Chapter III: Pairwise Sequence Alignment

Presentations use info from:

Jonathan Pevsner, Ph.D.
http://bioinfbook.org
pevsner@kennedykrieger.org
Bioinformatics and Functional Genomics
(3rd edition, ©2015 John Wiley & Sons, Ltd.)
You may use this PowerPoint for teaching purposes

dr. Stanislav Kolenčík
stanislav.kolencik@famnit.upr.si

What will you learn?

To define homology as well as orthologs and paralogs

To explain how PAM (accepted point mutation) matrices are derived

To contrast the utility of PAM and BLOSUM scoring matrices.

To define dynamic programming and explain how global and local pairwise alignments are performed

To perform pairwise alignment of protein or DNA sequences at the NCBI website

Pairwise sequence alignment is the most fundamental operation of bioinformatics

- It is used to decide if two proteins (or genes) 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.
- It is used in the analysis of genomes.

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

Protein alignment: often more informative than DNA alignment

Sequence alignment: protein sequences can be more informative than DNA

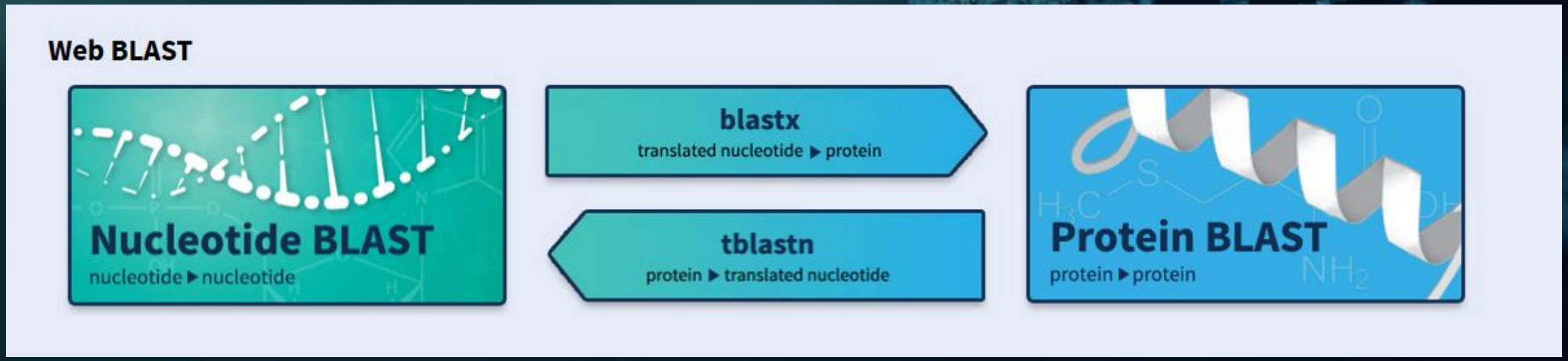
- Protein is more informative (20 vs 4 characters);
- Many amino acids share related biophysical properties
- Codons are degenerate: changes in the third position
- Often do not alter the amino acid that is specified
- Protein sequences offer a longer “look-back” time

Sequence alignment: protein sequences can be more informative than DNA

Example:

->searching for plant globins using human beta globin **DNA** yields no matches;

->searching for plant globins using human beta globin **protein** yields many matches.



Pairwise sequence alignment is the most fundamental operation of bioinformatics

Many times, DNA alignments are appropriate:

- to study non-coding regions of DNA (e.g., introns or intergenic regions)
- database searching
- to study DNA polymorphisms
- genome sequencing relies on DNA analysis

Definitions

Homology

- Similarity attributed to descent from a common ancestor.

