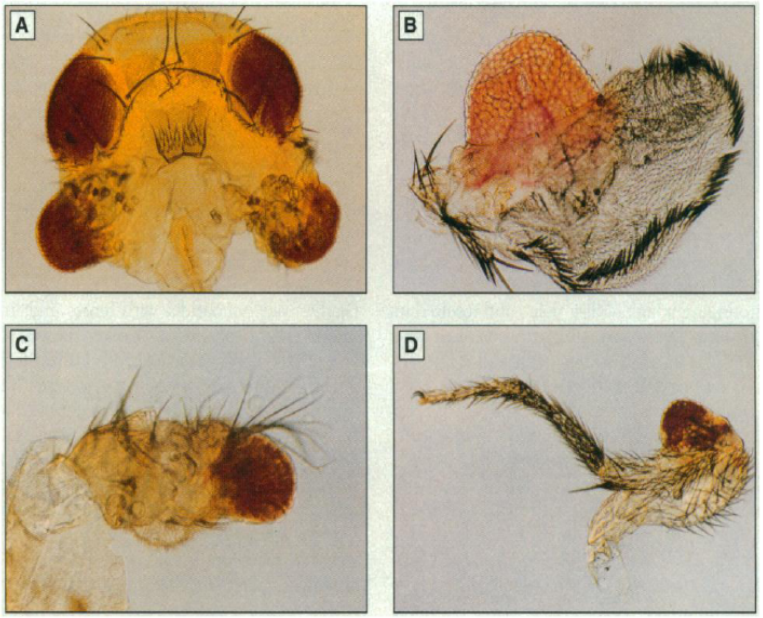
Sequence Alignment

The Story of *eyeless*

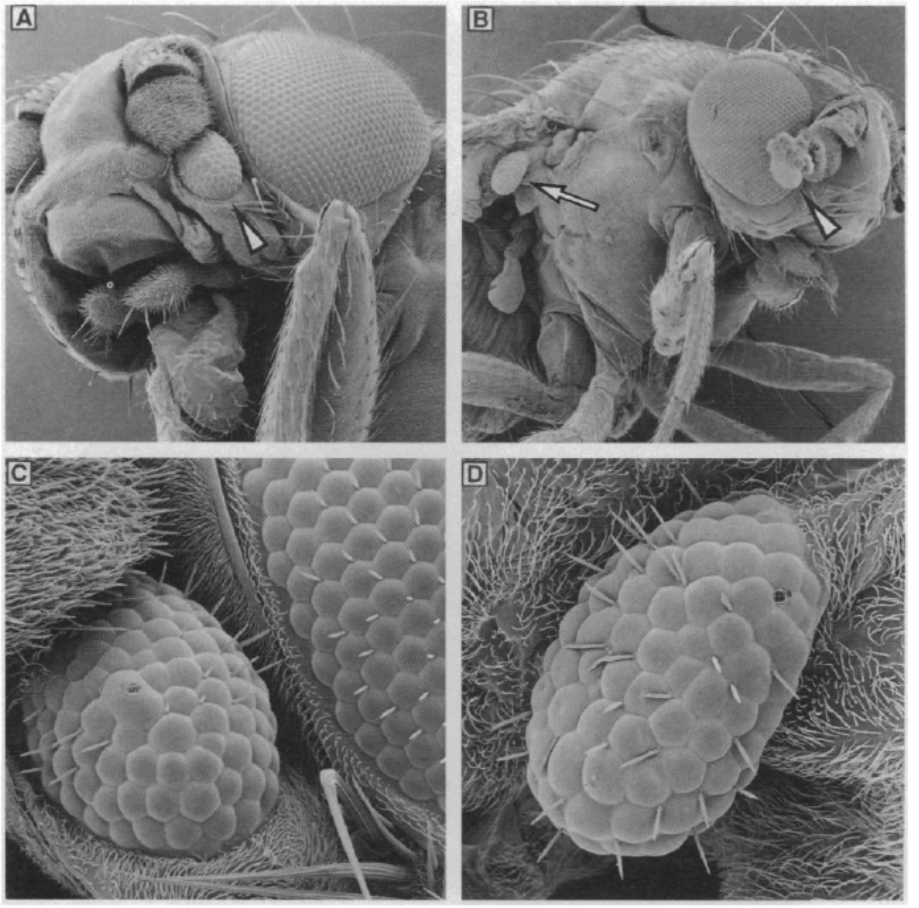
Wild type

eyeless

Muahahahaha....

