# Protein bioinformatics: evolution

Tuesday, April 11, 2006

Protein Bioinformatics
260.655
Jonathan Pevsner
pevsner@kennedykrieger.org

## Outline

Sean Prigge described properties of amino acids, and an example of a multiple sequence alignment (globins).

Today we will discuss amino acid properties, and protein relatedness from an evolutionary perspective.

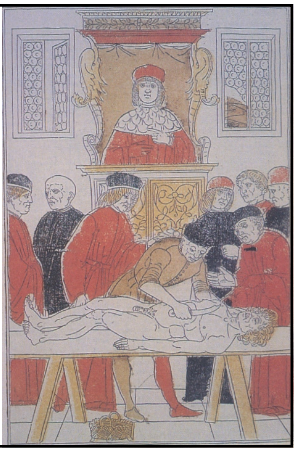
## Outline

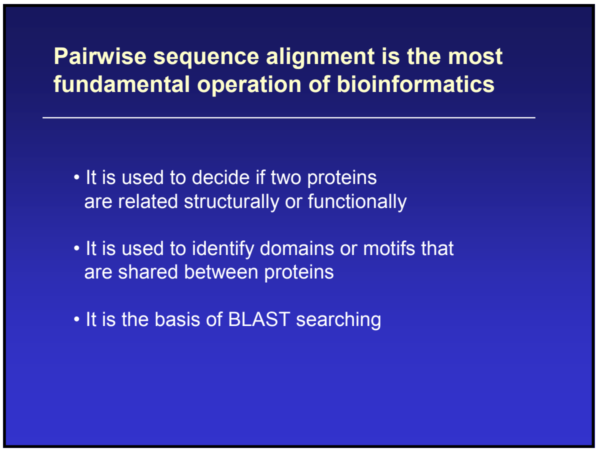
1. Pairwise alignment of proteins

2. Scoring matrices: how related are amino acids?

3. Multiple sequence alignment of proteins

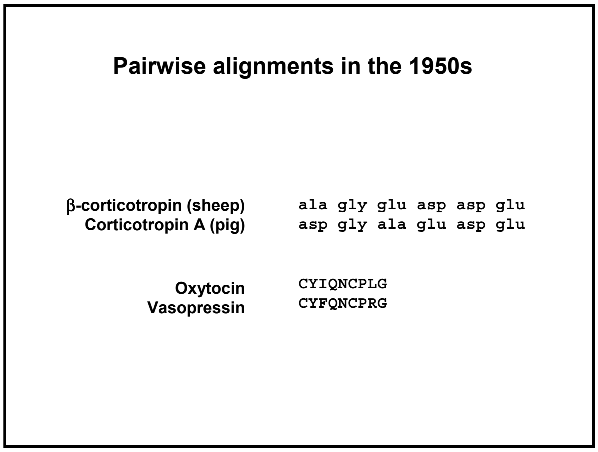
4. From multiple sequence alignment to phylogenetic tree





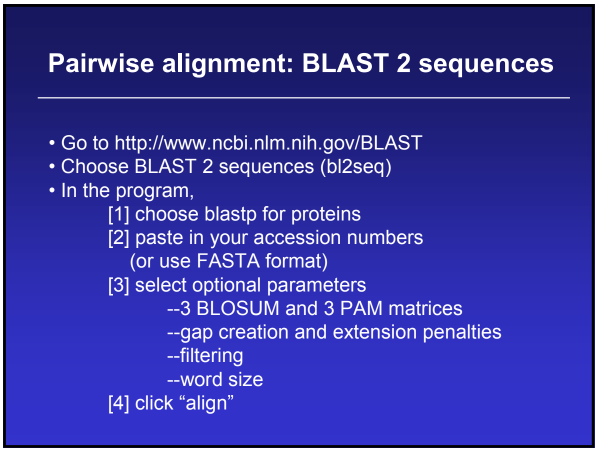
## Pairwise sequence alignment is the most fundamental operation of bioinformatics

- It is used to decide if two proteins are related structurally or functionally
- It is used to identify domains or motifs that are shared between proteins
- It is the basis of BLAST searching



## Pairwise alignments in the 1950s

|  |  |
|---|---|
| **β-corticotropin (sheep)** | `ala gly glu asp asp glu` |
| **Corticotropin A (pig)** | `asp gly ala glu asp glu` |
| **Oxytocin** | `CYIQNCPLG` |
| **Vasopressin** | `CYFQNCPRG` |



## Pairwise alignment: BLAST 2 sequences

- Go to http://www.ncbi.nlm.nih.gov/BLAST
- Choose BLAST 2 sequences (bl2seq)
- In the program,

  [1] choose blastp for proteins

  [2] paste in your accession numbers
  (or use FASTA format)

  [3] select optional parameters

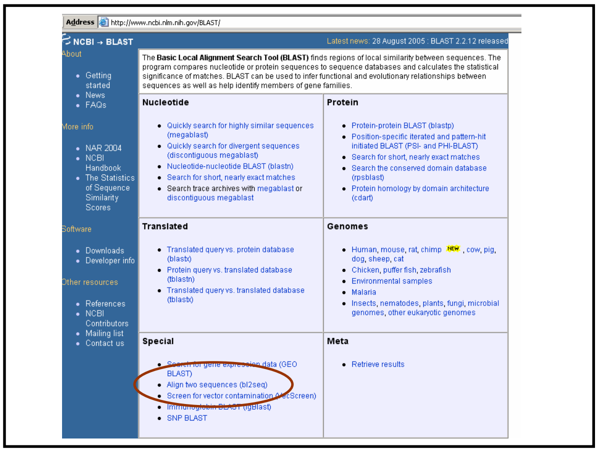
  --3 BLOSUM and 3 PAM matrices

  --gap creation and extension penalties

  --filtering

  --word size

  [4] click “align”



Address http://www.ncbi.nlm.nih.gov/BLAST/

NCBI → BLAST

Latest news: 28 August 2005 : BLAST 2.2.12 released

About

- Getting started
- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

**Nucleotide**

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

**Protein**

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

**Translated**

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

**Genomes**

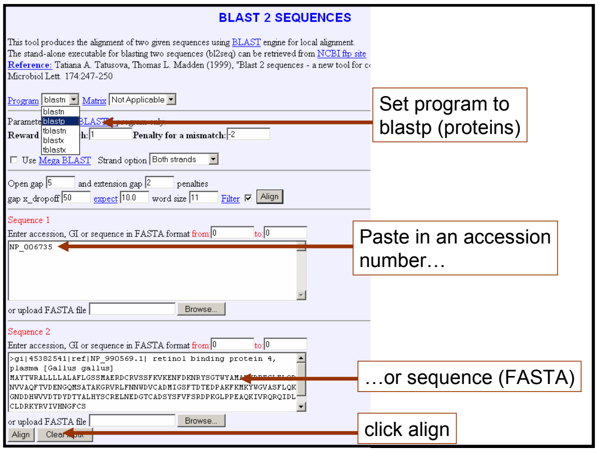
- Human, mouse, rat, chimp NEW , cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Environmental samples
- Malaria
- Insects, nematodes, plants, fungi, microbial genomes, other eukaryotic genomes

**Special**

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (igBlast)
- SNP BLAST

**Meta**

- Retrieve results



## BLAST 2 SEQUENCES

This tool produces the alignment of two given sequences using BLAST engine for local alignment.
The stand-alone executable for blasting two sequences (bl2seq) can be retrieved from NCBI ftp site
**Reference:** Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for c
Microbiol Lett. 174:247-250

Program blastn Matrix Not Applicable

Parameters for BLASTN program only
**Reward for a match:** 1 **Penalty for a mismatch:** -2

□ Use Mega BLAST Strand option Both strands

Open gap 5 and extension gap 2 penalties
gap x_dropoff 50 expect 10.0 word size 11 Filter ☑ Align

Sequence 1
Enter accession, GI or sequence in FASTA format from: 0 to: 0

```
NP_006735
```

or upload FASTA file Browse...

Sequence 2
Enter accession, GI or sequence in FASTA format from: 0 to: 0

```
>gi|45382541|ref|NP_990569.1| retinol binding protein 4,
plasma [Gallus gallus]
MAYTWRALLLLALAFLGSSMAERDCRVSSFKVKENFDKNRYSGTWYAMAKKDPEGLFLQD
NVVAQFTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQK
GNDDHWVVDTDYDTYALHYSCRELNEDGTCADSYSFVFSRDPKGLPPEAQKIVRQRQIDL
CLDRKYRVIVHNGFCS
```

or upload FASTA file Browse...

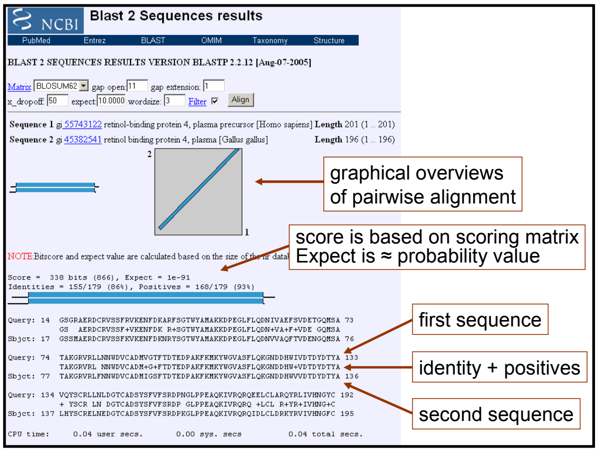
Align Clear Input

Set program to blastp (proteins)

Paste in an accession number…

…or sequence (FASTA)

click align



NCBI **Blast 2 Sequences results**

BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.12 [Aug-07-2005]

Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 10.0000 wordsize: 3 Filter ☑ Align

**Sequence 1** gi 55743122 retinol-binding protein 4, plasma precursor [Homo sapiens] **Length** 201 (1 .. 201)
**Sequence 2** gi 45382541 retinol binding protein 4, plasma [Gallus gallus] **Length** 196 (1 .. 196)

NOTE:Bitscore and expect value are calculated based on the size of the nr datab

```
Score =  338 bits (866), Expect = 1e-91
Identities = 155/179 (86%), Positives = 168/179 (93%)

Query: 14   GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSA 73
             GS  AERDCRVSSF+VKENFDK R+SGTWYAMAKKDPEGLFLQDN+VA+F+VDE GQMSA
Sbjct: 17   GSSMAERDCRVSSFKVKENFDKNRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGQMSA 76

Query: 74   TAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYA 133
             TAKGRVRL NNWDVCADM+G+FTDTEDPAKFKMKYWGVASFLQKGNDDHW+VDTDYDTYA
Sbjct: 77   TAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDTYA 136

Query: 134 VQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYC 192
             + YSCR LN DGTCADSYSFVFSRDP GLPPEAQKIVRQRQ +LCL R+YR+IVHNG+C
Sbjct: 137 LHYSCRELNEDGTCADSYSFVFSRDPKGLPPEAQKIVRQRQIDLCLDRKYRVIVHNGFC 195

CPU time:     0.04 user secs.      0.00 sys. secs        0.04 total secs.
```

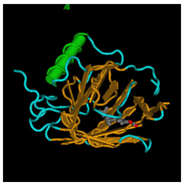
graphical overviews of pairwise alignment

score is based on scoring matrix
Expect is ≈ probability value

first sequence

identity + positives

second sequence



retinol-binding protein 4
(NP_006735)

β-lactoglobulin
(P02754)

## Definitions

**Pairwise alignment**
The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

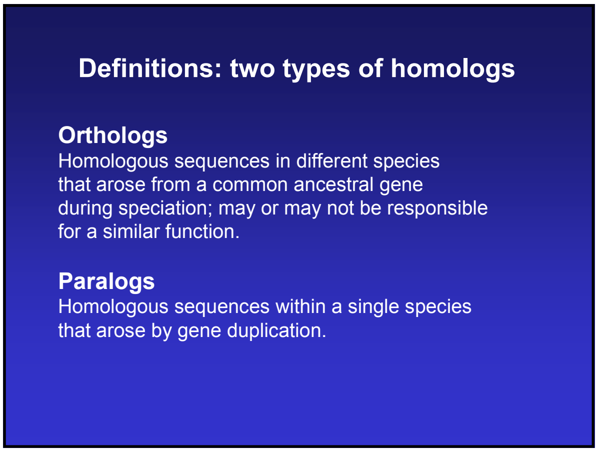
## Definitions

**Homology**
Similarity attributed to descent from a common ancestor.

**Identity**

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

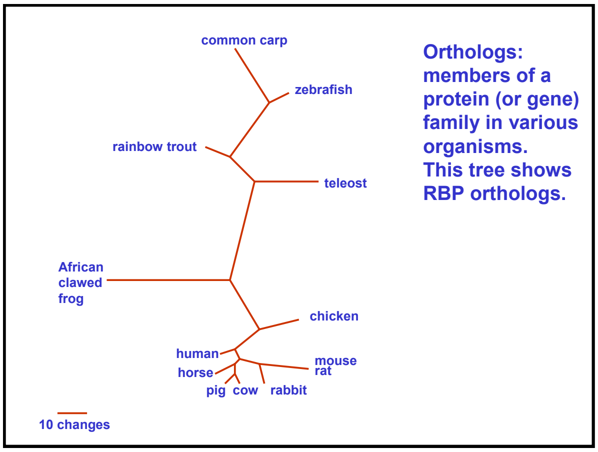
```
RBP:        26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVA 59
                + K++ + ++  GTW++MA   +    L +   A
glycodelin: 23  QTKQDLELPKLAGTWHSMAMA-TNNISLMATLKA 55
```



