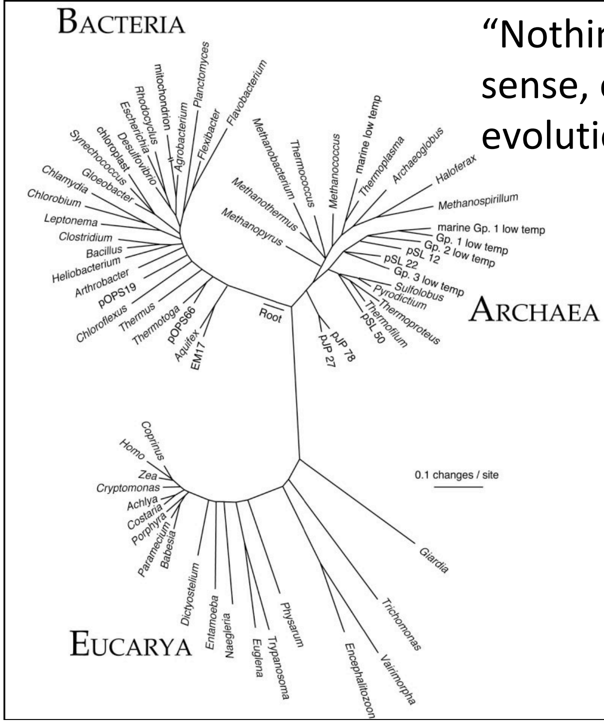
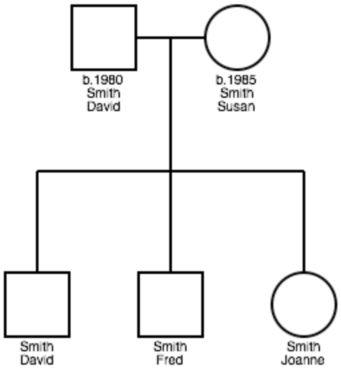


"Nothing in biology makes sense, except in the light of evolution."

-- Theodosius Dobzhansky

Consider basic inheritance …



- DNA is replicated and passed on to the next generation.
- But this is not an error-free process.

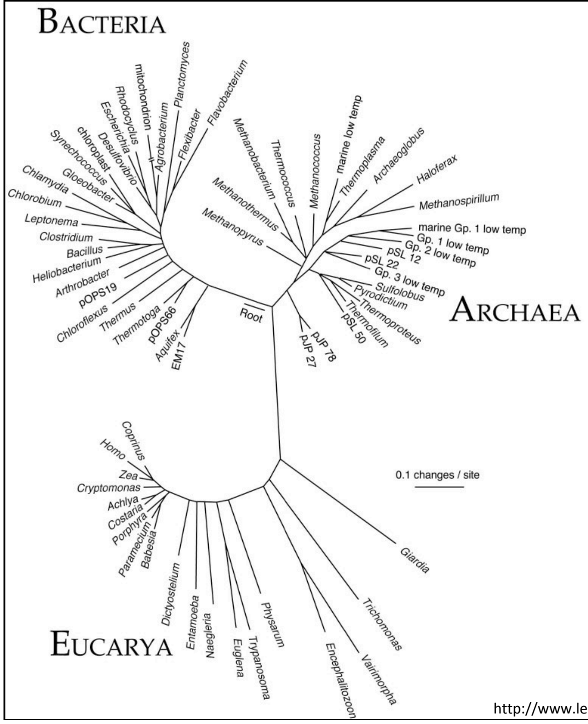
What kinds of errors?

DNA sequence alterations

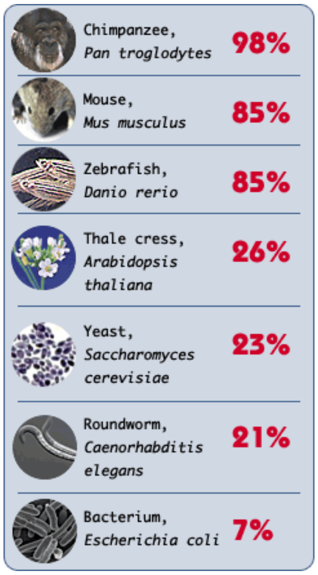
- Substitutions: A**C**GA → A**G**GA
- Insertions: ACGA → AC**CGGA**GA
- Deletions: **ACGG**AGA → AGA
- Transpositions: A**CGG**AGA → AAG**CGG**A
  (move a subsequence)
- Inversions: AC**GGA**GA → AC**TCC**GA

We seek a method to measure
sequence similarity, or how closely
sequences
resemble each other.