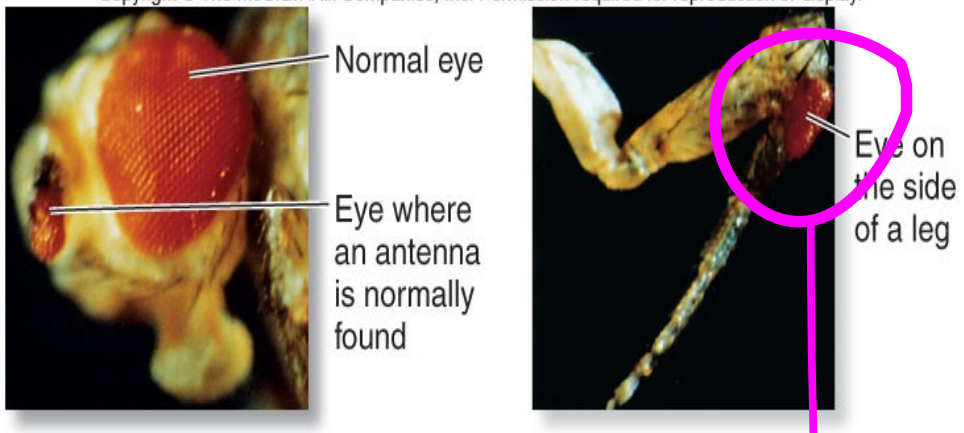


SCIENCE • VOL. 267 • 24 MARCH 1995 1789



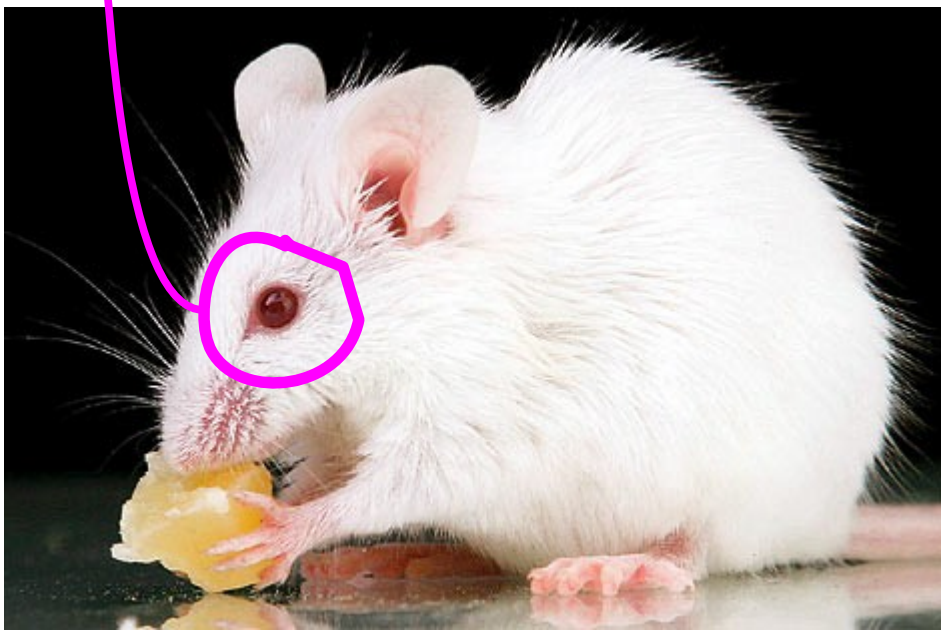
Ectopic expression of Drosophila *eyeless* gene