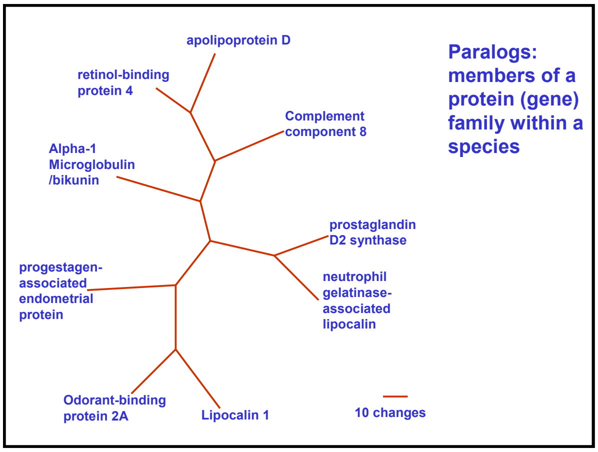
## Definitions: two types of homologs

**Orthologs**
Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

**Paralogs**
Homologous sequences within a single species that arose by gene duplication.

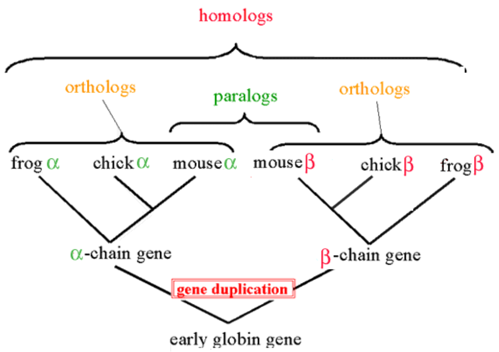


**Orthologs: members of a protein (or gene) family in various organisms. This tree shows RBP orthologs.**



apolipoprotein D
retinol-binding protein 4
Complement component 8
Alpha-1 Microglobulin /bikunin
prostaglandin D2 synthase
neutrophil gelatinase-associated lipocalin
progestagen-associated endometrial protein
Odorant-binding protein 2A
Lipocalin 1
10 changes

**Paralogs: members of a protein (gene) family within a species**



Source: NCBI website

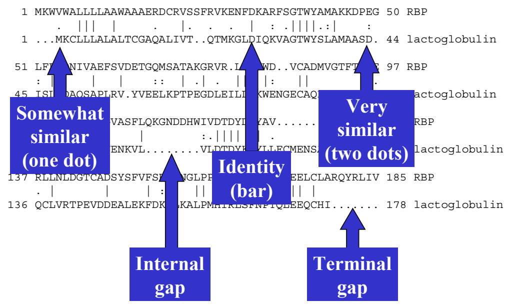
## Definitions

**Similarity**
The extent to which nucleotide or protein sequences are related. It is based upon identity plus conservation.

**Identity**
The extent to which two sequences are invariant.

**Conservation**
Changes at a specific position of an amino acid sequence that preserve the physico-chemical properties of the original residue.

## Pairwise alignment of retinol-binding protein 4 and β-lactoglobulin: explaining the dots and dashes



```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
       .  |||  |      .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LF  NIVAEFSVDETGQMSATAKGRVR.L  WD..VCADMVGTFT  E 97 RBP
    : |     |   |     ::   | .| .  ||  |:    ||        .
 45 ISL DAQSAPLRV.YVEELKPTPEGDLEILL  KWENGECAQ            actoglobulin
                       VASFLQKGNDDHWIVDTDYD  YAV......          RBP
                            |         :.||||  |  .
                       ENKVL........VLDTDYK  YLLFCMENSA         lactoglobulin
137 RLLNLDGTCADSYSFVFS  NGLPP           EELCLARQYRLIV 185 RBP
     . |        |      | :    ||            || |
136 QCLVRTPEVDDEALEKFDK  KALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```

Somewhat similar (one dot)

Identity (bar)

Very similar (two dots)

Internal gap

Terminal gap

## Gaps

- Positions at which a letter is paired with a null are called gaps.
- Gap scores are typically negative.
- Since a single mutational event may cause the insertion or deletion of more than one residue, the presence of a gap is ascribed more significance than the length of the gap.
- In BLAST, it is rarely necessary to change gap values from the default.

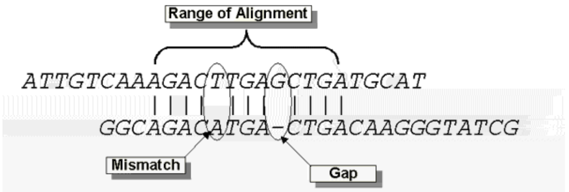
## Pairwise alignment of retinol-binding protein from human (top) and rainbow trout (*O. mykiss*): two closely related proteins

```
  1 .MKWVWALLLLA.AWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 48
     ::   ||  ||  ||   .||.||. .| :|||:.|:.| |||.|||||
  1 MLRICVALCALATCWA...QDCQVSNIQVMQNFDRSRYTGRWYAVAKKDP 47
             .         .         .         .         .
 49 EGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTED 98
     |||| ||:||:|||||.|.|.||| ||| :||||:.||.| ||| || |
 48 VGLFLLDNVVAQFSVDESGKMTATAHGRVIILNNWEMCANMFGTFEDTPD 97
             .         .         .         .         .
 99 PAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADS 148
    ||||||:||| ||:|| ||||||::||||| ||: |||| ..||||| |
 98 PAKFKMRYWGAASYLQTGNDDHWVIDTDYDNYAIHYSCREVDLDGTCLDG 147
             .         .         .         .         .
149 YSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL 199
    |||:||| | || || ||||  :..|:|   .|| : | |:|:
148 YSFIFSRHPTGLRPEDQKIVTDKKKEICFLGKYRRVGHTGFCESS...... 192
```

## General approach to pairwise alignment

- Choose two sequences
- Select an algorithm that generates a score
- Allow gaps (insertions, deletions)
- Score reflects degree of similarity
- Alignments can be global (Needleman and Wunsch, 1970) or local (Smith and Waterman, 1981)
- Estimate probability that the alignment occurred by chance

## Calculation of an alignment score



$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$

$\text{Score} = \text{Max}(S)$

Source: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Scores2.html

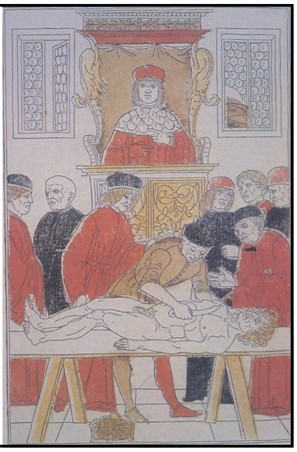
## Outline

1. Pairwise alignment of proteins

2. Scoring matrices: how related are amino acids?

3. Multiple sequence alignment of proteins

4. From multiple sequence alignment to phylogenetic tree



## How do we decide what scores to assign in pairwise alignments?

- Zuckerkandl and Pauling (1965) made a multiple sequence alignment of hemoglobin and myoglobin from primates, horse, cattle, pig, lamprey, and carp. They made a “scoring matrix.”

- Margaret Dayhoff and colleagues (1960s, 1970s) studied dozens of families of proteins to create scoring matrices that describe the relationship of well-conserved (or poorly-conserved) protein families.

## Multiple sequence alignment of glyceraldehyde 3-phosphate dehydrogenases

```
fly       GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human     GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant     GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly       KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human     KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant     KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast     KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon  KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human     GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant     GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast     GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon  GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

**Studying conserved (and nonconserved) residues in closely related families may reveal “rules” for amino acid substitutions accepted by natural selection**

## Multiple sequence alignment of human lipocalin paralogs

```
~~~~~EIQDVSGTWYAMTVDREFPEMNLESVTPMTLTTL.GGNLEAKVTM    lipocalin 1
LSFTLEEEDITGTWYAMVVDKDFPEDRRRKVSPVKVTALGGGNLEATFTF    odorant-binding protein 2a
TKQDLELPKLAGTWHSMAMATNNISLMATLKAPLRVHITSEDNLEIVLHR    progestagen-assoc. endo.
VQENFDVNKYLGRWYEIEKIPTTFENGRCIQANYSLMENGNQELRADGTV    apolipoprotein D
VKENFDKARFSGTWYAMAKDPEGLFLQDNIVAEFSVDETGNWDVCADGTF    retinol-binding protein
LQQNFQDNQFQGKWYVVGLAGNAI.LREDKDPQKMYATIDKSYNVTSVLF    neutrophil gelatinase-ass.
VQPNFQQDKFLGRWFSAGLASNSSWLREKKAALSMCKSVDGGLNLTSTFL    prostaglandin D2 synthase
VQENFNISRIYGKWYNLAIGSTCPWMDRMTVSTLVLGEGEAEISMTSTRW    alpha-1-microglobulin
PKANFDAQQFAGTWLLVAVGSACRFLQRAEATTLHVAPQGSTFRKLD...    complement component 8
```

**Studying conserved (and nonconserved) residues in distantly related families is also informative**

## Substituent residue

(percentage of total residue sites at which the substituent occurs)

Sequence (original amino acid)

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | ■ | | | 28 | | | 31 | 33 | | | | | | | | 31 | | | | |
| R | | ■ | | | 🟥 | | | 🟥 | 50 | | | 58 | 🟥 | 🟥 | 🟥 | 25 | | 🟥 | 🟥 | |
| N | 33 | | ■ | 47 | 🟥 | | | | 33 | 🟥 | | 33 | 🟥 | 🟥 | | 33 | 33 | 🟥 | | |
| D | 44 | | 22 | ■ | 🟥 | | 47 | 34 | 22 | | | 28 | 🟥 | 🟥 | | 25 | | 🟥 | 🟥 | |
| C | (66) | 🟥 | 🟥 | 🟥 | ■ | 🟥 | 🟥 | 🟥 | | 🟥 | | | 🟥 | | 🟥 | | | 🟥 | 🟥 | |
| Q | | | | 56 | 🟥 | ■ | 30 | 🟥 | 40 | 🟥 | | 70 | 🟥 | 🟥 | | | | 🟥 | 🟥 | |
| E | 50 | | | 44 | 🟥 | | ■ | 38 | | | | 41 | 🟥 | 🟥 | 24 | | | | 🟥 | |
| G | 51 | 🟥 | | 33 | 🟥 | 🟥 | 30 | ■ | | | | 27 | 🟥 | 🟥 | | 36 | | | | |
| H | | | | 26 | | | | | ■ | | 26 | 30 | | | | 22 | 22 | 🟥 | | |
| I | 39 | | 🟥 | | 🟥 | 🟥 | | | | ■ | 58 | | | 🟥 | | | | 🟥 | 🟥 | 46 |
| L | 21 | | | | | | | | | 23 | ■ | 23 | | 28 | 🟥 | | | | | 30 |
| K | 23 | 21 | | 28 | | | 31 | 23 | | | 21 | ■ | | | | 21 | | | | |
| M | 22 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | | 22 | 89 | | ■ | 22 | 🟥 | 🟥 | | | | 45 |
| F | | 🟥 | 🟥 | 🟥 | | 🟥 | 🟥 | 🟥 | 22 | 🟥 | 61 | | | ■ | 🟥 | | | | | |
| P | 50 | 🟥 | | 43 | 🟥 | | 57 | 43 | | | 🟥 | 21 | 🟥 | 🟥 | ■ | | | 🟥 | 🟥 | 🟥 |
| S | 49 | | | 24 | | | 24 | 36 | | | | 24 | 🟥 | | | ■ | 40 | 🟥 | 🟥 | |
| T | 32 | | | | | | 28 | 24 | | | | 24 | | | | 52 | ■ | 🟥 | 🟥 | |
| W | (40) | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | | | 🟥 | 🟥 | (40) | | | (60) | 🟥 | 🟥 | 🟥 | ■ | | 🟥 |
| Y | | 🟥 | | 🟥 | 🟥 | 🟥 | 🟥 | | (33) | 🟥 | | | | (50) | 🟥 | 🟥 | 🟥 | | ■ | |
| V | 36 | | | | | | | | | 21 | 43 | 21 | | | 🟥 | | | 🟥 | | ■ |

🟥 substitution never observed
□ substitution rarely observed (<20%)
▨ very conservative substitution (>40%)

## PAM matrices:<br>Point-accepted mutations

PAM matrices are based on global alignments of closely related proteins.

The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.

Other PAM matrices are extrapolated from PAM1.

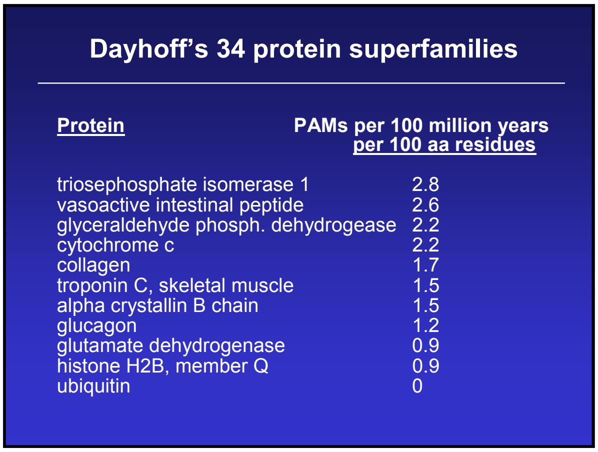
All the PAM data come from closely related proteins (>85% amino acid identity)

## Dayhoff's 34 protein superfamilies

| **Protein** | **PAMs per 100 million years per 100 aa residues** |
|---|---|
| Ig kappa chain | 37 |
| kappa casein | 33 |
| luteinizing hormone b | 30 |
| lactalbumin | 27 |
| complement component 3 | 27 |
| epidermal growth factor | 26 |
| proopiomelanocortin | 21 |
| pancreatic ribonuclease | 21 |
| haptoglobin alpha | 20 |
| serum albumin | 19 |
| phospholipase A2, group IB | 19 |
| prolactin | 17 |
| carbonic anhydrase C | 16 |
| hemoglobin α | 12 |
| hemoglobin β | 12 |

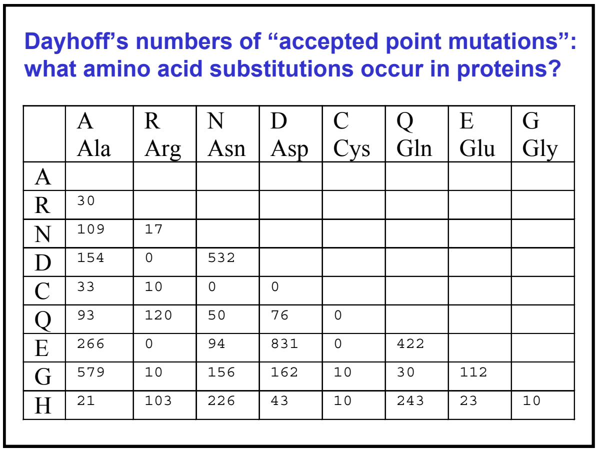
## Dayhoff's 34 protein superfamilies

| **Protein** | **PAMs per 100 million years per 100 aa residues** |
|---|---|
| apolipoprotein A-II | 10 |
| lysozyme | 9.8 |
| gastrin | 9.8 |
| myoglobin | 8.9 |
| nerve growth factor | 8.5 |
| myelin basic protein | 7.4 |
| thyroid stimulating hormone b | 7.4 |
| parathyroid hormone | 7.3 |
| parvalbumin | 7.0 |
| trypsin | 5.9 |
| insulin | 4.4 |
| calcitonin | 4.3 |
| arginine vasopressin | 3.6 |
| adenylate kinase 1 | 3.2 |



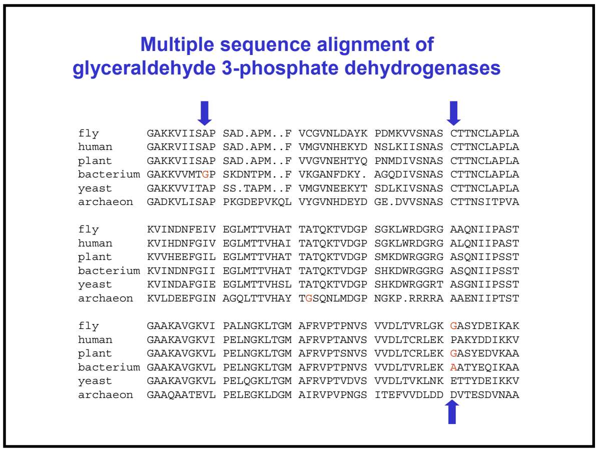
## Dayhoff’s 34 protein superfamilies

| Protein | PAMs per 100 million years per 100 aa residues |
|---|---|
| triosephosphate isomerase 1 | 2.8 |
| vasoactive intestinal peptide | 2.6 |
| glyceraldehyde phosph. dehydrogease | 2.2 |
| cytochrome c | 2.2 |
| collagen | 1.7 |
| troponin C, skeletal muscle | 1.5 |
| alpha crystallin B chain | 1.5 |
| glucagon | 1.2 |
| glutamate dehydrogenase | 0.9 |
| histone H2B, member Q | 0.9 |
| ubiquitin | 0 |



## Dayhoff’s numbers of “accepted point mutations”: what amino acid substitutions occur in proteins?

| | A<br>Ala | R<br>Arg | N<br>Asn | D<br>Asp | C<br>Cys | Q<br>Gln | E<br>Glu | G<br>Gly |
|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | |
| R | 30 | | | | | | | |
| N | 109 | 17 | | | | | | |
| D | 154 | 0 | 532 | | | | | |
| C | 33 | 10 | 0 | 0 | | | | |
| Q | 93 | 120 | 50 | 76 | 0 | | | |
| E | 266 | 0 | 94 | 831 | 0 | 422 | | |
| G | 579 | 10 | 156 | 162 | 10 | 30 | 112 | |
| H | 21 | 103 | 226 | 43 | 10 | 243 | 23 | 10 |



## Multiple sequence alignment of glyceraldehyde 3-phosphate dehydrogenases