Sequence Alignment



Sequence similarity to humans



Chimpanzee, *Pan troglodytes*	98%
Mouse, *Mus musculus*	85%
Zebrafish, *Danio rerio*	85%
Thale cress, *Arabidopsis thaliana*	26%
Yeast, *Saccharomyces cerevisiae*	23%
Roundworm, *Caenorhabditis elegans*	21%
Bacterium, *Escherichia coli*	7%

http://www.learner.org/interactives/dna/project3.html

A reoccurring theme …
sequence alignment

Sequence alignment (being able to measure the similarity between sequences) is necessary for:

* gene finding
* sequence assembly
* prediction of function
* assess evolutionary relationships
* database searching
* mapping to a reference

Arguably the cornerstone of computational genomics.

Overview

- What does it *mean* to *align* sequences?
- How do we cast sequence alignment as a *computational problem?*
- What *algorithms* exist for solving this computational problem?

Sequence comparison overview

- Problem: Find the “best” alignment between a query sequence and a target sequence.
- To solve this problem, we need
  - a method for scoring alignments, and
  - an algorithm for finding the alignment with the best score.
  - a methods to assess whether an alignment is significant
  - best if we can do this “quickly”

Sequence alignment - definition

Sequence alignment is an arrangement of two or more sequences, highlighting their similarity.

For example: the sequences below are padded with **gaps** (dashes) so that wherever possible, columns contain **identical characters** from the sequences involved

```
tcctctgcctctgccatcat---caaccccaaagt
|||| ||| ||||| |||||   ||||||||||||
tcctgtgcatctgcaatcatgggcaaccccaaagt
```

Alignment type

S_1 = FTFTALILLAVAV
S_2 = FTALLLAAV

- Global
  - the whole of each sequence is included, end to end

```
FTFTALILLAVAV
F--TAL-LLA-AV
```

Needleman-Wunch

- Local
  - only the best matching parts of each sequence

```
FTALILLA
FTAL-LLA
```

Smith-Waterman

- Glocal
  - global in query (small), local in reference (big)

```
FTFTALILL-AVAV
  FTAL-LLAAV
```

Scoring alignments

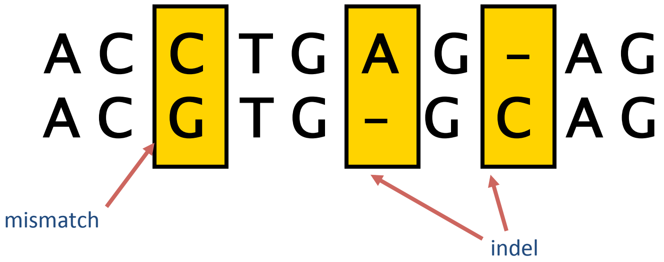
```
GAATC        GAAT-C      -GAAT-C
CATAC        C-ATAC      C-A-TAC

GAATC-       GAAT-C      GA-ATC
CA-TAC       CA-TAC      CATA-C
```

- We need a way to measure the quality of a candidate alignment.

- Alignment scores consist of two parts: a substitution matrix, and a gap penalty.

Percent Sequence Identity

The extent to which two nucleotide or amino acid sequences are invariant



```
A C C T G A G – A G
A C G T G – G C A G
```

mismatch

indel

70% identical

But depending purely on percent identity fails to consider # gaps and length of alignment!