.....hahahahaha

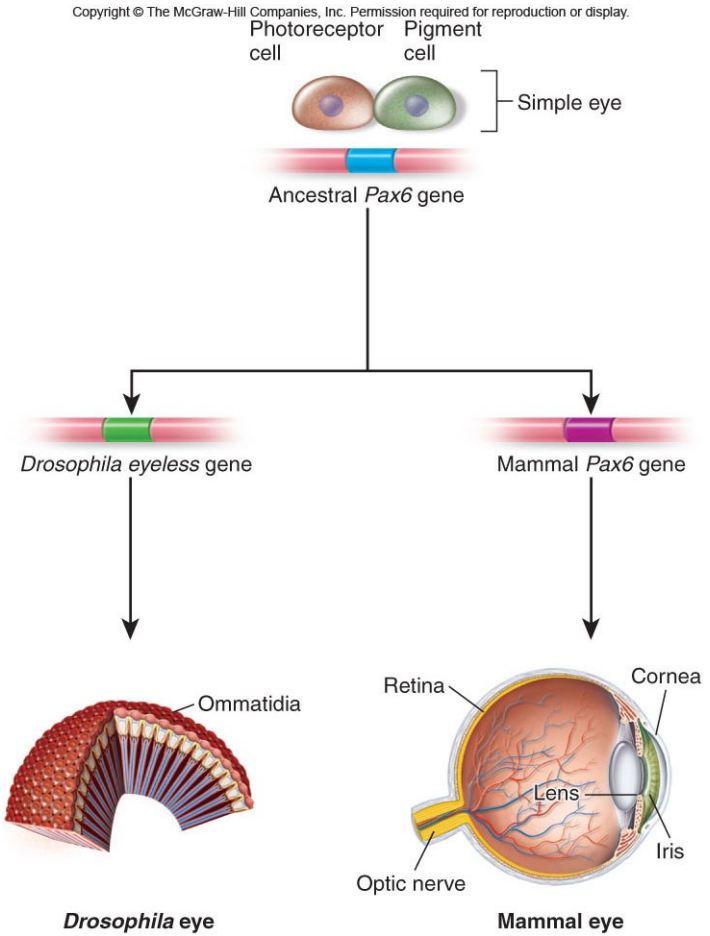


(a) Abnormal expression of *Drosophila eyeless* gene

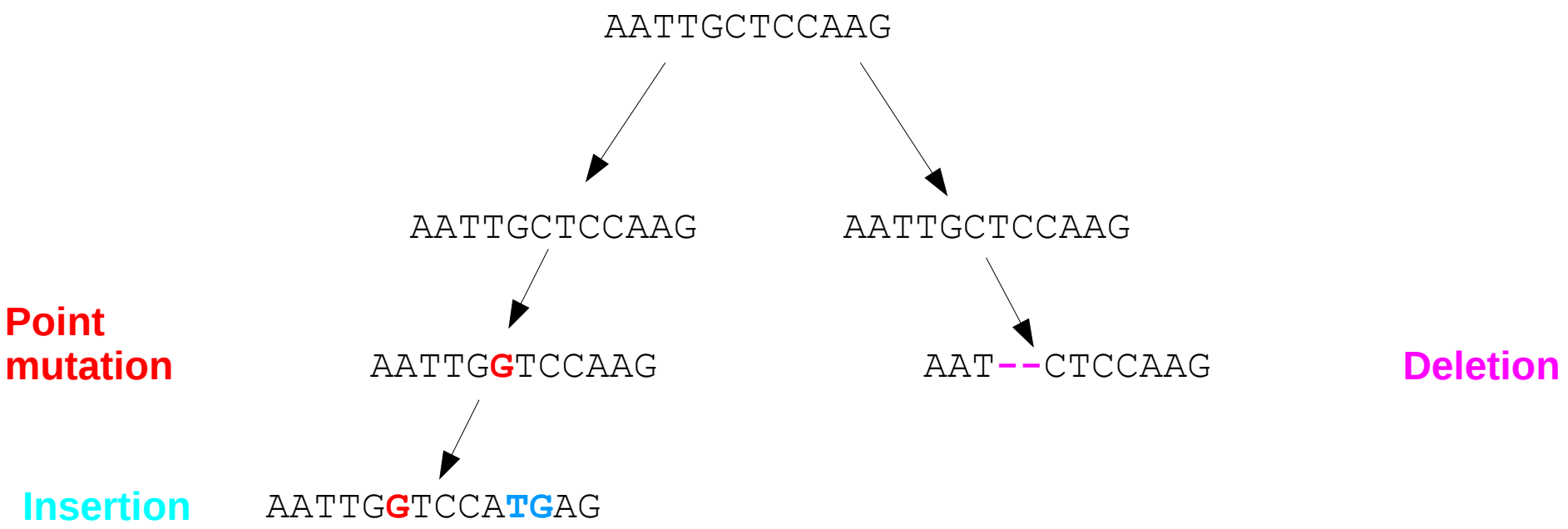
(b) Abnormal expression of mouse *Pax6* gene in a fruit fly leg