```
fly       GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human     GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant     GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly       KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human     KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant     KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast     KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon  KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human     GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant     GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast     GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon  GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

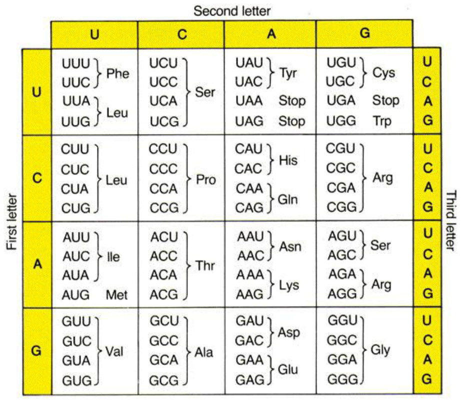
## The relative mutability of amino acids

| | | | |
|---|---|---|---|
| Asn | 134 | His | 66 |
| Ser | 120 | Arg | 65 |
| Asp | 106 | Lys | 56 |
| Glu | 102 | Pro | 56 |
| Ala | 100 | Gly | 49 |
| Thr | 97 | Tyr | 41 |
| Ile | 96 | Phe | 41 |
| Met | 94 | Leu | 40 |
| Gln | 93 | Cys | 20 |
| Val | 74 | Trp | 18 |

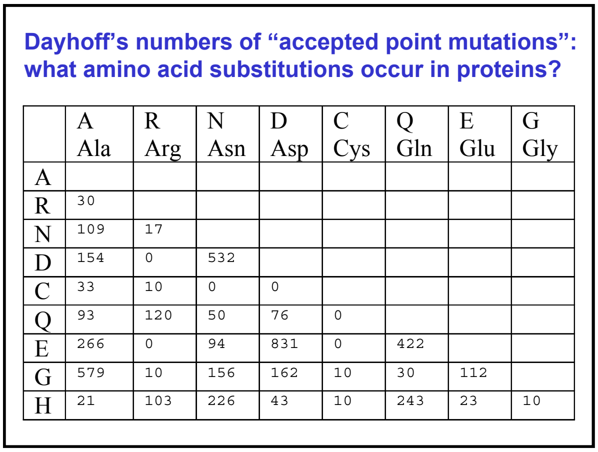
## Normalized frequencies of amino acids

| | | | |
|---|---|---|---|
| Gly | 8.9% | Arg* | 4.1% |
| Ala | 8.7% | Asn | 4.0% |
| Leu* | 8.5% | Phe | 4.0% |
| Lys | 8.1% | Gln | 3.8% |
| Ser* | 7.0% | Ile | 3.7% |
| Val | 6.5% | His | 3.4% |
| Thr | 5.8% | Cys | 3.3% |
| Pro | 5.1% | Tyr | 3.0% |
| Glu | 5.0% | Met† | 1.5% |
| Asp | 4.7% | Trp† | 1.0% |

blue*=6 codons; red†=1 codon

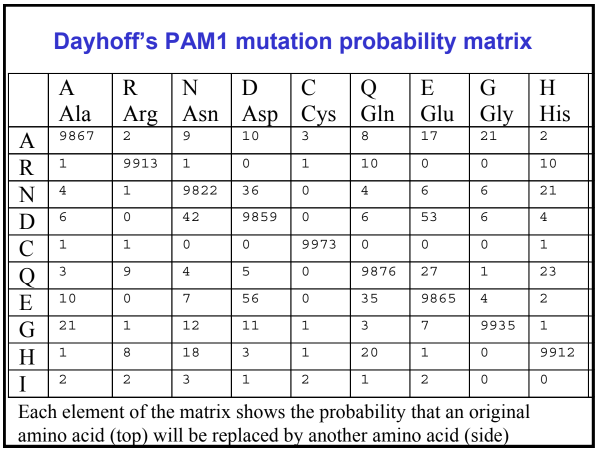


| | Second letter | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| First letter | U | | C | | A | | G | | Third letter |
| U | UUU | Phe | UCU | Ser | UAU | Tyr | UGU | Cys | U |
| | UUC | | UCC | | UAC | | UGC | | C |
| | UUA | Leu | UCA | | UAA | Stop | UGA | Stop | A |
| | UUG | | UCG | | UAG | Stop | UGG | Trp | G |
| C | CUU | Leu | CCU | Pro | CAU | His | CGU | Arg | U |
| | CUC | | CCC | | CAC | | CGC | | C |
| | CUA | | CCA | | CAA | Gln | CGA | | A |
| | CUG | | CCG | | CAG | | CGG | | G |
| A | AUU | Ile | ACU | Thr | AAU | Asn | AGU | Ser | U |
| | AUC | | ACC | | AAC | | AGC | | C |
| | AUA | | ACA | | AAA | Lys | AGA | Arg | A |
| | AUG | Met | ACG | | AAG | | AGG | | G |
| G | GUU | Val | GCU | Ala | GAU | Asp | GGU | Gly | U |
| | GUC | | GCC | | GAC | | GGC | | C |
| | GUA | | GCA | | GAA | Glu | GGA | | A |
| | GUG | | GCG | | GAG | | GGG | | G |



## Dayhoff’s numbers of “accepted point mutations”: what amino acid substitutions occur in proteins?

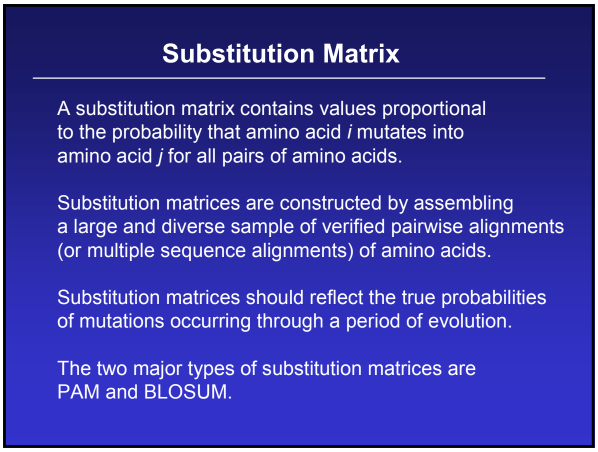
| | A<br>Ala | R<br>Arg | N<br>Asn | D<br>Asp | C<br>Cys | Q<br>Gln | E<br>Glu | G<br>Gly |
|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | |
| R | 30 | | | | | | | |
| N | 109 | 17 | | | | | | |
| D | 154 | 0 | 532 | | | | | |
| C | 33 | 10 | 0 | 0 | | | | |
| Q | 93 | 120 | 50 | 76 | 0 | | | |
| E | 266 | 0 | 94 | 831 | 0 | 422 | | |
| G | 579 | 10 | 156 | 162 | 10 | 30 | 112 | |
| H | 21 | 103 | 226 | 43 | 10 | 243 | 23 | 10 |



## Dayhoff’s PAM1 mutation probability matrix

| | A<br>Ala | R<br>Arg | N<br>Asn | D<br>Asp | C<br>Cys | Q<br>Gln | E<br>Glu | G<br>Gly | H<br>His |
|---|---|---|---|---|---|---|---|---|---|
| A | 9867 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 |
| R | 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 10 |
| N | 4 | 1 | 9822 | 36 | 0 | 4 | 6 | 6 | 21 |
| D | 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 |
| C | 1 | 1 | 0 | 0 | 9973 | 0 | 0 | 0 | 1 |
| Q | 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 23 |
| E | 10 | 0 | 7 | 56 | 0 | 35 | 9865 | 4 | 2 |
| G | 21 | 1 | 12 | 11 | 1 | 3 | 7 | 9935 | 1 |
| H | 1 | 8 | 18 | 3 | 1 | 20 | 1 | 0 | 9912 |
| I | 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 |

Each element of the matrix shows the probability that an original amino acid (top) will be replaced by another amino acid (side)



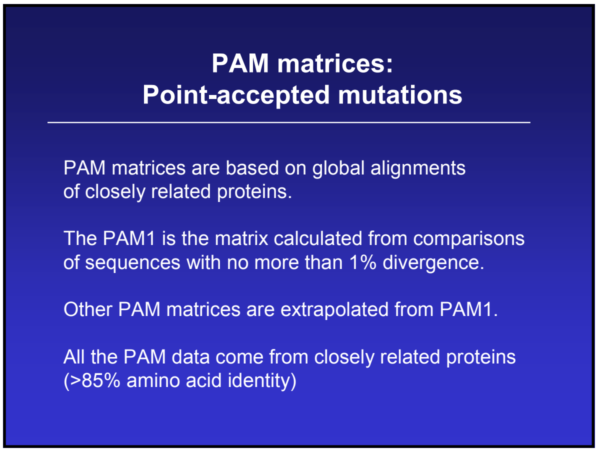
## Substitution Matrix

A substitution matrix contains values proportional to the probability that amino acid *i* mutates into amino acid *j* for all pairs of amino acids.

Substitution matrices are constructed by assembling a large and diverse sample of verified pairwise alignments (or multiple sequence alignments) of amino acids.

Substitution matrices should reflect the true probabilities of mutations occurring through a period of evolution.

The two major types of substitution matrices are PAM and BLOSUM.



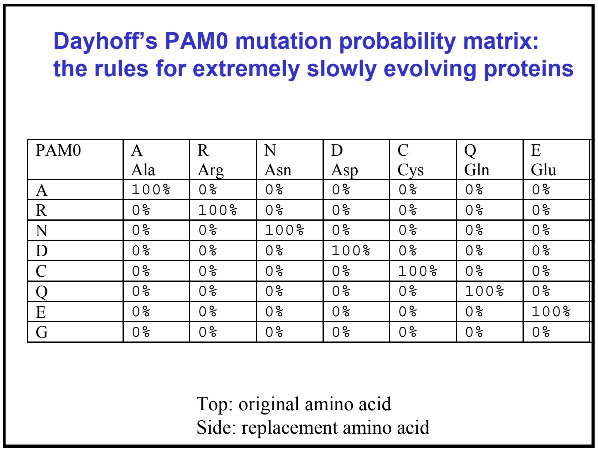
## PAM matrices: Point-accepted mutations

PAM matrices are based on global alignments of closely related proteins.

The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.

Other PAM matrices are extrapolated from PAM1.