Scoring an Alignment

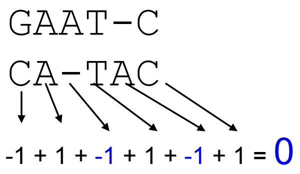
- the score of an alignment is the sum of the scores for pairs of aligned characters plus the scores for gaps
- example: given the following alignment

```
VAHV---D--DMPNALSALSDLHAHKL
AIQLQVTGVVVTDATLKNLGSVHVSKG
```

- we would score it by
  S(V,A) + S(A,I) + S(H,Q) + S(V,L) + S(gap,Q) + S(gap,V) …

Scoring alignments

- Simplest scoring scheme:
  - Match (+1)
  - Mismatch (-1)
  - Gap (-1)

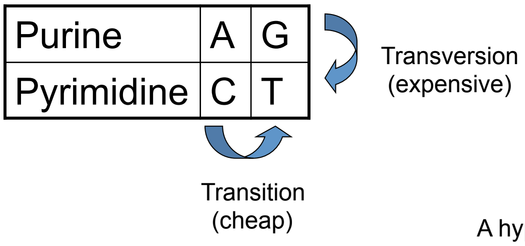


```
GAAT-C
CA-TAC
```

-1 + 1 + -1 + 1 + -1 + 1 = 0

NOTE THAT THESE SCORES ARE ARBITRARY.

Alternatively use substitution matrix



Purine	A	G
Pyrimidine	C	T

Transversion
(expensive)

Transition
(cheap)

These scores are more biologically inspired, BUT still arbitrary!

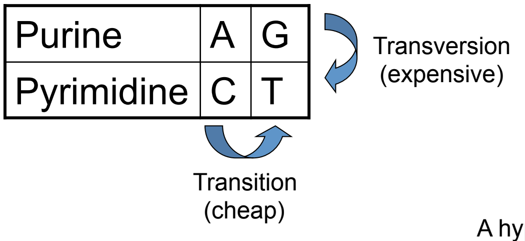
```
GAATC
CATAC
```

-5 + 10 + -5 + -5 + 10 = 5

A hypothetical substitution matrix:

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

So can score any alignment



Purine	A	G
Pyrimidine	C	T

Transversion
(expensive)

Transition
(cheap)

```
GAAT-C
CA-TAC
```

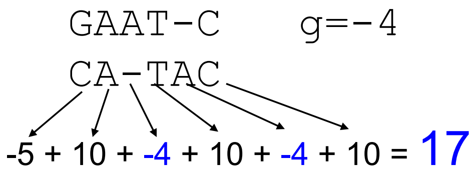
-5 + 10 + ? + 10 + ? + 10 = ?

A hypothetical substitution matrix:

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Scoring gaps

- Linear gap penalty: every gap receives a score of g.



```
GAAT-C    g=-4
CA-TAC
```

-5 + 10 + -4 + 10 + -4 + 10 = 17

My gap penalty implicitly reflects an opinion on gaps given my scoring matrix. So it must "scale" with my arbitrary system.

Additive scoring scheme => independence

Can invent intuitive scoring schemes, but what is the guiding principles?

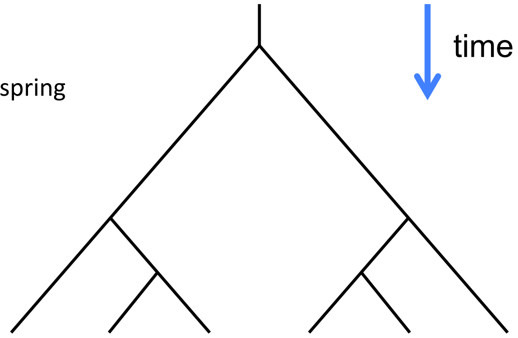
Want assign a score that gives a relative likelihood that the sequences are related (as opposed to unrelated)

How do we develop a non-arbitrary scoring scheme?

- Lets think about this carefully:
  - every type of mutation has an associated probability
  - If we knew those probabilities, we could use them as a scoring scheme
  - BUT … because of selection, not all positions have the same *observed* mutation rate
  - So we need to start by considering whether two sequences derive from a common anscestor …

Homology implies a relationship between sequences, specifically vertical transmission