Identity

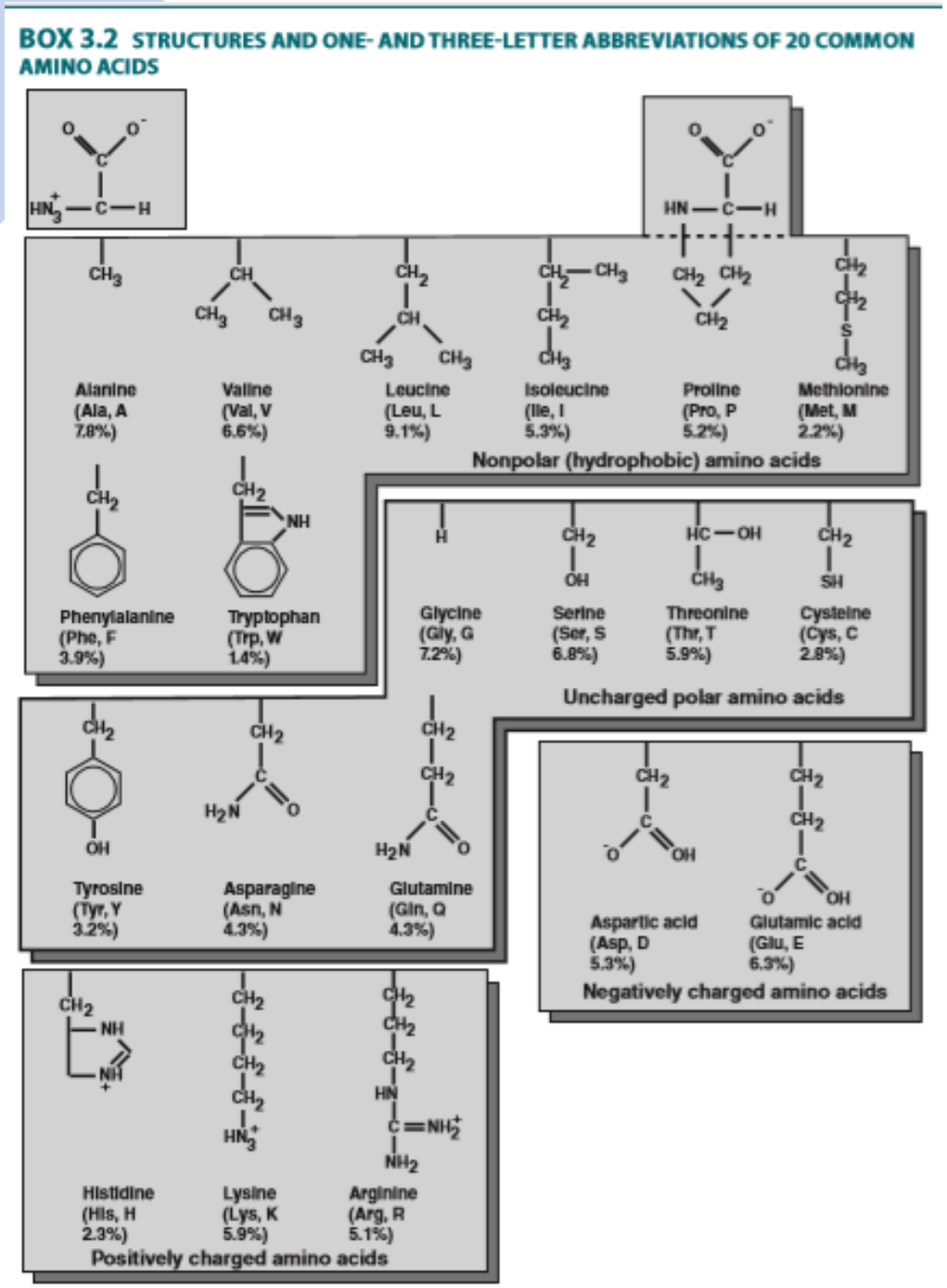
- The extent to which two (nucleotide or amino acid) sequences are invariant.