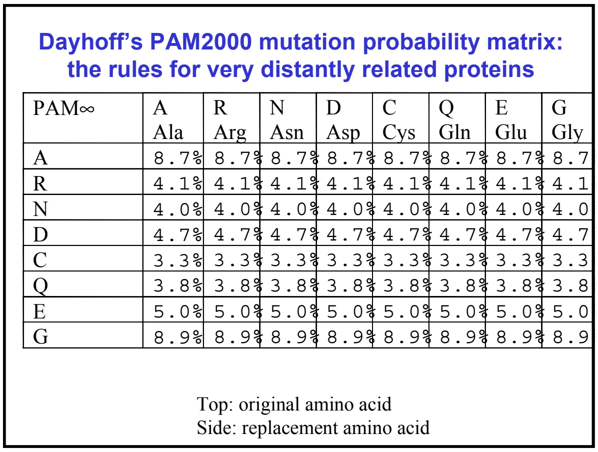
All the PAM data come from closely related proteins (>85% amino acid identity)



## Dayhoff's PAM0 mutation probability matrix: the rules for extremely slowly evolving proteins

| PAM0 | A<br>Ala | R<br>Arg | N<br>Asn | D<br>Asp | C<br>Cys | Q<br>Gln | E<br>Glu |
|---|---|---|---|---|---|---|---|
| A | 100% | 0% | 0% | 0% | 0% | 0% | 0% |
| R | 0% | 100% | 0% | 0% | 0% | 0% | 0% |
| N | 0% | 0% | 100% | 0% | 0% | 0% | 0% |
| D | 0% | 0% | 0% | 100% | 0% | 0% | 0% |
| C | 0% | 0% | 0% | 0% | 100% | 0% | 0% |
| Q | 0% | 0% | 0% | 0% | 0% | 100% | 0% |
| E | 0% | 0% | 0% | 0% | 0% | 0% | 100% |
| G | 0% | 0% | 0% | 0% | 0% | 0% | 0% |

Top: original amino acid
Side: replacement amino acid



## Dayhoff's PAM2000 mutation probability matrix: the rules for very distantly related proteins

| PAM∞ | A<br>Ala | R<br>Arg | N<br>Asn | D<br>Asp | C<br>Cys | Q<br>Gln | E<br>Glu | G<br>Gly |
|---|---|---|---|---|---|---|---|---|
| A | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7 |
| R | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1 |
| N | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0 |
| D | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7 |
| C | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3 |
| Q | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8 |
| E | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0 |
| G | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9 |

Top: original amino acid
Side: replacement amino acid

## PAM250 mutation probability matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 13 | 6 | 9 | 9 | 5 | 8 | 9 | 12 | 6 | 8 | 6 | 7 | 7 | 4 | 11 | 11 | 11 | 2 | 4 | 9 |
| R | 3 | 17 | 4 | 3 | 2 | 5 | 3 | 2 | 6 | 3 | 2 | 9 | 4 | 1 | 4 | 4 | 3 | 7 | 2 | 2 |
| N | 4 | 4 | 6 | 7 | 2 | 5 | 6 | 4 | 6 | 3 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | 2 | 3 | 3 |
| D | 5 | 4 | 8 | 11 | 1 | 7 | 10 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| C | 2 | 1 | 1 | 1 | 52 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 2 |
| Q | 3 | 5 | 5 | 6 | 1 | 10 | 7 | 3 | 7 | 2 | 3 | 5 | 3 | 1 | 4 | 3 | 3 | 1 | 2 | 3 |
| E | 5 | 4 | 7 | 11 | 1 | 9 | 12 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| G | 12 | 5 | 10 | 10 | 4 | 7 | 9 | 27 | 5 | 5 | 4 | 6 | 5 | 3 | 8 | 11 | 9 | 2 | 3 | 7 |
| H | 2 | 5 | 5 | 4 | 2 | 7 | 4 | 2 | 15 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 2 |
| I | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 10 | 6 | 2 | 6 | 5 | 2 | 3 | 4 | 1 | 3 | 9 |
| L | 6 | 4 | 4 | 3 | 2 | 6 | 4 | 3 | 5 | 15 | 34 | 4 | 20 | 13 | 5 | 4 | 6 | 6 | 7 | 13 |
| K | 6 | 18 | 10 | 8 | 2 | 10 | 8 | 5 | 8 | 5 | 4 | 24 | 9 | 2 | 6 | 8 | 8 | 4 | 3 | 5 |
| M | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 6 | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| F | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 5 | 6 | 1 | 4 | 32 | 1 | 2 | 2 | 4 | 20 | 3 |
| P | 7 | 5 | 5 | 4 | 3 | 5 | 4 | 5 | 5 | 3 | 3 | 4 | 3 | 2 | 20 | 6 | 5 | 1 | 2 | 4 |
| S | 9 | 6 | 8 | 7 | 7 | 6 | 7 | 9 | 6 | 5 | 4 | 7 | 5 | 3 | 9 | 10 | 9 | 4 | 4 | 6 |
| T | 8 | 5 | 6 | 6 | 4 | 5 | 5 | 6 | 4 | 6 | 4 | 6 | 5 | 3 | 6 | 8 | 11 | 2 | 3 | 6 |
| W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 55 | 1 | 0 |
| Y | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 15 | 1 | 2 | 2 | 3 | 31 | 2 |
| V | 7 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 15 | 10 | 4 | 10 | 5 | 5 | 5 | 7 | 2 | 4 | 17 |

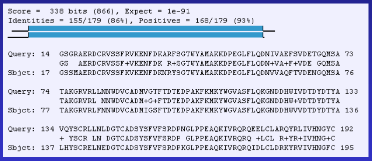
Top: original amino acid
Side: replacement amino acid

## PAM250 log odds scoring matrix

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 2 | | | | | | | | | | | | | | | | | | | |
| **R** | -2 | 6 | | | | | | | | | | | | | | | | | | |
| **N** | 0 | 0 | 2 | | | | | | | | | | | | | | | | | |
| **D** | 0 | -1 | 2 | 4 | | | | | | | | | | | | | | | | |
| **C** | -2 | -4 | -4 | -5 | 12 | | | | | | | | | | | | | | | |
| **Q** | 0 | 1 | 1 | 2 | -5 | 4 | | | | | | | | | | | | | | |
| **E** | 0 | -1 | 1 | 3 | -5 | 2 | 4 | | | | | | | | | | | | | |
| **G** | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | | | | | | | | | | | | |
| **H** | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | | | | | | | | | | | |
| **I** | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | | | | | | | | | | |
| **L** | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | -2 | 6 | | | | | | | | | |
| **K** | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | | | | | | | | |
| **M** | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | | | | | | | |
| **F** | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | | | | | | |
| **P** | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | | | | | |
| **S** | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | | | | |
| **T** | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | | | |
| **W** | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | | |
| **Y** | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | |
| **V** | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |
| | **A** | **R** | **N** | **D** | **C** | **Q** | **E** | **G** | **H** | **I** | **L** | **K** | **M** | **F** | **P** | **S** | **T** | **W** | **Y** | **V** |

## Why do we go from a mutation probability matrix to a log odds matrix?

- We want a scoring matrix so that when we do a pairwise alignment (or a BLAST search) we know what score to assign to two aligned amino acid residues.



```
Score =  338 bits (866), Expect = 1e-91
Identities = 155/179 (86%), Positives = 168/179 (93%)

Query: 14  GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSA 73
            GS  AERDCRVSSF+VKENFDK R+SGTWYAMAKKDPEGLFLQDN+VA+F+VDE GQMSA
Sbjct: 17  GSSMAERDCRVSSFKVKENFDKNRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGQMSA 76

Query: 74  TAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYA 133
            TAKGRVRL NNWDVCADM+G+FTDTEDPAKFKMKYWGVASFLQKGNDDHW+VDTDYDTYA
Sbjct: 77  TAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDTYA 136

Query: 134 VQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYC 192
            + YSCR LN DGTCADSYSFVFSRDP GLPPEAQKIVRQRQ +LCL R+YR+IVHNG+C
Sbjct: 137 LHYSCRELNEDGTCADSYSFVFSRDPKGLPPEAQKIVRQRQIDLCLDRKYRVIVHNGFC 195
```

- Logarithms are easier to use for a scoring system. They allow us to sum the scores of aligned residues (rather than having to multiply them).

## How do we go from a mutation probability matrix to a log odds matrix?

- The cells in a log odds matrix consist of an “odds ratio”:

$$\frac{\text{the probability that an alignment is authentic}}{\text{the probability that the alignment was random}}$$

The score S for an alignment of residues a,b is given by:

$S(a,b) = 10 \log_{10} (M_{ab}/p_b)$

As an example, for tryptophan,

$S(a,\text{tryptophan}) = 10 \log_{10} (0.55/0.010) = 17.4$

## Normalized frequencies of amino acids

|  | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 13 | 6 | 9 | 9 | 5 | 8 | 9 | 12 | 6 | 8 | 6 | 7 | 7 | 4 | 11 | 11 | 11 | 2 | 4 | 9 |
| R | 3 | 17 | 4 | 3 | 2 | 5 | 3 | 2 | 6 | 3 | 2 | 9 | 4 | 1 | 4 | 4 | 3 | 7 | 2 | 2 |
| N | 4 | 4 | 6 | 7 | 2 | 5 | 6 | 4 | 6 | 3 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | 2 | 3 | 3 |
| D | 5 | 4 | 8 | 11 | 1 | 7 | 10 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| C | 2 | 1 | 1 | 1 | 52 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 2 |
| Q | 3 | 5 | 5 | 6 | 1 | 10 | 7 | 3 | 7 | 2 | 3 | 5 | 3 | 1 | 4 | 3 | 3 | 1 | 2 | 3 |
| E | 5 | 4 | 7 | 11 | 1 | 9 | 12 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| G | 12 | 5 | 10 | 10 | 4 | 7 | 9 | 27 | 5 | 5 | 4 | 6 | 5 | 3 | 8 | 11 | 9 | 2 | 3 | 7 |
| H | 2 | 5 | 5 | 4 | 2 | 7 | 4 | 2 | 15 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 2 |
| I | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 10 | 6 | 2 | 6 | 5 | 2 | 3 | 4 | 1 | 3 | 9 |
| L | 6 | 4 | 4 | 3 | 2 | 6 | 4 | 3 | 5 | 15 | 34 | 4 | 20 | 13 | 5 | 4 | 6 | 6 | 7 | 13 |
| K | 6 | 18 | 10 | 8 | 2 | 10 | 8 | 5 | 8 | 5 | 4 | 24 | 9 | 2 | 6 | 8 | 8 | 4 | 3 | 5 |
| M | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 6 | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| F | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 5 | 6 | 1 | 4 | 32 | 1 | 2 | 2 | 4 | 20 | 3 |
| P | 7 | 5 | 5 | 4 | 3 | 5 | 4 | 5 | 5 | 3 | 3 | 4 | 3 | 2 | 20 | 6 | 5 | 1 | 2 | 4 |
| S | 9 | 6 | 8 | 7 | 7 | 6 | 7 | 9 | 6 | 5 | 4 | 7 | 5 | 3 | 9 | 10 | 9 | 4 | 4 | 6 |
| T | 8 | 5 | 6 | 6 | 4 | 5 | 5 | 6 | 4 | 6 | 4 | 6 | 5 | 3 | 6 | 8 | 11 | 2 | 3 | 6 |
| W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 55 | 1 | 0 |
| Y | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 15 | 1 | 2 | 2 | 3 | 31 | 2 |
| V | 7 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 15 | 10 | 4 | 10 | 5 | 5 | 5 | 7 | 2 | 4 | 17 |

|  |  |
|---|---|
| Arg | 4.1% |
| Asn | 4.0% |
| Phe | 4.0% |
| Gln | 3.8% |
| Ile | 3.7% |
| His | 3.4% |
| Cys | 3.3% |
| Tyr | 3.0% |
| Met | 1.5% |
| Trp | 1.0% |

## What do the numbers mean in a log odds matrix?

$S(a,\text{tryptophan}) = 10 \log_{10} (0.55/0.010) = 17.4$

A score of +17 for tryptophan means that this alignment is 50 times more likely than a chance alignment of two Trp residues.

$S(a,b) = 10 \log_{10} (M_{ab}/p_b)$<br>
$S(a,b) = 17$<br>
Probability of replacement $(M_{ab}/p_b) = x$<br>
Then<br>
$17 = 10 \log_{10} x$<br>
$1.7 = \log_{10} x$<br>
$10^{1.7} = x$<br>
$50 = x$

## What do the numbers mean in a log odds matrix?

A score of –10 indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one tenth as frequent as the chance alignment of these amino acids.

A score of 0 is neutral.

A score of +2 indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance.