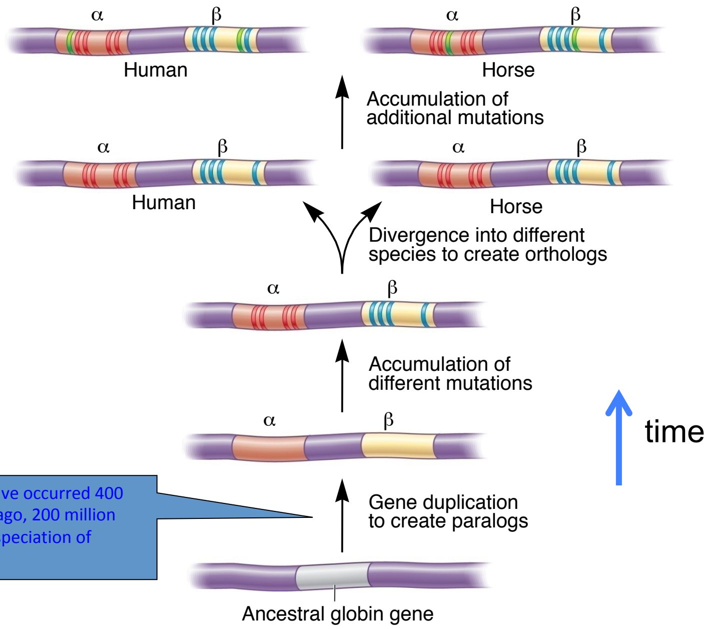
Parents to offspring



Homologous Genes

Two genes are said to be homologous if they are derived from the same ancestral gene

- Orthologous genes or orthologs are homologous genes that arose by a speciation event.
- Paralogous genes or paralogs are homologous genes found that arose by gene duplication.

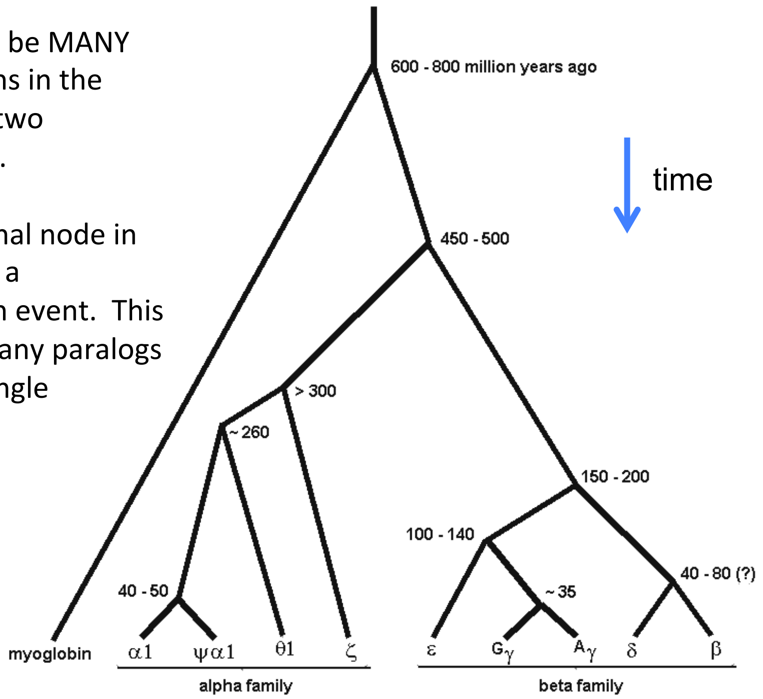

α
β
Human
α
β
Horse
Accumulation of additional mutations
α
β
Human
α
β
Horse
Divergence into different species to create orthologs
α
β
Accumulation of different mutations
α
β
time
Gene duplication to create paralogs
Ancestral globin gene

Thought to have occurred 400 million years ago, 200 million years before speciation of mammals