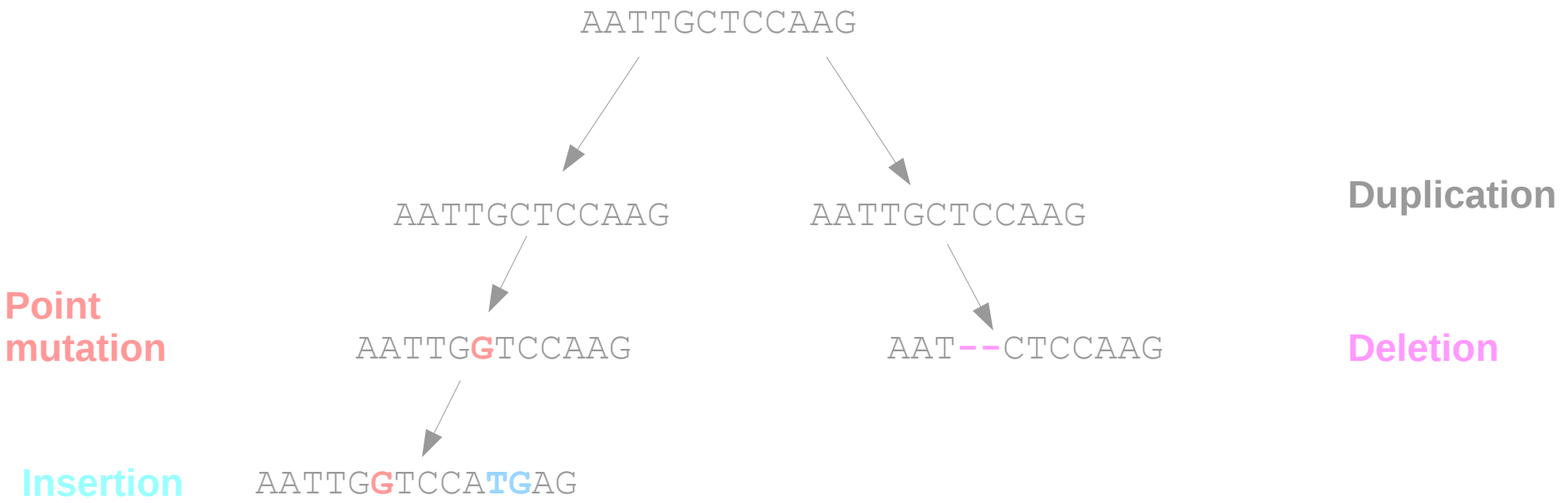
Homology: Common ancestry

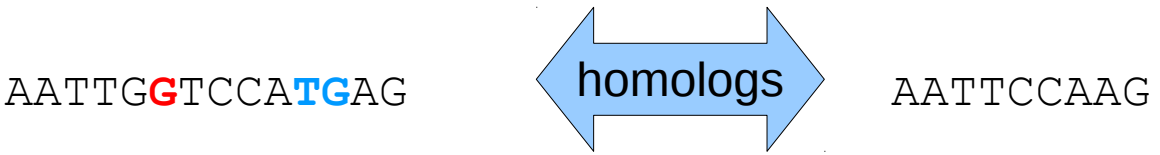


Sequence Evolution Events

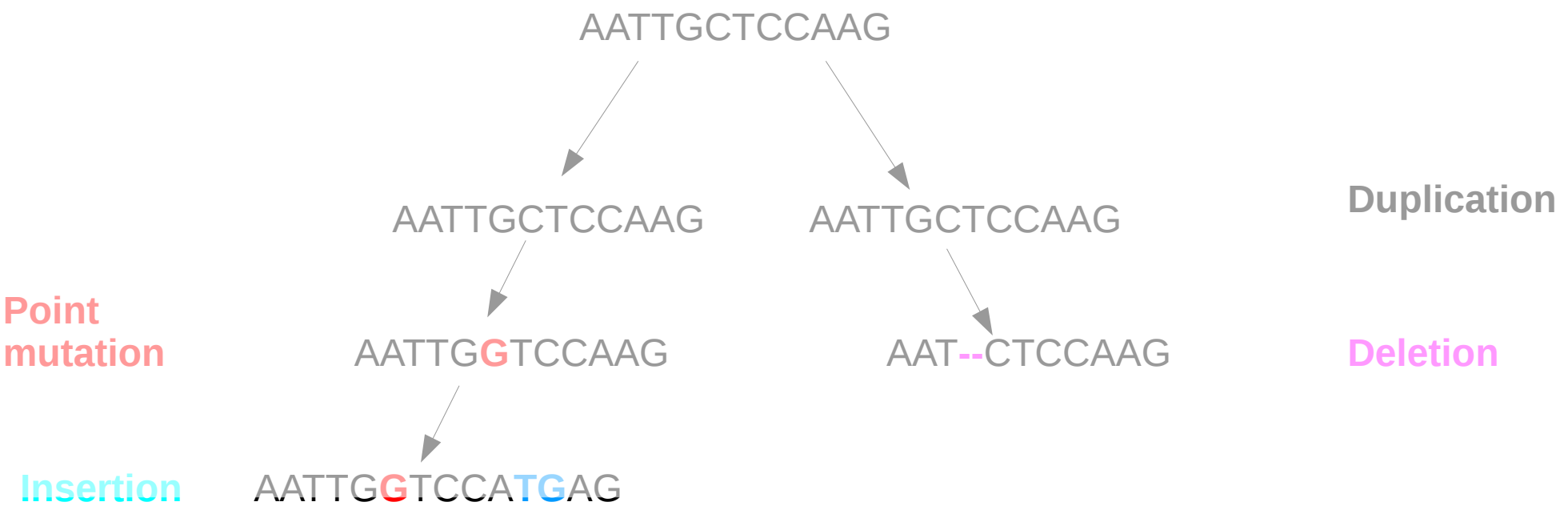


Sequence Evolution Events





Sequence Evolution Events



AATTGGTCCATGAG
AAT--CTCCA--AG

Sequence alignment:
reconstructing evolutionary events since LUCA

Sequence Alignment for

- Predicting function
- Searching sequence databases
- Gene finding (not by length alone...)
- Genome construction / sequence assembly

Local & Global Alignments

Global: indicative of whole gene/protein homology

Local: Common substructure in two different genes / proteins

```
LGPSSKQTGKGS-SRIWDN
|    |  |||   |  |
LN-ITKSAGKGAIMRLGDA
```

```
------QTGKGS-SRIWDN
|     |  |||   |  |
------SAGKGAIMRLGDA
```

Cost of Evolutionary Events

$c(-, a) = c(a,-) = -1$ insertion/deletion (indel)

$c(a,b) = -1$ when: $a \neq b$ mutation

$c(a,b) = 1$ when: $a = b$

Cost of Evolutionary Events

$c(-, a) = c(a,-) = -1$ insertion/deletion (indel)

$c(a,b) = -1$ when: $a \neq b$ mutation

$c(a,b) = 1$ when: $a = b$

Also known as **scoring** or **cost function:** mutations and indels are more **Costly** than evolutionary “inertia”