## PAM10 log odds scoring matrix

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 7 | | | | | | | | | | | | | | | | | | | |
| **R** | -10 | 9 | | | | | | | | | | | | | | | | | | |
| **N** | -7 | -9 | 9 | | | | | | | | | | | | | | | | | |
| **D** | -6 | -17 | -1 | 8 | | | | | | | | | | | | | | | | |
| **C** | -10 | -11 | -17 | -21 | 10 | | | | | | | | | | | | | | | |
| **Q** | -7 | -4 | -7 | -6 | -20 | 9 | | | | | | | | | | | | | | |
| **E** | -5 | -15 | -5 | 0 | -20 | -1 | 8 | | | | | | | | | | | | | |
| **G** | -4 | -13 | -6 | -6 | -13 | -10 | -7 | 7 | | | | | | | | | | | | |
| **H** | -11 | -4 | -2 | -7 | -10 | -2 | -9 | -13 | 10 | | | | | | | | | | | |
| **I** | -8 | -8 | -8 | -11 | -9 | -11 | -8 | -17 | -13 | 9 | | | | | | | | | | |
| **L** | -9 | -12 | -10 | -19 | -21 | -8 | -13 | -14 | -9 | -4 | 7 | | | | | | | | | |
| **K** | -10 | -2 | -4 | -8 | -20 | -6 | -7 | -10 | -10 | -9 | -11 | 7 | | | | | | | | |
| **M** | -8 | -7 | -15 | -17 | -20 | -7 | -10 | -12 | -17 | -3 | -2 | -4 | 12 | | | | | | | |
| **F** | -12 | -12 | -12 | -21 | -19 | -19 | -20 | -12 | -9 | -5 | -5 | -20 | -7 | 9 | | | | | | |
| **P** | -4 | -7 | -9 | -12 | -11 | -6 | -9 | -10 | -7 | -12 | -10 | -10 | -11 | -13 | 8 | | | | | |
| **S** | -3 | -6 | -2 | -7 | -6 | -8 | -7 | -4 | -9 | -10 | -12 | -7 | -8 | -9 | -4 | 7 | | | | |
| **T** | -3 | -10 | -5 | -8 | -11 | -9 | -9 | -10 | -11 | -5 | -10 | -6 | -7 | -12 | -7 | -2 | 8 | | | |
| **W** | -20 | -5 | -11 | -21 | -22 | -19 | -23 | -21 | -10 | -20 | -9 | -18 | -19 | -7 | -20 | -8 | -19 | 13 | | |
| **Y** | -11 | -14 | -7 | -17 | -7 | -18 | -11 | -20 | -6 | -9 | -10 | -12 | -17 | -1 | -20 | -10 | -9 | -8 | 10 | |
| **V** | -5 | -11 | -12 | -11 | -9 | -10 | -10 | -9 | -9 | -1 | -5 | -13 | -4 | -12 | -9 | -10 | -6 | -22 | -10 | 8 |
| | **A** | **R** | **N** | **D** | **C** | **Q** | **E** | **G** | **H** | **I** | **L** | **K** | **M** | **F** | **P** | **S** | **T** | **W** | **Y** | **V** |

BLOSUM 80 BLOSUM 62 BLOSUM 45

PAM 1 PAM 120 PAM 250

*Less divergent* ← → *More divergent*

More conserved

Less conserved

Rat versus mouse protein

Rat versus bacterial protein

## Comparing two proteins with a PAM1 matrix gives completely different results than PAM250!

Consider two distantly related proteins. A PAM40 matrix is not forgiving of mismatches, and penalizes them severely. Using this matrix you can find almost no match.

```
hsrbp,  136 CRLLNLDGTC
btlact,   3 CLLLALALTC
            * ** *  **
```

A PAM250 matrix is very tolerant of mismatches.

```
24.7% identity in 81 residues overlap; Score: 77.0; Gap frequency: 3.7%
  rbp4 26 RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDV
btlact 21 QTMKGLDIQKVAGTWYSLAMAASD-ISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWEN
                *     ****  *        * *       *               **  *

  rbp4 86 --CADMVGTFTDTEDPAKFKM
btlact 80 GECAQKKIIAEKTKIPAVFKI
            **        *  ** **
```

## PAM: “Accepted point mutation”

• Two proteins with 50% identity may have 80 changes per 100 residues. Why? Because any residue can be subject to back mutations.

• Proteins with 20% to 25% identity are in the “twilight zone” and may be statistically significantly related.

• PAM or “accepted point mutation” refers to the “hits” or matches between two sequences (Dayhoff & Eck, 1968)

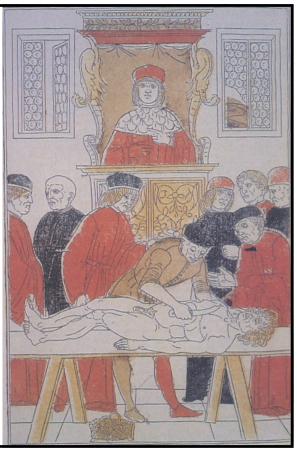
**Outline**

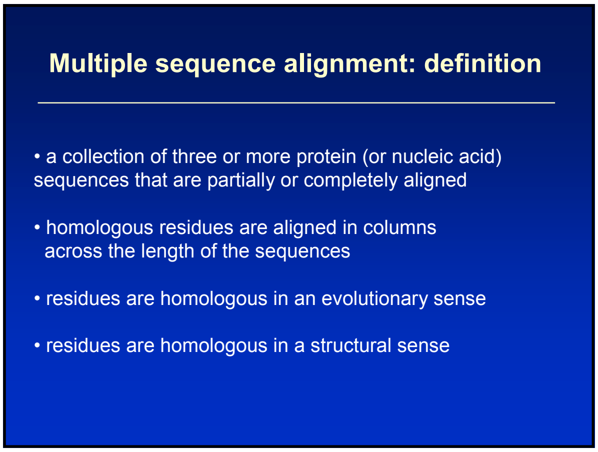
1. Pairwise alignment of proteins

2. Scoring matrices: how related are amino acids?

3. Multiple sequence alignment of proteins

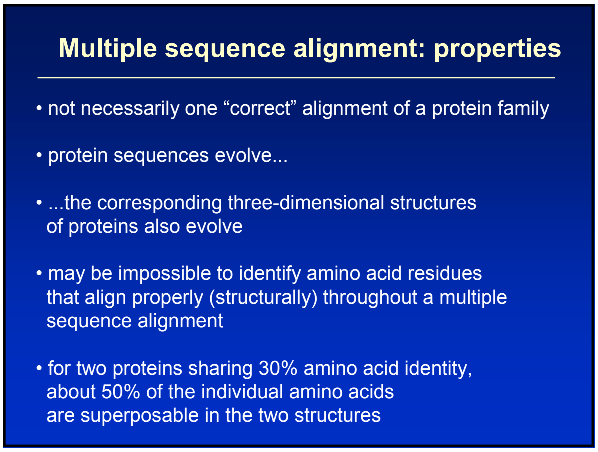
4. From multiple sequence alignment to phylogenetic tree





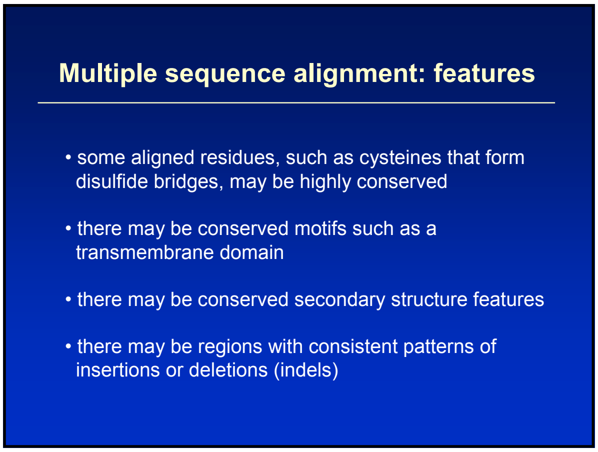
## Multiple sequence alignment: definition

- a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned
- homologous residues are aligned in columns across the length of the sequences
- residues are homologous in an evolutionary sense
- residues are homologous in a structural sense



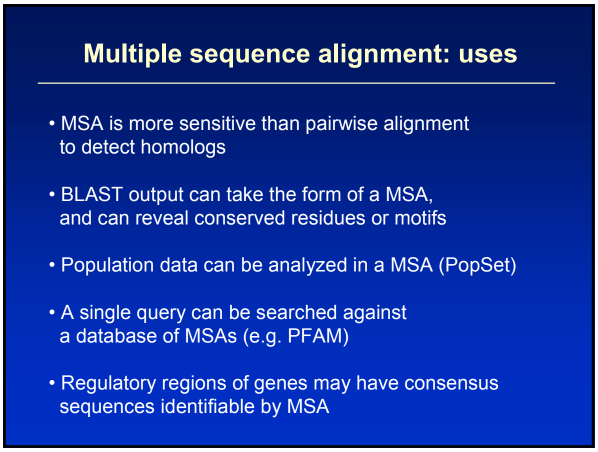
## Multiple sequence alignment: properties

- not necessarily one “correct” alignment of a protein family
- protein sequences evolve...
- ...the corresponding three-dimensional structures of proteins also evolve
- may be impossible to identify amino acid residues that align properly (structurally) throughout a multiple sequence alignment
- for two proteins sharing 30% amino acid identity, about 50% of the individual amino acids are superposable in the two structures



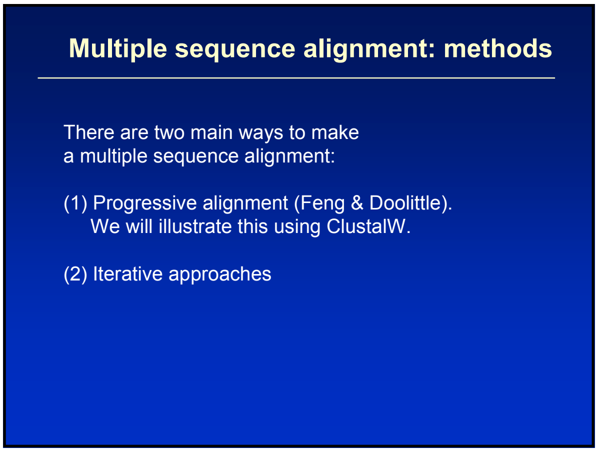
## Multiple sequence alignment: features

- some aligned residues, such as cysteines that form disulfide bridges, may be highly conserved
- there may be conserved motifs such as a transmembrane domain
- there may be conserved secondary structure features
- there may be regions with consistent patterns of insertions or deletions (indels)



## Multiple sequence alignment: uses

- MSA is more sensitive than pairwise alignment to detect homologs
- BLAST output can take the form of a MSA, and can reveal conserved residues or motifs
- Population data can be analyzed in a MSA (PopSet)
- A single query can be searched against a database of MSAs (e.g. PFAM)
- Regulatory regions of genes may have consensus sequences identifiable by MSA

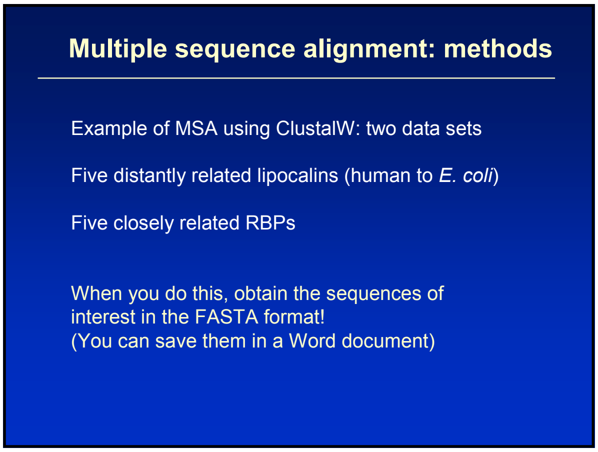


## Multiple sequence alignment: methods

There are two main ways to make a multiple sequence alignment:

(1) Progressive alignment (Feng & Doolittle).
We will illustrate this using ClustalW.

(2) Iterative approaches



## Multiple sequence alignment: methods

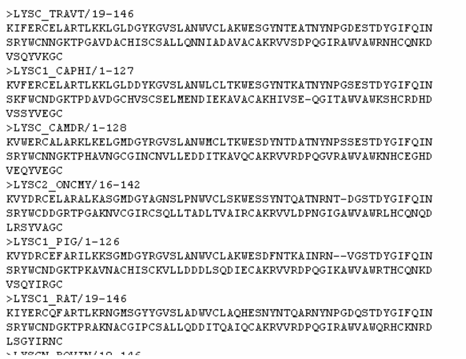
Example of MSA using ClustalW: two data sets

Five distantly related lipocalins (human to *E. coli*)

Five closely related RBPs

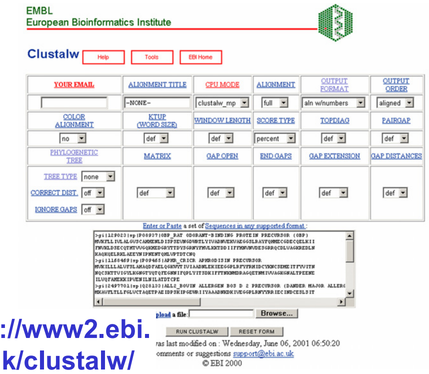
When you do this, obtain the sequences of interest in the FASTA format!
(You can save them in a Word document)

## The input for ClustalW: a group of sequences (DNA or protein) in the FASTA format



```
>LYSC_TRAVT/19-146
KIFERCELARTLKKLGLDGYKGVSLANWVCLAKWESGYNTEATNYNPGDESTDYGIFQIN
SRYWCNNGKTPGAVDACHISCSALLQNNIADAVACAKRVVSDPQGIRAWVAWRNHCQNKD
VSQYVKGC
>LYSC1_CAPHI/1-127
KVFERCELARTLKKLGLDDYKGVSLANWLCLTKWESGYNTKATNYNPGSESTDYGIFQIN
SKFWCNDGKTPDAVDGCHVSCSELMENDIEKAVACAKHIVSE-QGITAWVAWKSHCRDHD
VSSYVEGC
>LYSC_CAMDR/1-128
KVWERCALARKLKELGMDGYRGVSLANWMCLTKWESDYNTDATNYNPSSESTDYGIFQIN
SRYWCNNGKTPHAVNGCGINCNVLLEDDITKAVQCAKRVVRDPQGVRAWVAWKNHCEGHD
VEQYVEGC
>LYSC2_ONCMY/16-142
KVYDRCELARALKASGMDGYAGNSLPNWVCLSKWESSYNTQATNRNT-DGSTDYGIFQIN
SRYWCDDGRTPGAKNVCGIRCSQLLTADLTVAIRCAKRVVLDPNGIGAWVAWRLHCQNQD
LRSYVAGC
>LYSC1_PIG/1-126
KVYDRCEFARILKKSGMDGYRGVSLANWVCLAKWESDFNTKAINRN--VGSTDYGIFQIN
SRYWCNDGKTPKAVNACHISCKVLLDDDLSQDIECAKRVVRDPQGIKAWVAWRTHCQNKD
VSQYIRGC
>LYSC1_RAT/19-146
KIYERCQFARTLKRNGMSGYYGVSLADWVCLAQHESNYNTQARNYNPGDQSTDYGIFQIN
SRYWCNDGKTPRAKNACGIPCSALLQDDITQAIQCAKRVVRDPQGIRAWVAWQRHCKNRD
LSGYIRNC
>LYSCN_BOVIN/19-146
```

## Use ClustalW to do a progressive MSA



EMBL
European Bioinformatics Institute

Clustalw

| | | | | | |
|---|---|---|---|---|---|
| YOUR EMAIL | ALIGNMENT TITLE | CPU MODE | ALIGNMENT | OUTPUT FORMAT | OUTPUT ORDER |
| | -NONE- | clustalw_mp | full | aln w/numbers | aligned |
| COLOR ALIGNMENT | KTUP (WORD SIZE) | WINDOW LENGTH | SCORE TYPE | TOPDIAG | PAIRGAP |
| no | def | def | percent | def | def |
| PHYLOGENETIC TREE | MATRIX | GAP OPEN | END GAPS | GAP EXTENSION | GAP DISTANCES |
| TREE TYPE none<br>CORRECT DIST. off<br>IGNORE GAPS off | def | def | def | def | def |