Evolution of paralogous and orthologous genes

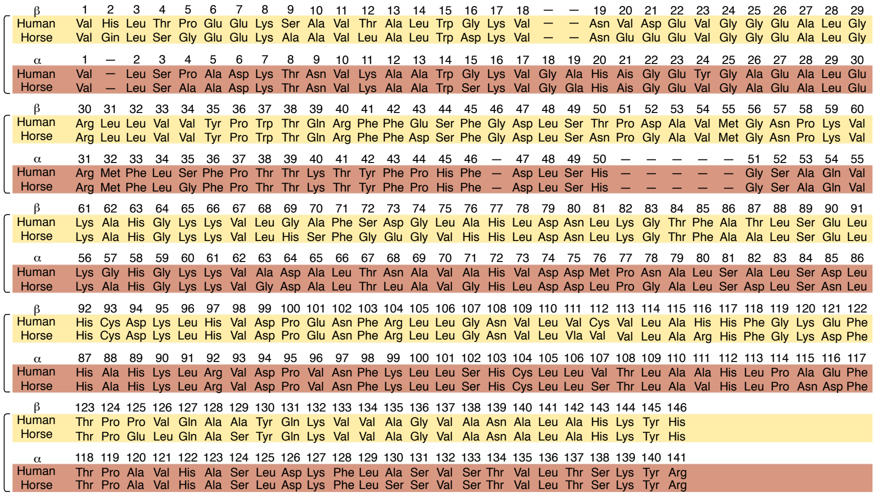
There may be MANY duplications in the history of two sequences.

Each internal node in this tree is a duplication event. This leads to many paralogs within a single species!!