Homologous Camels

GAMAL

JAMAL

GAMLA

KAMELOS

CAMEL

Exercise

BIOLOGY
LIMNOLOGY

Exercise

```
--BIOLOGY
LIMNOLOGY
```

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

Exercise

BI--OLOGY
LIMNOLOGY

-1+1+-1+-1 +1+ 1+1+1+1 = 3

Exercise

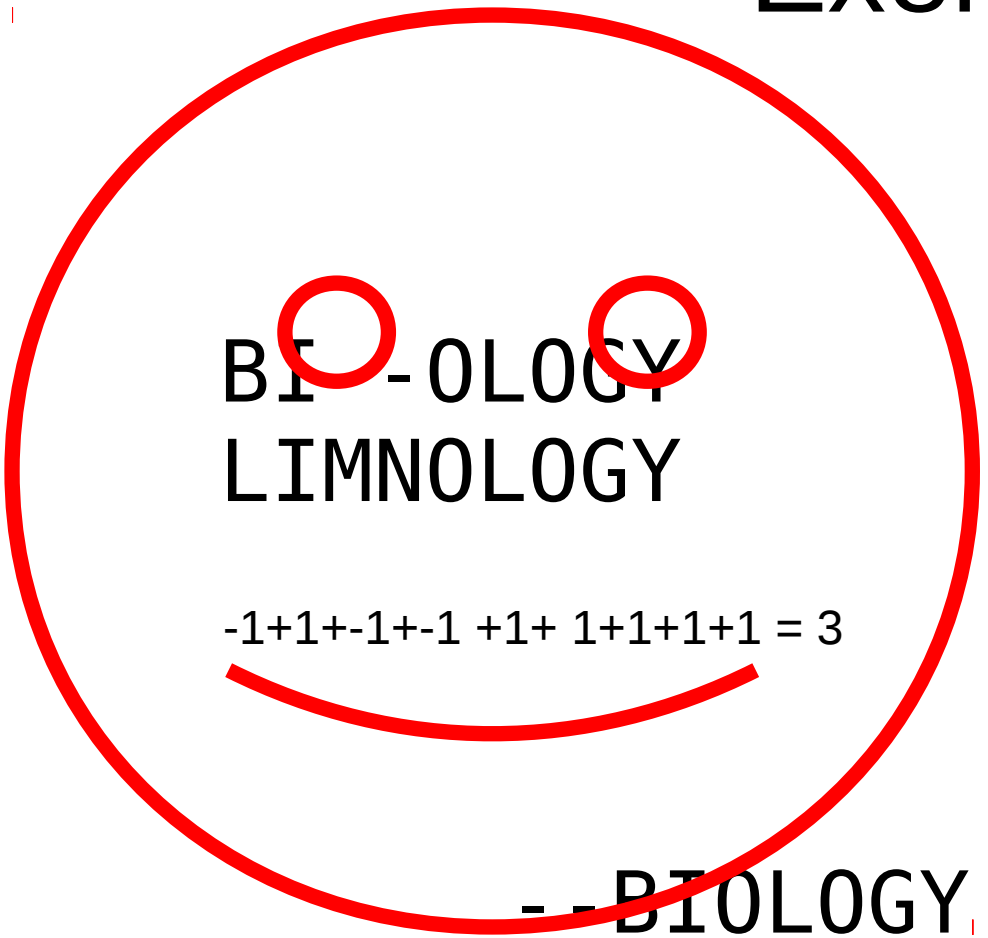
```
BI--OLOGY
LIMNOLOGY
```

-1+1+-1+-1 +1+ 1+1+1+1 = 3

```
  --BIOLOGY
LIMNOLOGY
```

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

Exercise



BI-OLOGY
LIMNOLOGY

-1+1+-1+-1 +1+ 1+1+1+1 = 3

--BIOLOGY
LIMNOLOGY

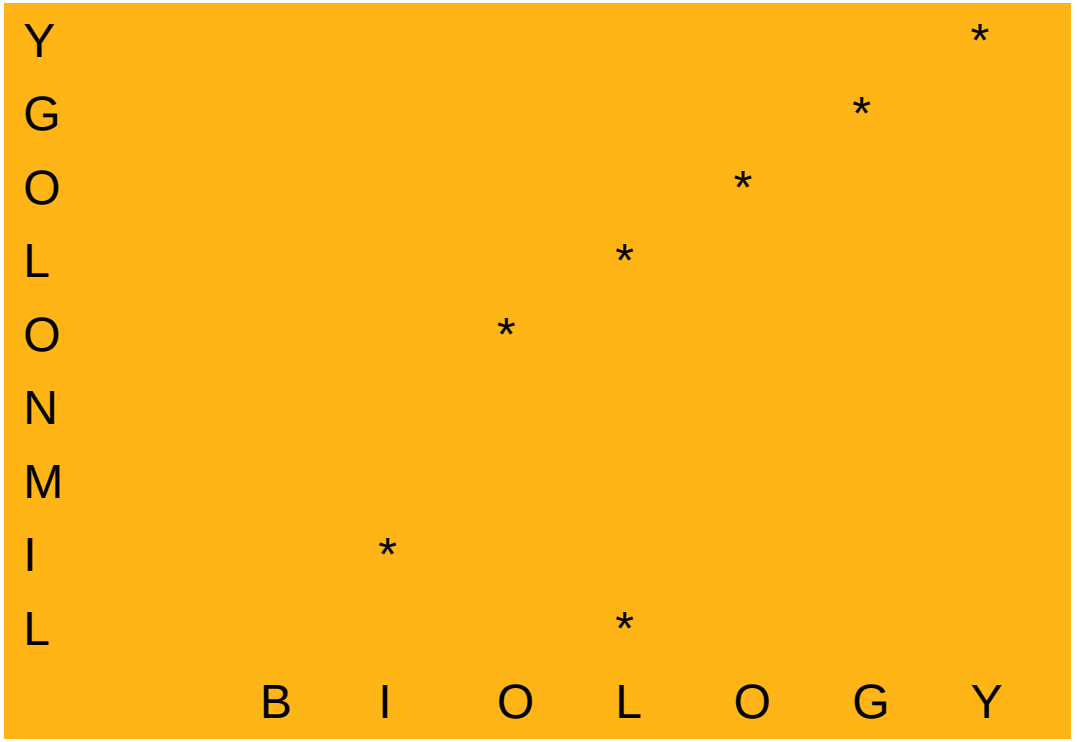
-1+-1+-1+-1+1+1+1+1+1 = 1

Improving the cost function

c(-, a) = c(a,-) = -1		insertion/deletion (indel)
c(-) = -0.2		extension
c(a,b) = -1	when: a≠b	mutation
c(a,b) = 1	when: a=b	

Also known as **cost function:**
Point mutations are costly
Starting indels is **costly**
Extending indels cost less than starting them. (Why?)