Enter or Paste a set of Sequences in any supported format :

```
>gi|129023|sp|P08937|OBP_RAT ODORANT-BINDING PROTEIN PRECURSOR (OBP)
MVKFLLIVLALGVSCAHHENLDISPSEVNGDWRTLYIVADNVEKVAEGGSLRAYFQHMECGDECQELKII
FNVKLDSECQTHTVVGQKHEDGRYTTDYSGRNYFHVLKKTDDIIFFHNVNVDESGRRQCDLVAGKREDLN
KAQKQELRKLAEEYNIPNENTQHLVPTDTCNQ
>gi|1168469|sp|P09465|APHR_CRICR APHRODISIN PRECURSOR
MVKILLLALVFSLAHAQDFAELQGKWYTIVIAADNLEKIEEGGPLRFYFRHIDCYKNCSEMEITFYVITN
NQCSKTTVIGYLKGNGTYQTQFEGNNIFQPLYITSDKIFFTNKNMDRAGQETNMIVVAGKGNALTPEENE
ILVQFAHEKKIPVENILNILATDTCPE
>gi|2497701|sp|Q28133|ALL2_BOVIN ALLERGEN BOS D 2 PRECURSOR (DANDER MAJOR ALLERG
MKAVFLTLLFGLVCTAQETPAEIDPSKIPGEWRIIYAAADNKDKIVEGGPLRNYYRRIECINDCESLSIT
```

Upload a file: Browse...

was last modified on : Wednesday, June 06, 2001 06:50:20

comments or suggestions support@ebi.ac.uk

© EBI 2000

http://www2.ebi.ac.uk/clustalw/

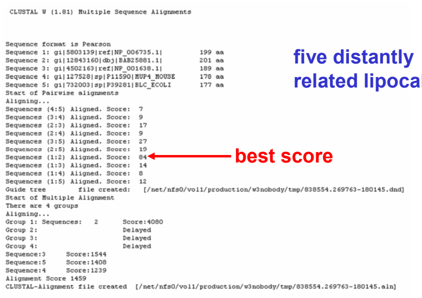
## Feng-Doolittle MSA occurs in 3 stages

[1] Do a set of global pairwise alignments (Needleman and Wunsch’s dynamic programming algorithm)

[2] Create a guide tree

[3] Progressively align the sequences

## Progressive MSA stage 1 of 3: generate global pairwise alignments

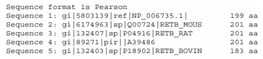


```
 CLUSTAL W (1.81) Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: gi|5803139|ref|NP_006735.1|        199 aa
Sequence 2: gi|12843160|dbj|BAB25881.1|        201 aa
Sequence 3: gi|4502163|ref|NP_001638.1|        189 aa
Sequence 4: gi|127528|sp|P11590|MUP4_MOUSE      178 aa
Sequence 5: gi|732003|sp|P39281|BLC_ECOLI       177 aa
Start of Pairwise alignments
Aligning...
Sequences (4:5) Aligned. Score:  7
Sequences (3:4) Aligned. Score:  9
Sequences (2:3) Aligned. Score:  17
Sequences (2:4) Aligned. Score:  9
Sequences (3:5) Aligned. Score:  27
Sequences (2:5) Aligned. Score:  19
Sequences (1:2) Aligned. Score:  84
Sequences (1:3) Aligned. Score:  14
Sequences (1:4) Aligned. Score:  8
Sequences (1:5) Aligned. Score:  12
Guide tree        file created:   [/net/nfs0/vol1/production/w3nobody/tmp/838554.269763-180145.dnd]
Start of Multiple Alignment
There are 4 groups
Aligning...
Group 1: Sequences:   2      Score:4080
Group 2:                     Delayed
Group 3:                     Delayed
Group 4:                     Delayed
Sequence:3     Score:1544
Sequence:5     Score:1408
Sequence:4     Score:1239
Alignment Score 1459
CLUSTAL-Alignment file created  [/net/nfs0/vol1/production/w3nobody/tmp/838554.269763-180145.aln]
```

five distantly related lipocalins

best score

## Progressive MSA stage 1 of 3: generate global pairwise alignments



```
Sequence format is Pearson
Sequence 1: gi|5803139|ref|NP_006735.1|         199 aa
Sequence 2: gi|6174963|sp|Q00724|RETB_MOUS      201 aa
Sequence 3: gi|132407|sp|P04916|RETB_RAT        201 aa
Sequence 4: gi|89271|pir||A39486                201 aa
Sequence 5: gi|132403|sp|P18902|RETB_BOVIN      183 aa
```

Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score: 84
Sequences (1:3) Aligned. Score: 84
Sequences (1:4) Aligned. Score: 91
Sequences (1:5) Aligned. Score: 92
Sequences (2:3) Aligned. Score: 99
Sequences (2:4) Aligned. Score: 86
Sequences (2:5) Aligned. Score: 85
Sequences (3:4) Aligned. Score: 85
Sequences (3:5) Aligned. Score: 84
Sequences (4:5) Aligned. Score: 96

five closely related lipocalins

best score

## Number of pairwise alignments needed

For *n* sequences, (*n*-1)(*n*) / 2

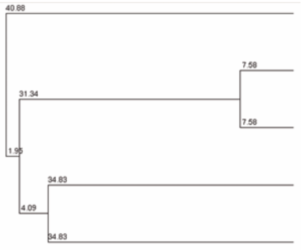
For 5 sequences, (4)(5) / 2 = 10

## Feng-Doolittle stage 2: guide tree

- Convert similarity scores to distance scores
- A tree shows the distance between objects
- Use UPGMA (defined below)
- ClustalW provides a syntax to describe the tree

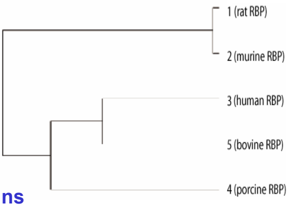
## Progressive MSA stage 2 of 3: generate a guide tree calculated from the distance matrix

```
(
(
(
gi|5803139|ref|NP_006735.1|:0.09385,
gi|12843160|dbj|BAB25881.1|:0.05691)
:0.34199,
gi|127528|sp|P11590|MUP4_MOUSE:0.48994)
:0.06118,
gi|4502163|ref|NP_001638.1|:0.35805,
gi|732003|sp|P39281|BLC_ECOLI:0.36511);
```



## Progressive MSA stage 2 of 3: generate guide tree

```
(
(
gi|5803139|ref|NP_006735.1|:0.04284,
(
gi|6174963|sp|Q00724|RETB_MOUS:0.00075,
gi|132407|sp|P04916|RETB_RAT:0.00423)
:0.10542)
:0.01900,
gi|89271|pir||A39486:0.01924,
gi|132403|sp|P18902|RETB_BOVIN:0.01902);
```

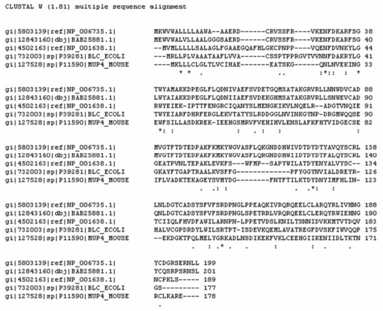


five closely related lipocalins

## Feng-Doolittle stage 3: progressive alignment

- Make a MSA based on the order in the guide tree
- Start with the two most closely related sequences
- Then add the next closest sequence
- Continue until all sequences are added to the MSA
- Rule: “once a gap, always a gap.”

## Progressive MSA stage 3 of 3: progressively align the sequences following the branch order of the tree

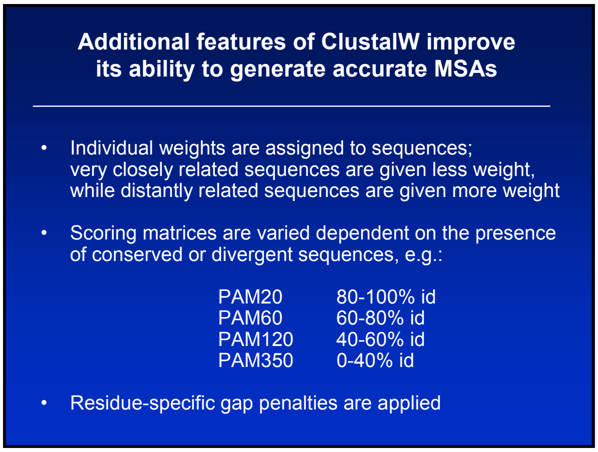


```
CLUSTAL W (1.81) multiple sequence alignment


gi|5803139|ref|NP_006735.1|         MKWVWALLLLAAWA--AAERD------CRVSSFR----VKENFDKARFSG 38
gi|12843160|dbj|BAB25881.1|         MEWVWALVLLAALGGGSAERD------CRVSSFR----VKENFDKARFSG 40
gi|4502163|ref|NP_001638.1|         --MVMLLLLSALAGLFGAAEGQAFHLGKCPNPP----VQENFDVNKYLG 44
gi|732003|sp|P39281|BLC_ECOLI       ---MRLLPLVAAATAAFLVVA------CSSPTPPRGVTVVNNFDAKRYLG 41
gi|127528|sp|P11590|MUP4_MOUSE      ----MKLLLCLGLTLVCIHAE-------EATSKG------QNLNVEKING 33
                                          * *  .                  ..        :*::  :  *

gi|5803139|ref|NP_006735.1|         TWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCAD 88
gi|12843160|dbj|BAB25881.1|         LWYAIAKKDPEGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCAD 90
gi|4502163|ref|NP_001638.1|         RWYEIEK-IPTTFENGRCIQANYSLMENGKIKVLNQELR--ADGTVNQIE 91
gi|732003|sp|P39281|BLC_ECOLI       TWYEIARFDHRFERGLEKVTATYSLRDDGGLNVINKGYNP-DRGMWQQSE 90
gi|127528|sp|P11590|MUP4_MOUSE      EWFSILLASDKREK-IEEHGSMRVFVEHIHVLENSLAFKFHTVIDGECSE 82
                                     *: :               :   . :   :                  :

gi|5803139|ref|NP_006735.1|         MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL 138
gi|12843160|dbj|BAB25881.1|         MVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRL 140
gi|4502163|ref|NP_001638.1|         GEATPVNLTEPAKLEVKFS---WFMP--SAPYWILATDYENYALVYSC-- 134
gi|732003|sp|P39281|BLC_ECOLI       GKAYFTGAPTRAALKVSFFG-------------PFYGGYNVIALDREYR- 126
gi|127528|sp|P11590|MUP4_MOUSE      IFLVADKTEKAGEYSVMYDG--------FNTFTILKTDYDNYIMFHLIN- 123
                                               .  .: :                .  .*:   :

gi|5803139|ref|NP_006735.1|         LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNG 188
gi|12843160|dbj|BAB25881.1|         QNLDGTCADSYSFVFSRDPNGLSPETRRLVRQRQEELCLERQYRWIEHNG 190
gi|4502163|ref|NP_001638.1|         TCIIQLFHVDFAWILARNPN-LPPETVDSLKNILTSNNIDVKKMTVTDQV 183
gi|732003|sp|P39281|BLC_ECOLI       HALVCGPDRDYLWILSRTPT-ISDEVKQEMLAVATREGFDVSKFIWVQQP 175
gi|127528|sp|P11590|MUP4_MOUSE      --EKDGKTFQLMELYGRKADLNSDIKEKFVKLCEEHGIIKENIIDLTKTN 171
                                             .   : .* .   .      :        :  .     .

gi|5803139|ref|NP_006735.1|         YCDGRSERNLL 199
gi|12843160|dbj|BAB25881.1|         YCQSRPSRNSL 201
gi|4502163|ref|NP_001638.1|         NCPKLS----- 189
gi|732003|sp|P39281|BLC_ECOLI       GS--------- 177
gi|127528|sp|P11590|MUP4_MOUSE      RCLKARE---- 178
                                     .
```

## Clustal W alignment of 5 closely related lipocalins

```
CLUSTAL W (1.82) multiple sequence alignment


gi|89271|pir||A39486                MEWVWALVLLAALGSAQAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 50
gi|132403|sp|P18902|RETB_BOVIN      ------------------ERDCRVSSFRVKENFDKARFAGTWYAMAKKDP 32
gi|5803139|ref|NP_006735.1|         MKWVWALLLLAAW--AAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 48
gi|6174963|sp|Q00724|RETB_MOUS      MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDP 50
gi|132407|sp|P04916|RETB_RAT        MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDP 50
                                                      ********************:* ***:*****

gi|89271|pir||A39486                EGLFLQDNIVAEFSVDENGHMSATAKGRVRLLNNWDVCADMVGTFTDTED 100
gi|132403|sp|P18902|RETB_BOVIN      EGLFLQDNIVAEFSVDENGHMSATAKGRVRLLNNWDVCADMVGTFTDTED 82
gi|5803139|ref|NP_006735.1|         EGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTED 98
gi|6174963|sp|Q00724|RETB_MOUS      EGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTED 100
gi|132407|sp|P04916|RETB_RAT        EGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTED 100
                                    *********:*******.*:************.**:**************

gi|89271|pir||A39486                PAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAAQYSCRLQNLDGTCADS 150
gi|132403|sp|P18902|RETB_BOVIN      PAKFKMKYWGVASFLQKGNDDHWIIDTDYETFAVQYSCRLLNLDGTCADS 132
gi|5803139|ref|NP_006735.1|         PAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADS 148
gi|6174963|sp|Q00724|RETB_MOUS      PAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRLQNLDGTCADS 150
gi|132407|sp|P04916|RETB_RAT        PAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRLQNLDGTCADS 150
                                    ****************:*******:****:*:* ****** *********
```