So how can we determine if sequences are orthologs or paralogs…

Hemoglobin

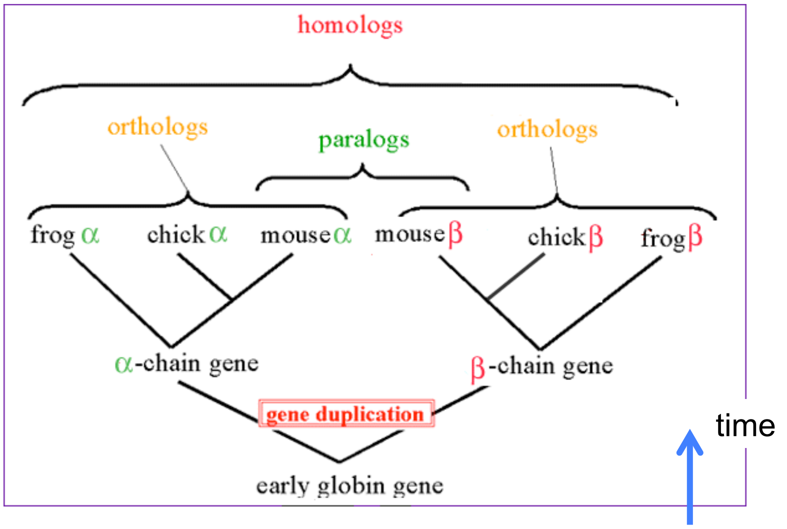


β	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	—	—	19	20	21	22	23	24	25	26	27	28	29
Human	Val	His	Leu	Thr	Pro	Glu	Glu	Lys	Ser	Ala	Val	Thr	Ala	Leu	Trp	Gly	Lys	Val	—	—	Asn	Val	Asp	Glu	Val	Gly	Gly	Glu	Ala	Leu	Gly
Horse	Val	Gin	Leu	Ser	Gly	Glu	Glu	Lys	Ala	Ala	Val	Leu	Ala	Leu	Trp	Asp	Lys	Val	—	—	Asn	Glu	Glu	Glu	Val	Gly	Gly	Glu	Ala	Leu	Gly
α	1	—	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Human	Val	—	Leu	Ser	Pro	Ala	Asp	Lys	Thr	Asn	Val	Lys	Ala	Ala	Trp	Gly	Lys	Val	Gly	Ala	His	Ais	Gly	Glu	Tyr	Gly	Ala	Glu	Ala	Leu	Glu
Horse	Val	—	Leu	Ser	Ala	Ala	Asp	Lys	Thr	Asn	Val	Lys	Ala	Ala	Trp	Ser	Lys	Val	Gly	Gla	His	Ais	Gly	Glu	Val	Gly	Ala	Glu	Ala	Leu	Glu
β	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Human	Arg	Leu	Leu	Val	Val	Tyr	Pro	Trp	Thr	Gln	Arg	Phe	Phe	Glu	Ser	Phe	Gly	Asp	Leu	Ser	Thr	Pro	Asp	Ala	Val	Met	Gly	Asn	Pro	Lys	Val
Horse	Arg	Leu	Leu	Val	Val	Tyr	Pro	Trp	Thr	Gln	Arg	Phe	Phe	Asp	Ser	Phe	Gly	Asp	Leu	Ser	Asn	Pro	Gly	Ala	Val	Met	Gly	Asn	Pro	Lys	Val
α	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	—	47	48	49	50	—	—	—	—	—	51	52	53	54	55
Human	Arg	Met	Phe	Leu	Ser	Phe	Pro	Thr	Thr	Lys	Thr	Tyr	Phe	Pro	His	Phe	—	Asp	Leu	Ser	His	—	—	—	—	—	Gly	Ser	Ala	Gln	Val
Horse	Arg	Met	Phe	Leu	Gly	Phe	Pro	Thr	Thr	Lys	Thr	Tyr	Phe	Pro	His	Phe	—	Asp	Leu	Ser	His	—	—	—	—	—	Gly	Ser	Ala	Gln	Val
β	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91
Human	Lys	Ala	His	Gly	Lys	Lys	Val	Leu	Gly	Ala	Phe	Ser	Asp	Gly	Leu	Ala	His	Leu	Asp	Asn	Leu	Lys	Gly	Thr	Phe	Ala	Thr	Leu	Ser	Glu	Leu
Horse	Lys	Ala	His	Gly	Lys	Lys	Val	Leu	His	Ser	Phe	Gly	Glu	Gly	Val	His	His	Leu	Asp	Asn	Leu	Lys	Gly	Thr	Phe	Ala	Ala	Leu	Ser	Glu	Leu
α	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86
Human	Lys	Gly	His	Gly	Lys	Lys	Val	Ala	Asp	Ala	Leu	Thr	Asn	Ala	Val	Ala	His	Val	Asp	Asp	Met	Pro	Asn	Ala	Leu	Ser	Ala	Leu	Ser	Asp	Leu
Horse	Lys	Ala	His	Gly	Lys	Lys	Val	Gly	Asp	Ala	Leu	Thr	Leu	Ala	Val	Gly	His	Leu	Asp	Asp	Leu	Pro	Gly	Ala	Leu	Ser	Asp	Leu	Ser	Asn	Leu
β	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122
Human	His	Cys	Asp	Lys	Leu	His	Val	Asp	Pro	Glu	Asn	Phe	Arg	Leu	Leu	Gly	Asn	Val	Leu	Val	Cys	Val	Leu	Ala	His	His	Phe	Gly	Lys	Glu	Phe
Horse	His	Cys	Asp	Lys	Leu	His	Val	Asp	Pro	Glu	Asn	Phe	Arg	Leu	Leu	Gly	Asn	Val	Leu	Vla	Val	Val	Leu	Ala	Arg	His	Phe	Gly	Lys	Asp	Phe
α	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117
Human	His	Ala	His	Lys	Leu	Arg	Val	Asp	Pro	Val	Asn	Phe	Lys	Leu	Leu	Ser	His	Cys	Leu	Leu	Val	Thr	Leu	Ala	Ala	His	Leu	Pro	Ala	Glu	Phe
Horse	His	Ala	His	Lys	Leu	Arg	Val	Asp	Pro	Val	Asn	Phe	Lys	Leu	Leu	Ser	His	Cys	Leu	Leu	Ser	Thr	Leu	Ala	Val	His	Leu	Pro	Asn	Asp	Phe
β	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146							
Human	Thr	Pro	Pro	Val	Gln	Ala	Ala	Tyr	Gln	Lys	Val	Val	Ala	Gly	Val	Ala	Asn	Ala	Leu	Ala	His	Lys	Tyr	His							
Horse	Thr	Pro	Glu	Leu	Gln	Ala	Ser	Tyr	Gln	Lys	Val	Val	Ala	Gly	Val	Ala	Asn	Ala	Leu	Ala	His	Lys	Tyr	His							
α	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141							
Human	Thr	Pro	Ala	Val	His	Ala	Ser	Leu	Asp	Lys	Phe	Leu	Ala	Ser	Val	Ser	Thr	Val	Leu	Thr	Ser	Lys	Tyr	Arg							
Horse	Thr	Pro	Ala	Val	His	Ala	Ser	Leu	Asp	Lys	Phe	Leu	Ser	Ser	Val	Ser	Thr	Val	Leu	Thr	Ser	Lys	Tyr	Arg							