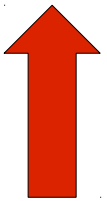
The Dotplot



Back to DNA

- Scoring matrix or substitution matrix describes the cost of mutations

	A	T	G	C	-
A	1	-1	-1	-1	-1
T	-1	1	-1	-1	-1
G	-1	-1	1	-1	-1
C	-1	-1	-1	1	-1
-	-1	-1	-1	-1	N/D



The simplest matrix:
IDENTITY MATRIX

```
Human Chr.21 37218151-37247016,+
Chimp Chr.22 37250843-37279887,+

37218151 GGAGCAATGCACAAGCTCTCTAGTTC------------------------ 37218176
         ||||||||||||||||||||||||||
37250843 GGAGCAATGCACAAGCTCTCTAGTTCGCAAGGTAAAGGTGCGACCTTTCT 37250892

37218177 ------------------------CCCAAGGTAAAGGTGCCGACCTTCCTGGGG 37218206
                                 | ||||||||||||||||||||||||||||
37250893 GGGGGCAGCACGTCCAGTTCCTCAAGGTAAAGGTGCCGACCTTCCTGGGG 37250942

37218207 GCAGCACGTCCAGTTCCCCAAGGTAAAGGTGCCGACCTTCCTAGGGGTGG 37218256
         |||||||||||||||| ||||||||||||||||||||||||||||||||
37250943 GCAGCACGTCCAGTTACCCAAGGTAAAGGTGCCGACCTTCCTAGGGGTGG 37250992

37218257 CACATCCCACCTTCTTCTGCTTCCTCTGGACTCACTGGGGGGACCAAAAG 37218306
         |||||||||||||||||||||||||||| ||||||||||||||||||||
37250993 CACATCCCACCTTCTTCTGCTTCCTCTGGCCTCACTGGGGGGACCAAAAG 37251042

37218307 GGGTCCACTGGCAGGCAACTCTGCAGGGGTGGAGGGTGGGGGGGCTGGCTC 37218356
         |||||||||||||||||||||||||||||||||| ||||||||||||||
37251043 GGGTCCACTGGCAGGCAACTCTGCAGGGGTGGAGGCTGGGGGGCTGGCTC 37251092

37218357 TGCCCCACACACAGATTCAGGTCCCAGCCGGTCCTAACCCCGCCAGCTGCC 37218406
```

Proteins: serious business

- Protein sequences: 20 (21) amino acids
- Some amino acids tend to be more conserved
- Some amino acids substitute others more readily. (Why?)
- Hence: protein substitution matrices are *quantitative* rather than *qualitative*

Scoring matrices for Proteins

- Simulate the evolutionary cost of mutations and indels in proteins
- Generated by a combination of:
  - Observations
  - Statistical model
- Often “repeats” the known biochemistry

PAM Matrix

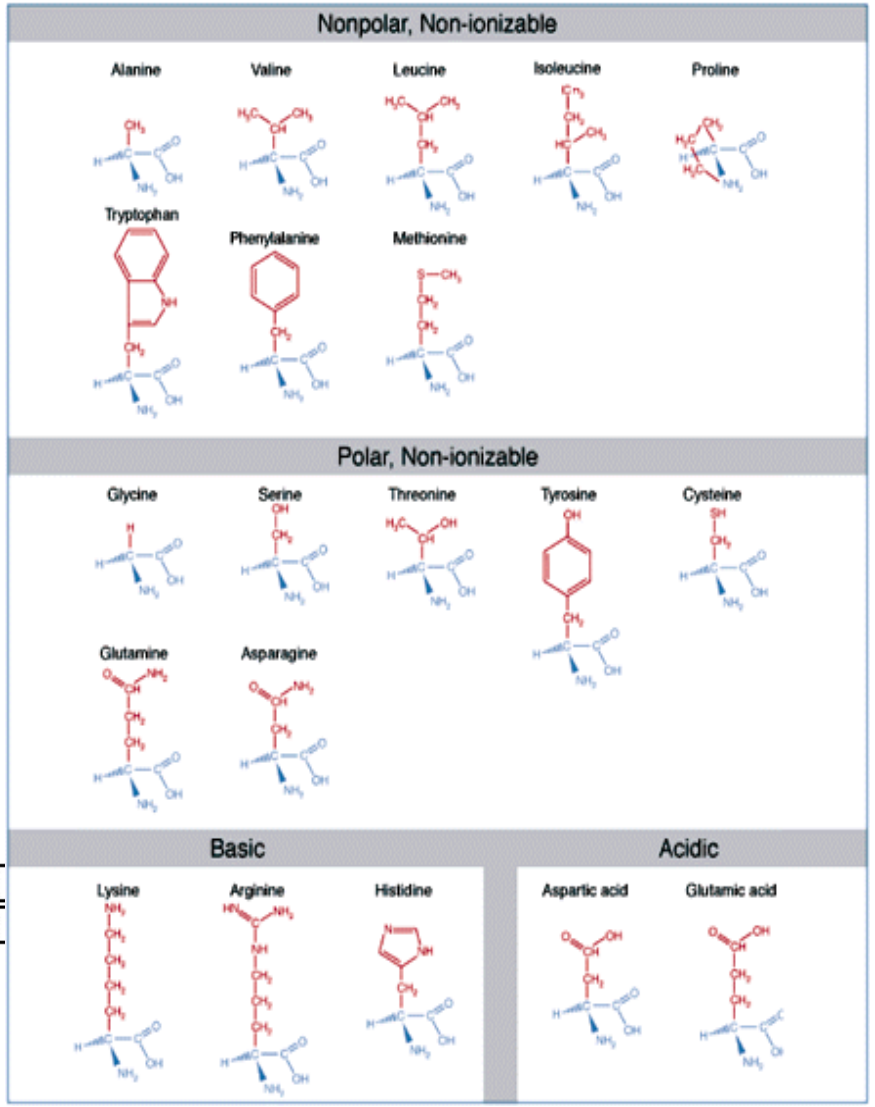
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