* asterisks indicate identity in a column

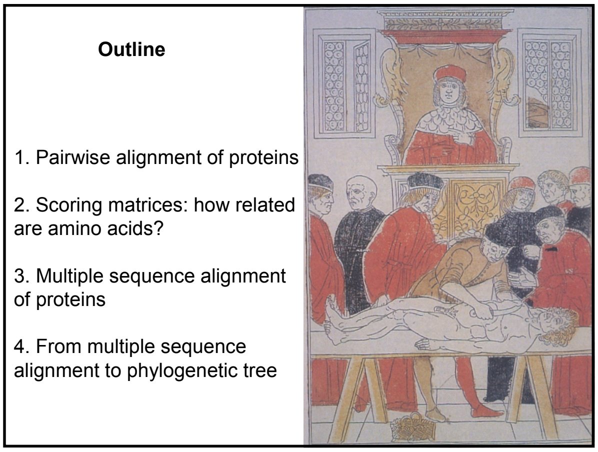


## Additional features of ClustalW improve its ability to generate accurate MSAs

- Individual weights are assigned to sequences; very closely related sequences are given less weight, while distantly related sequences are given more weight
- Scoring matrices are varied dependent on the presence of conserved or divergent sequences, e.g.:

| | |
|---|---|
| PAM20 | 80-100% id |
| PAM60 | 60-80% id |
| PAM120 | 40-60% id |
| PAM350 | 0-40% id |

- Residue-specific gap penalties are applied



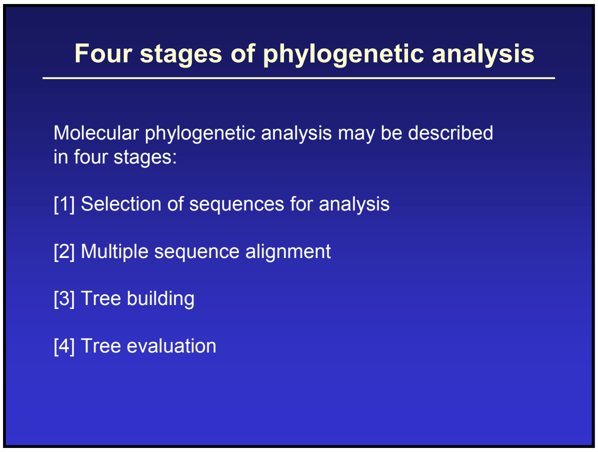
**Outline**

1. Pairwise alignment of proteins

2. Scoring matrices: how related are amino acids?

3. Multiple sequence alignment of proteins

4. From multiple sequence alignment to phylogenetic tree



## Four stages of phylogenetic analysis

Molecular phylogenetic analysis may be described in four stages:

[1] Selection of sequences for analysis

[2] Multiple sequence alignment

[3] Tree building

[4] Tree evaluation

## Stage 1: Use of DNA, RNA, or protein

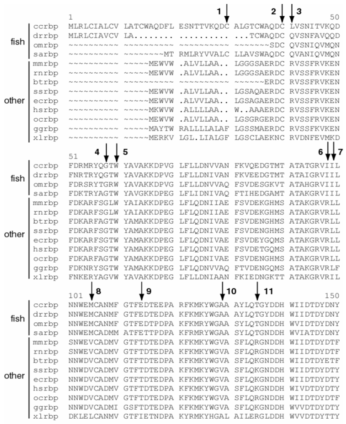
For some phylogenetic studies, it may be preferable to use protein instead of DNA sequences. With DNA, one can also study synonymous versus nonsynonymous mutations, noncoding DNA, pseudogenes, etc.

## Stage 2: Multiple sequence alignment

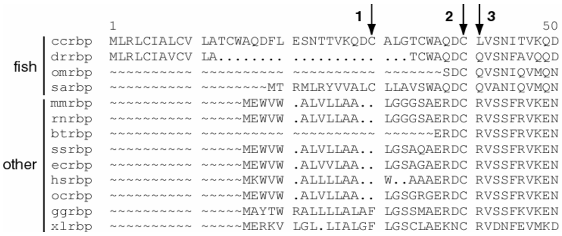
The fundamental basis of a phylogenetic tree is a multiple sequence alignment.

(If there is a misalignment, or if a nonhomologous sequence is included in the alignment, it will still be possible to generate a tree.)

Consider the following alignment of 13 orthologous retinol-binding proteins.



```
                                                1          2  3
             1                                                      50
      ccrbp  MLRLCIALCV LATCWAQDFL ESNTTVKQDC ALGTCWAQDC LVSNITVKQD
fish  drrbp  MLRLCIAVCV LA........ .......... ...TCWAQDC QVSNFAVQQD
      omrbp  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~SDC QVSNIQVMQN
      sarbp  ~~~~~~~~~~ ~~~~~~~~MT RMLRYVVALC LLAVSWAQDC QVANIQVMQN
      mmrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGGGSAERDC RVSSFRVKEN
      rnrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGGGSAERDC RVSSFRVKEN
      btrbp  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~ERDC RVSSFRVKEN
      ssrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGSAQAERDC RVSSFRVKEN
other ecrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVVLAA.. LGSAGAERDC RVSSFRVKEN
      hsrbp  ~~~~~~~~~~ ~~~~~MKWVW .ALLLLAA.. W..AAAERDC RVSSFRVKEN
      ocrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGSGRGERDC RVSSFRVKEN
      ggrbp  ~~~~~~~~~~ ~~~~~MAYTW RALLLLALAF LGSSMAERDC RVSSFKVKEN
      xlrbp  ~~~~~~~~~~ ~~~~~MERKV LGL.LIALGF LGSCLAEKNC RVDNFEVMKD

                  4  5                                      6  7
             51
      ccrbp  FDRMRYQGTW YAVAKKDPVG LFLLDNVVAN FKVQEDGTMT ATATGRVIIL
fish  drrbp  FNRTRYQGTW YAVAKKDPVG LFLLDNIVAN FKVEEDGTMT ATAIGRVIIL
      omrbp  FDRSRYTGRW YAVAKKDPVG LFLLDNVVAQ FSVDESGKVT ATAHGRVIIL
      sarbp  FDKTRYAGTW YAVGKKDPEG LFLIDNIVAQ FTIHEDGAMT ATAKGRVIIL
      mmrbp  FDKARFSGLW YAIAKKDPEG LFLQDNIIAE FSVDEKGHMS ATAKGRVRLL
      rnrbp  FDKARFSGLW YAIAKKDPEG LFLQDNIIAE FSVDEKGHMS ATAKGRVRLL
      btrbp  FDKARFAGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS ATAKGRVRLL
      ssrbp  FDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS ATAKGRVRLL
other ecrbp  FDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDEYGQMS ATAKGRVRLL
      hsrbp  FDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDETGQMS ATAKGRVRLL
      ocrbp  FDKARFAGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS ATAKGRVRLL
      ggrbp  FDKNRYSGTW YAMAKKDPEG LFLQDNVVAQ FTVDENGQMS ATAKGRVRLF
      xlrbp  FNKERYAGVW YAVAKKDPEG LFLLDNIAAN FKIEDNGKTT ATAKGRVRIL

                8          9                 10       11
             101                                                    150
      ccrbp  NNWEMCANMF GTFEDTEEPA RFKMKYWGAA AYLQTGYDDH WIIDTDYDNY
fish  drrbp  NNWEMCANMF GTFEDTEDPA KFKMKYWGAA AYLQTGYDDH WIIDTDYDNY
      omrbp  NNWEMCANMF GTFEDTPDPA KFKMRYWGAA SYLQTGNDDH WVIDTDYDNY
      sarbp  NNWEMCADMM ATFETTPDPA KFRMRYWGAA SYLQTGNDDH WVIATDYDNY
      mmrbp  SNWEVCADMV GTFTDTEDPA KFKMKYWGVA SFLQRGNDDH WIIDTDYDTF
      rnrbp  SNWEVCADMV GTFTDTEDPA KFKMKYWGVA SFLQRGNDDH WIIDTDYDTF
      btrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQKGNDDH WIIDTDYETF
      ssrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQKGNDDH WIIDTDYDTY
other ecrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQKGNDDH WIIDTDYDTY
      hsrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQKGNDDH WIVDTDYDTY
      ocrbp  NNWDVCADMV GTFTDTEDPA KFKMKYWGVA SFLQRGNDDH WIIDTDYDTF
      ggrbp  NNWDVCADMI GSFTDTEDPA KFKMKYWGVA SFLQKGNDDH WVVDTDYDTY
      xlrbp  DKLELCANMV GTFIETNDPA KYRMKYHGAL AILERGLDDH WVVDTDYTTY
```



```
             1                                                50
      ccrbp  MLRLCIALCV LATCWAQDFL ESNTTVKQDC ALGTCWAQDC LVSNITVKQD
      drrbp  MLRLCIAVCV LA........ .......... ...TCWAQDC QVSNFAVQQD
fish  omrbp  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~SDC QVSNIQVMQN
      sarbp  ~~~~~~~~~~ ~~~~~~~~MT RMLRYVVALC LLAVSWAQDC QVANIQVMQN
      mmrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGGGSAERDC RVSSFRVKEN
      rnrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGGGSAERDC RVSSFRVKEN
      btrbp  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~ERDC RVSSFRVKEN
      ssrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGSAQAERDC RVSSFRVKEN
other ecrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVVLAA.. LGSAGAERDC RVSSFRVKEN
      hsrbp  ~~~~~~~~~~ ~~~~~MKWVW .ALLLLAA.. W..AAAERDC RVSSFRVKEN
      ocrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGSGRGERDC RVSSFRVKEN
      ggrbp  ~~~~~~~~~~ ~~~~~MAYTW RALLLLALAF LGSSMAERDC RVSSFKVKEN
      xlrbp  ~~~~~~~~~~ ~~~~~MERKV LGL.LIALGF LGSCLAEKNC RVDNFEVMKD
```

Some positions of the multiple sequence alignment are invariant (arrow 2). Some positions distinguish fish RBP from all other RBPs (arrow 3).

## Stage 2: Multiple sequence alignment

[1] Confirm that all sequences are homologous

[2] Adjust gap creation and extension penalties as needed to optimize the alignment

[3] Restrict phylogenetic analysis to regions of the multiple sequence alignment for which data are available for all taxa (delete columns having incomplete data—even if a gap occurs in only one taxon).

[4] In this example, note that four RBPs are from fish, while the others are vertebrates that evolved more recently.

## Stage 3: Tree-building methods

We will discuss two tree-building methods: distance-based and character-based.

Distance-based methods involve a distance metric, such as the number of amino acid changes between the sequences, or a distance score. Examples of distance-based algorithms are UPGMA and neighbor-joining.

## Stage 3: Tree-building methods

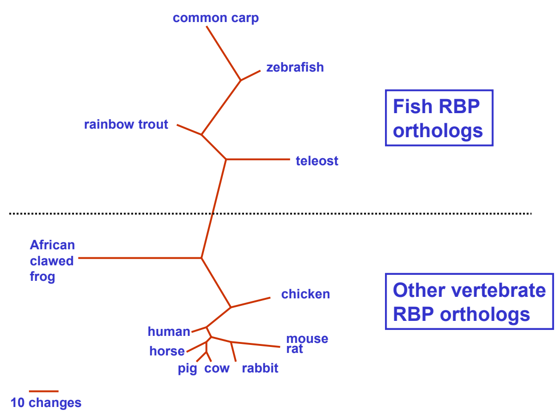
We will discuss two tree-building methods: distance-based and character-based.

Distance-based methods involve a distance metric, such as the number of amino acid changes between the sequences, or a distance score. Examples of distance-based algorithms are UPGMA and neighbor-joining.

Character-based methods include maximum parsimony and maximum likelihood. Parsimony analysis involves the search for the tree with the fewest amino acid (or nucleotide) changes that account for the observed differences between taxa.

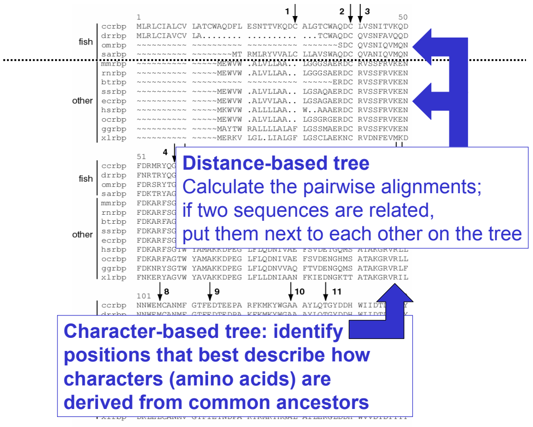
## Stage 3: Tree-building methods

We can introduce distance-based and character-based tree-building methods by referring to a tree of 13 orthologous retinol-binding proteins, and the multiple sequence alignment from which the tree was generated.