(a) Alignment of human and horse globin polypeptides

A comparison of the α- and β-globin polypeptides from humans and horses

If we have the tree, it's easy …



But from sequence alignment, how do we tell which proteins are homologs?

```
SEQ 1: GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
       G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
SEQ 2: GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

SEQ 1: GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
       ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
SEQ 2: NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

SEQ 1: GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
       GS+ + G     +D L  ++ H+ D+  A +AL D    ++AH+
SEQ 2: GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Are these proteins homologs?

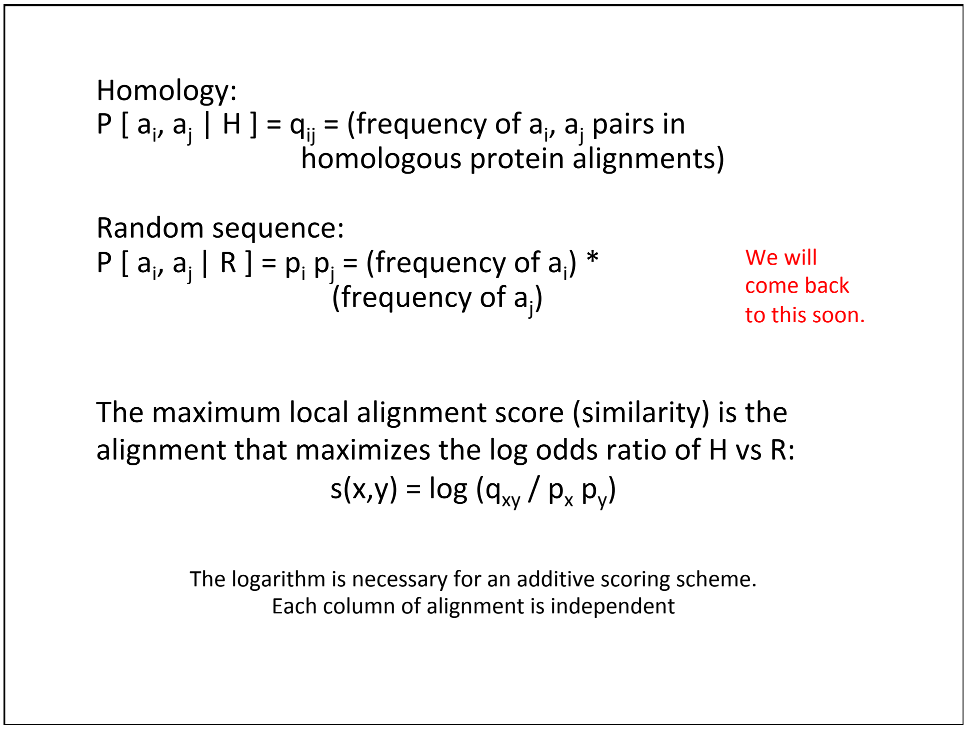
Both negatively charged

```
SEQ 1: GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
       G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
SEQ 2: GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

SEQ 1: GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
       ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
SEQ 2: NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

SEQ 1: GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
       GS+ + G     +D L  ++ H+ D+  A +AL D    ++AH+
SEQ 2: GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Caution: similarity **does NOT** equal homology!!

```
SEQ 1: GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
       G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
SEQ 2: GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

SEQ 1: GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
       ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
SEQ 2: NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

SEQ 1: GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
       GS+ + G     +D L  ++ H+ D+  A +AL D    ++AH+
SEQ 2: GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

But barring any additional information, it's often the best we can do.



Homology:
$P [a_i, a_j \mid H] = q_{ij}$ = (frequency of a_i, a_j pairs in homologous protein alignments)

Random sequence:
$P [a_i, a_j \mid R] = p_i \, p_j$ = (frequency of a_i) * (frequency of a_j)

We will come back to this soon.

The maximum local alignment score (similarity) is the alignment that maximizes the log odds ratio of H vs R:

$$s(x,y) = \log (q_{xy} / p_x \, p_y)$$

The logarithm is necessary for an additive scoring scheme.
Each column of alignment is independent