WLGYR
WIMCF

17+2-3-0-4
=12

WL-GYR
WIMCF-

17+2-4-3+7-4
=15



Margaret Dayhoff, 1978

PAM Matrix

WLGYR
WIMCF

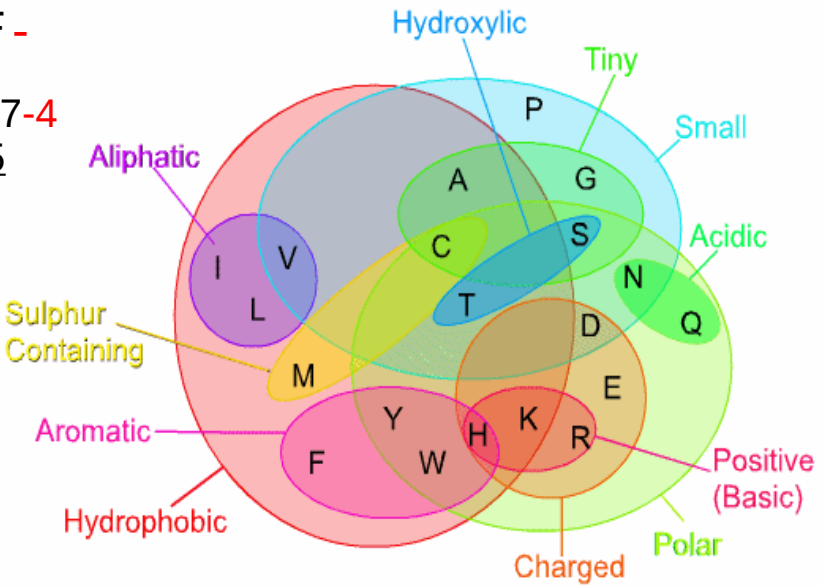
17+2-3-0-4
=12

WL-GYR
WIMCF-

17+2-4-3+7-4
=15

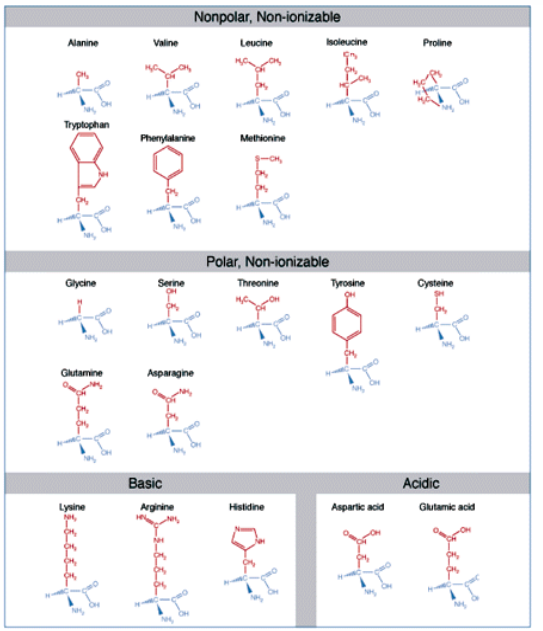
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Margaret Dayhoff, 1978



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)



PAM (Percent Accepted Mutations)

- Based on a database of 1572 changes in 71 groups of closely related proteins
- PAM1: Probability of 1 amino acid out of 100 will mutate

AA Change	PAM1
Phe ↔ Phe	0.9946
Phe ↔ Tyr	0.0021
Phe ↔ Leu	0.0013
Phe ↔ Ala	0.0002
Phe ↔ Arg	0.0001
...	....
SUM	1.000

PAM (Percent Accepted Mutations)

- Based on a database of 1572 changes in 71 groups of closely related proteins
- PAM1: Probability of 1 amino acid out of 100 will mutate
- PAM250 = PAM1 times itself 250 times

AA Change	PAM1	PAM250
Phe ↔ Phe	0.9946	0.32
Phe ↔ Tyr	0.0021	0.15
Phe ↔ Leu	0.0013	0.13
Phe ↔ Ala	0.0002	0.04
Phe ↔ Arg	0.0001	0.01
...	....	...
SUM	1.000	1,00

PAM

- Scoring matrix: 20x20
- Matrix entry M(*i,j*) = probability of *i* substituting *j*
- Diagonally symmetric M(i,j) = M(j,i)
- The higher the PAM number, the higher the evolutionary distance we can assess

PAM: Assumptions and Errors

- **Assumptions**
  - Replacement at any site depends only on the amino acid at that site and the probability given by the table (Markov model).
  - Sequences that are being compared have a representative amino acid composition.
- **Errors**
  - Many sequences depart from average composition.
  - Rare replacements were observed too infrequently to resolve relative probabilities accurately (for 36 pairs no replacements were observed!).
  - Errors in PAM1 are magnified in the extrapolation to PAM 250.
  - Distantly related sequences usually have stretches of conserved residues. This implies that replacement is not equally probable over entire sequence.