Fish RBP orthologs

Other vertebrate RBP orthologs



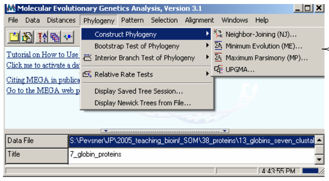
```
       1                                                 50
ccrbp  MLRLCIALCV LATCWAQDFL ESNTTVKQDC ALGTCWAQDC LVSNITVKQD
drrbp  MLRLCIAVCV LA........ .......... ...TCWAQDC QVSNFAVQQD
omrbp  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~SDC QVSNIQVMQN
sarbp  ~~~~~~~~~~ ~~~~~~~~MT RMLRYVVALC LLAVSWAQDC QVANIQVMQN
mmrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGGGSAERDC RVSSFRVKEN
rnrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGGGSAERDC RVSSFRVKEN
btrbp  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~ERDC RVSSFRVKEN
ssrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGSAQAERDC RVSSFRVKEN
ecrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVVLAA.. LGSAGAERDC RVSSFRVKEN
hsrbp  ~~~~~~~~~~ ~~~~~MKWVW .ALLLLAA.. W..AAAERDC RVSSFRVKEN
ocrbp  ~~~~~~~~~~ ~~~~~MEWVW .ALVLLAA.. LGSGRGERDC RVSSFRVKEN
ggrbp  ~~~~~~~~~~ ~~~~~MAYTW RALLLLALAF LGSSMAERDC RVSSFKVKEN
xlrbp  ~~~~~~~~~~ ~~~~~MERKV LGL.LIALGF LGSCLAEKNC RVDNFEVMKD

       51
ccrbp  FDRMRYQG
drrbp  FNRTRYQG
omrbp  FDRSRYTG
sarbp  FDKTRYAG
mmrbp  FDKARFSG
rnrbp  FDKARFSG
btrbp  FDKARFAG
ssrbp  FDKARFSG
ecrbp  FDKARFSG
hsrbp  FDKARFSGTW YAMAKKDPEG LFLQDNIVAE FSVDETGQMS ATAKGRVRLL
ocrbp  FDKARFAGTW YAMAKKDPEG LFLQDNIVAE FSVDENGHMS ATAKGRVRLL
ggrbp  FDKNRYSGTW YAMAKKDPEG LFLQDNVVAQ FTVDENGQMS ATAKGRVRLF
xlrbp  FNKERYAGVW YAVAKKDPEG LFLLDNIAAN FKIEDNGKTT ATAKGRVRIL

       101
ccrbp  NNWEMCANMF GTFEDTEEPA RFKMKYWGAA AYLQTGYDDH WIIDT
```

**Distance-based tree**
Calculate the pairwise alignments;
if two sequences are related,
put them next to each other on the tree

**Character-based tree: identify positions that best describe how characters (amino acids) are derived from common ancestors**

## How to use MEGA to make a tree

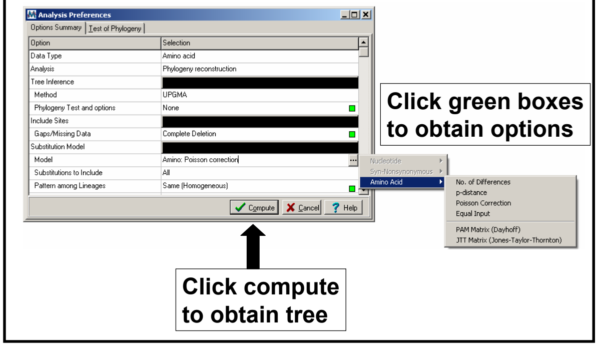
[1] Download MEGA for free (www.megasoftware.net)
[2] Enter a multiple sequence alignment (.meg) file
[3] Under the phylogeny menu, select one of these four methods…



Data File: S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_seven_clustal
Title: 7_globin_proteins

Neighbor-Joining (NJ)
Minimum Evolution (ME)
Maximum Parsimony (MP)
UPGMA

## Use of MEGA for a distance-based tree: UPGMA

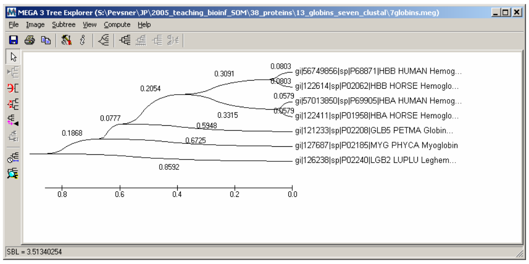


| Option | Selection |
|---|---|
| Data Type | Amino acid |
| Analysis | Phylogeny reconstruction |
| Tree Inference | |
| Method | UPGMA |
| Phylogeny Test and options | None 🟩 |
| Include Sites | |
| Gaps/Missing Data | Complete Deletion 🟩 |
| Substitution Model | |
| Model | Amino: Poisson correction |
| Substitutions to Include | All |
| Pattern among Lineages | Same (Homogeneous) 🟩 |

**Click green boxes to obtain options**

**Click compute to obtain tree**

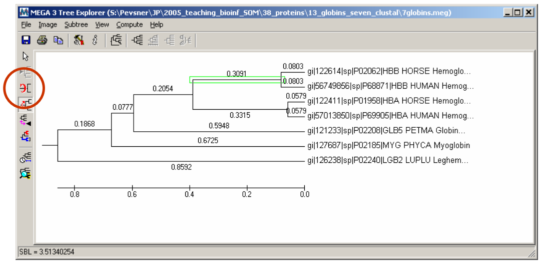
## Use of MEGA for a distance-based tree: UPGMA


MEGA 3 Tree Explorer (S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_seven_clustal\7globins.meg)
File Image Subtree View Compute Help
0.0803
0.3091
0.0803
0.2054
0.0579
0.0579
0.3315
0.0777
0.5948
0.1868
0.6725
0.8592
gi|56749856|sp|P68871|HBB HUMAN Hemog...
gi|122614|sp|P02062|HBB HORSE Hemoglo...
gi|57013850|sp|P69905|HBA HUMAN Hemog...
gi|122411|sp|P01958|HBA HORSE Hemoglo...
gi|121233|sp|P02208|GLB5 PETMA Globin...
gi|127687|sp|P02185|MYG PHYCA Myoglobin
gi|126238|sp|P02240|LGB2 LUPLU Leghem...
0.8
0.6
0.4
0.2
0.0

SBL = 3.51340254

**A variety of styles are available for tree display**

## Use of MEGA for a distance-based tree: UPGMA

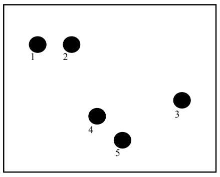

MEGA 3 Tree Explorer (S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_seven_clustal\7globins.meg)
File Image Subtree View Compute Help
0.0803
0.3091
0.0803
0.2054
0.0579
0.0777
0.0579
0.3315
0.1868
0.5948
0.6725
0.8592
gi|122614|sp|P02062|HBB HORSE Hemoglo...
gi|56749856|sp|P68871|HBB HUMAN Hemog...
gi|122411|sp|P01958|HBA HORSE Hemoglo...
gi|57013850|sp|P69905|HBA HUMAN Hemog...
gi|121233|sp|P02208|GLB5 PETMA Globin...
gi|127687|sp|P02185|MYG PHYCA Myoglobin
gi|126238|sp|P02240|LGB2 LUPLU Leghem...
0.8
0.6
0.4
0.2
0.0

SBL = 3.51340254

**Flipping branches around a node creates an equivalent topology**

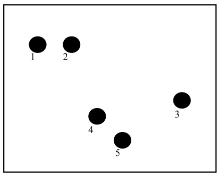
## Tree-building methods: UPGMA

UPGMA is
**u**nweighted **p**air **g**roup **m**ethod
using **a**rithmetic mean



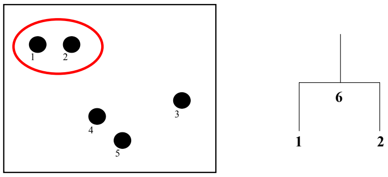
## Tree-building methods: UPGMA

Step 1: compute the pairwise distances of all the proteins. Get ready to put the numbers 1-5 at the bottom of your new tree.



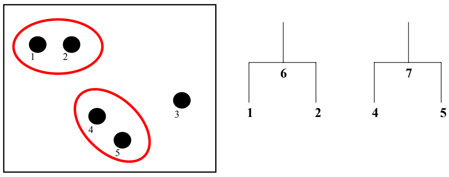
## Tree-building methods: UPGMA

Step 2: Find the two proteins with the smallest pairwise distance. Cluster them.



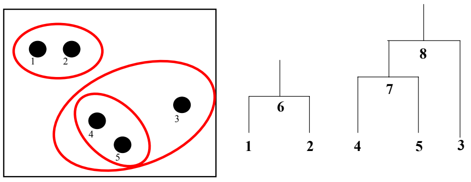
## Tree-building methods: UPGMA

Step 3: Do it again. Find the next two proteins with the smallest pairwise distance. Cluster them.



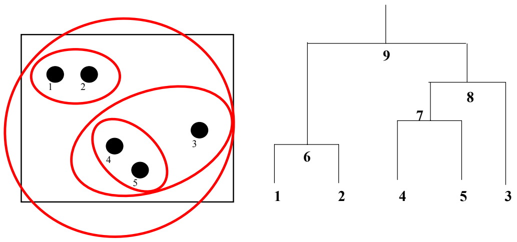
## Tree-building methods: UPGMA

Step 4: Keep going. Cluster.

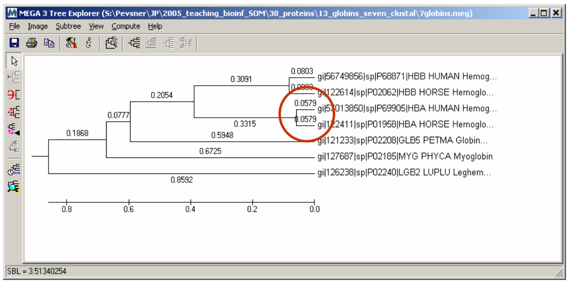


## Tree-building methods: UPGMA

Step 4: Last cluster! This is your tree.



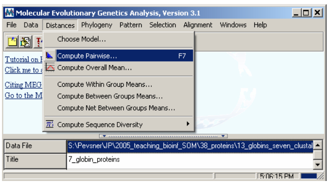
## MEGA for UPGMA: branch lengths reflect differences


MEGA 3 Tree Explorer (S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_seven_clustal\7globins.meg)

File
Image
Subtree
View
Compute
Help
0.3091
0.0803
0.2054
0.0777
0.1868
0.3315
0.0579
0.0579
0.5948
0.6725
0.8592
gi|56749856|sp|P68871|HBB HUMAN Hemog...
gi|122614|sp|P02062|HBB HORSE Hemoglo...
gi|57013850|sp|P69905|HBA HUMAN Hemog...
gi|122411|sp|P01958|HBA HORSE Hemoglo...
gi|121233|sp|P02208|GLB5 PETMA Globin...
gi|127687|sp|P02185|MYG PHYCA Myoglobin
gi|126238|sp|P02240|LGB2 LUPLU Leghem...
0.8
0.6
0.4
0.2
0.0

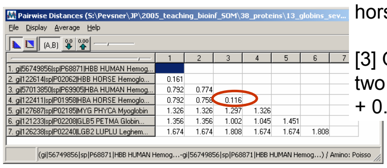
SBL = 3.51340254

## MEGA for UPGMA: branch lengths reflect differences


Molecular Evolutionary Genetics Analysis, Version 3.1
File Data Distances Phylogeny Pattern Selection Alignment Windows Help
Choose Model…
Compute Pairwise… F7
Compute Overall Mean…
Compute Within Group Means…
Compute Between Groups Means…
Compute Net Between Groups Means…
Compute Sequence Diversity
Tutorial on I
Click me to
Citing MEG
Go to the M
5:06:15 PM

Data File: S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_seven_clustal

Title: 7_globin_proteins


Pairwise Distances (S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_sev...
File Display Average Help
(A,B)
0.0
0.00

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. gi\|56749856\|sp\|P68871\|HBB HUMAN Hemog... | | | | | | | |
| 2. gi\|122614\|sp\|P02062\|HBB HORSE Hemoglo... | 0.161 | | | | | | |
| 3. gi\|57013850\|sp\|P69905\|HBA HUMAN Hemog... | 0.792 | 0.774 | | | | | |
| 4. gi\|122411\|sp\|P01958\|HBA HORSE Hemoglo... | 0.792 | 0.758 | 0.116 | | | | |
| 5. gi\|127687\|sp\|P02185\|MYG PHYCA Myoglobin | 1.326 | 1.326 | 1.297 | 1.326 | | | |
| 6. gi\|121233\|sp\|P02208\|GLB5 PETMA Globin... | 1.356 | 1.356 | 1.002 | 1.045 | 1.451 | | |
| 7. gi\|126238\|sp\|P02240\|LGB2 LUPLU Leghem... | 1.674 | 1.674 | 1.808 | 1.674 | 1.674 | 1.808 | |

(gi|56749856|sp|P68871|HBB HUMAN Hemog...-gi|56749856|sp|P68871|HBB HUMAN Hemog...) / Amino: Poisso

[1] From main MEGA menu, compute pairwise distances

[2] Note that the smallest distance is 0.116 (from human to horse hemoglobin).

[3] On the tree, these two taxa are 0.0579 + 0.0579 = 0.116 apart!

## Stage 4: Evaluating trees

The main criteria by which the accuracy of a phylogentic tree is assessed are consistency, efficiency, and robustness. Evaluation of accuracy can refer to an approach (e.g. UPGMA) or to a particular tree.

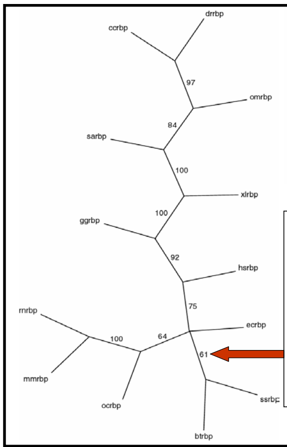
## Stage 4: Evaluating trees: bootstrapping

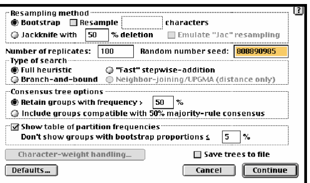
Bootstrapping is a commonly used approach to measuring the robustness of a tree topology. Given a branching order, how consistently does an algorithm find that branching order in a randomly permuted version of the original data set?

## Stage 4: Evaluating trees: bootstrapping

Bootstrapping is a commonly used approach to measuring the robustness of a tree topology. Given a branching order, how consistently does an algorithm find that branching order in a randomly permuted version of the original data set?

To bootstrap, make an artificial dataset obtained by randomly sampling columns from your multiple sequence alignment. Make the dataset the same size as the original. Do 100 (to 1,000) bootstrap replicates. Observe the percent of cases in which the assignment of clades in the original tree is supported by the bootstrap replicates. >70% is considered significant.




Resampling method
● Bootstrap □ Resample ___ characters
○ Jackknife with 50 % deletion □ Emulate "Jac" resampling
Number of replicates: 100 Random number seed: 808890985
Type of search
● Full heuristic ○ "Fast" stepwise-addition
○ Branch-and-bound ○ Neighbor-joining/UPGMA (distance only)
Consensus tree options
● Retain groups with frequency > 50 %
○ Include groups compatible with 50% majority-rule consensus
☑ Show table of partition frequencies
Don't show groups with bootstrap proportions ≤ 5 %

**In 61% of the bootstrap resamplings, ssrbp and btrbp (pig and cow RBP) formed a distinct clade. In 39% of the cases, another protein joined the clade (e.g. ecrbp), or one of these two sequences joined another clade.**