BLOSUM Matrix

- Generated from BLOCKS database
- Initial alignments determined by %ID in BLOCK
- BLOSUM85: alignments of 85% ID

```
Block BP00180A
DE   LIGASE SYNTHETASE CARBAMOYL
ACCC_BACSU|P49787   ( 104) GPSADAISKMG  39
ACCC_METJA|Q58626   ( 104) GPNPDAIEAMG  52
CARB_BACSU|P25994   ( 276) GIEGGCNVQLA  57
CARY_BACSU|P18185   ( 954) GTFASWMEQEG  51
DUR1_YEAST|P32528   ( 735) GPSGDIIRGLG  17
PCCA_HUMAN|P05165   (  27) GSVGYDPNEKT 100
PUR2_ECOLI|P15640   (  93) GPTAGAAQLEG  37
PURK_PSEAE|P72158   (   9) GQLGRMLALAG  26
PURT_PASHA|P46927   ( 261) GIFGVELFVCG  46
PYR1_DROME|P05990   (  33) GVGGEVVFQTG  18
SUCC_METJA|Q57663   (  52) GKAGGILFASN  54
YFIQ_ECOLI|P76594   ( 125) NSLGLLAPWQG  59
ACCC_ECOLI|P24182   ( 104) GPKAETIRLMG  27
ACCC_HAEIN|P43873   ( 104) GPTADVIRLMG  23
ACCC_PSEAE|P37798   ( 104) GPTAEVIRLMG  23
CARB_ECOLI|P00968   ( 397) ALRGLEVGATG  28
CPSM_HUMAN|P31327   ( 429) GSGGLSIGQAG   8
CPSM_RAT|P07756     ( 429) GSGGLSIGQAG   8
PYR1_DICDI|P20054   ( 372) GSGGLSIGQAG   8
PYR1_HUMAN|P27708   ( 400) GSGGLSIGQAG   8
PYR1_YEAST|P07259   ( 445) GSGGLSIGQAG   8
COA1_HUMAN|Q13085   (  60) SDLGISALQDG  23
```

Deriving BLOSUM

BABA
AAAC
AACC
AABA
AACC
AABC

aa	Observed (qi) frequency
A	14/24
B	4/24
C	6/24

Mutation type	Observed mutations (qij)	Expected mutations (eij)
A ↔ A	26/60 = 0.433	(14/24)x(14/24) =0.34
A ↔ B	8/60 = 0.1333	(14/24)x(4/24) = 0.0972
A ↔ C		(14/24)x(6/24) = 0.1459
B ↔ B	3/60 = 0.05	
B ↔ C		
C ↔ C		

Deriving BLOSUM

$$q_{ij} = \frac{f_{ij}}{\sum_{i}^{20} (f_{ij})}$$

$$e_{ij} = q_i \times q_j$$

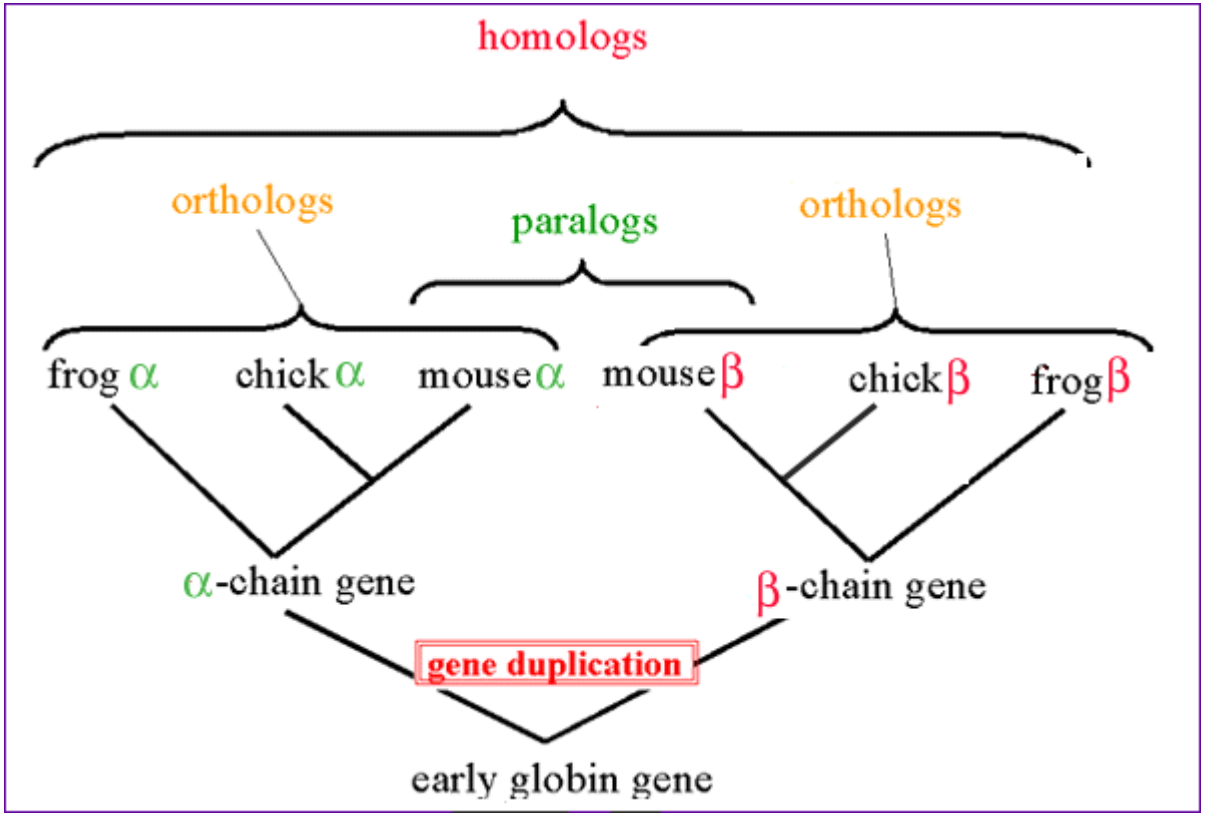
$$M_{ij} = 2\log_2\left(\frac{q_{ij}}{e_{ij}}\right)$$

	A	B	C
A	q_{AA}/e_{AA} = $2\log_2(0.433/0.34)$ = **0.697**		
B			
C			

BLOSUM vs. PAM

	PAM	BLOSUM
Derived from	Full-length proteins	Local blocks
Evolutionary model	Markovian & extrapolation	None implicit

Homology, Orthology, Paralogy



My Brain Hurts...