Similarity

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

Conservation

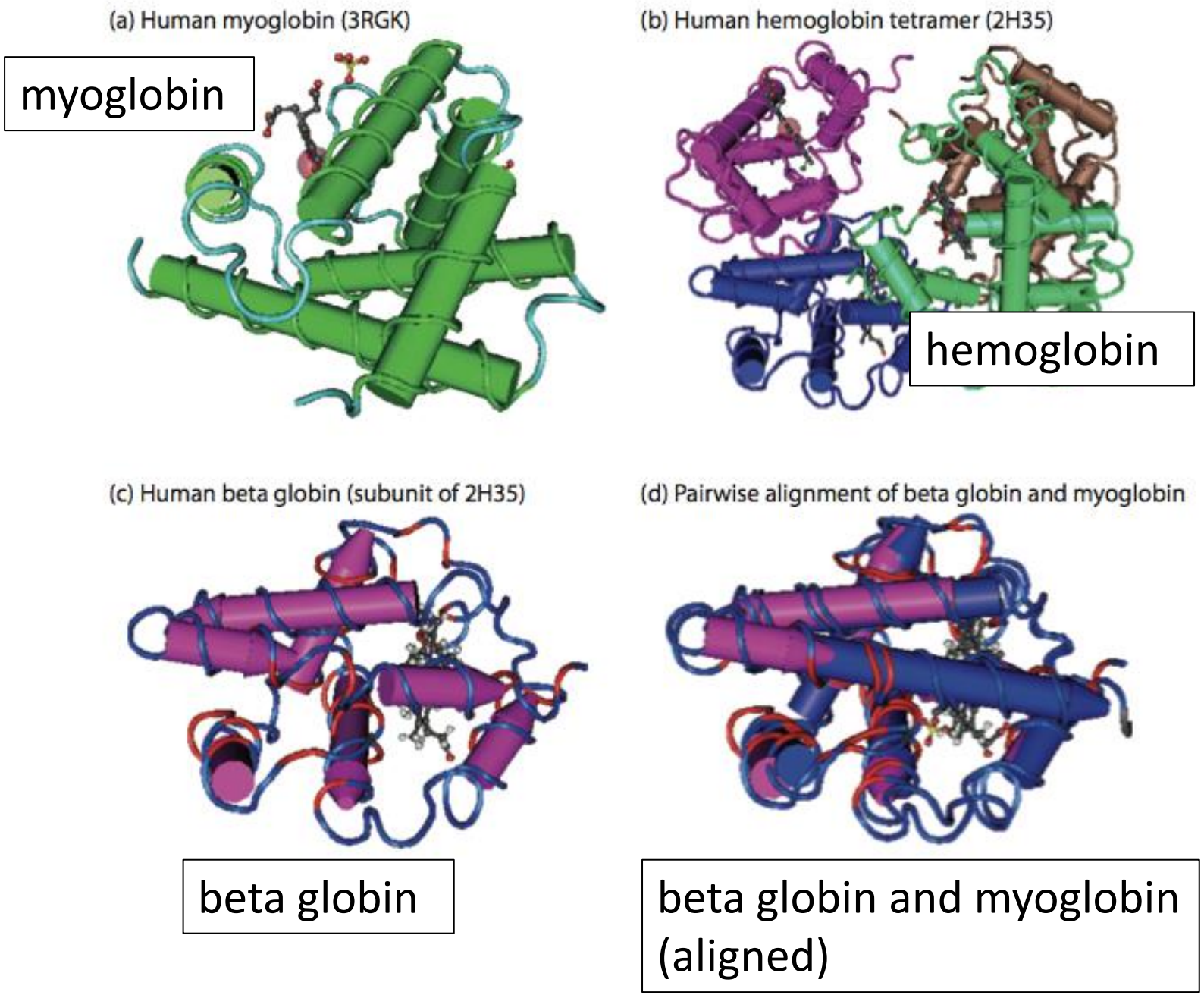
- Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physio-chemical properties of the original residue.

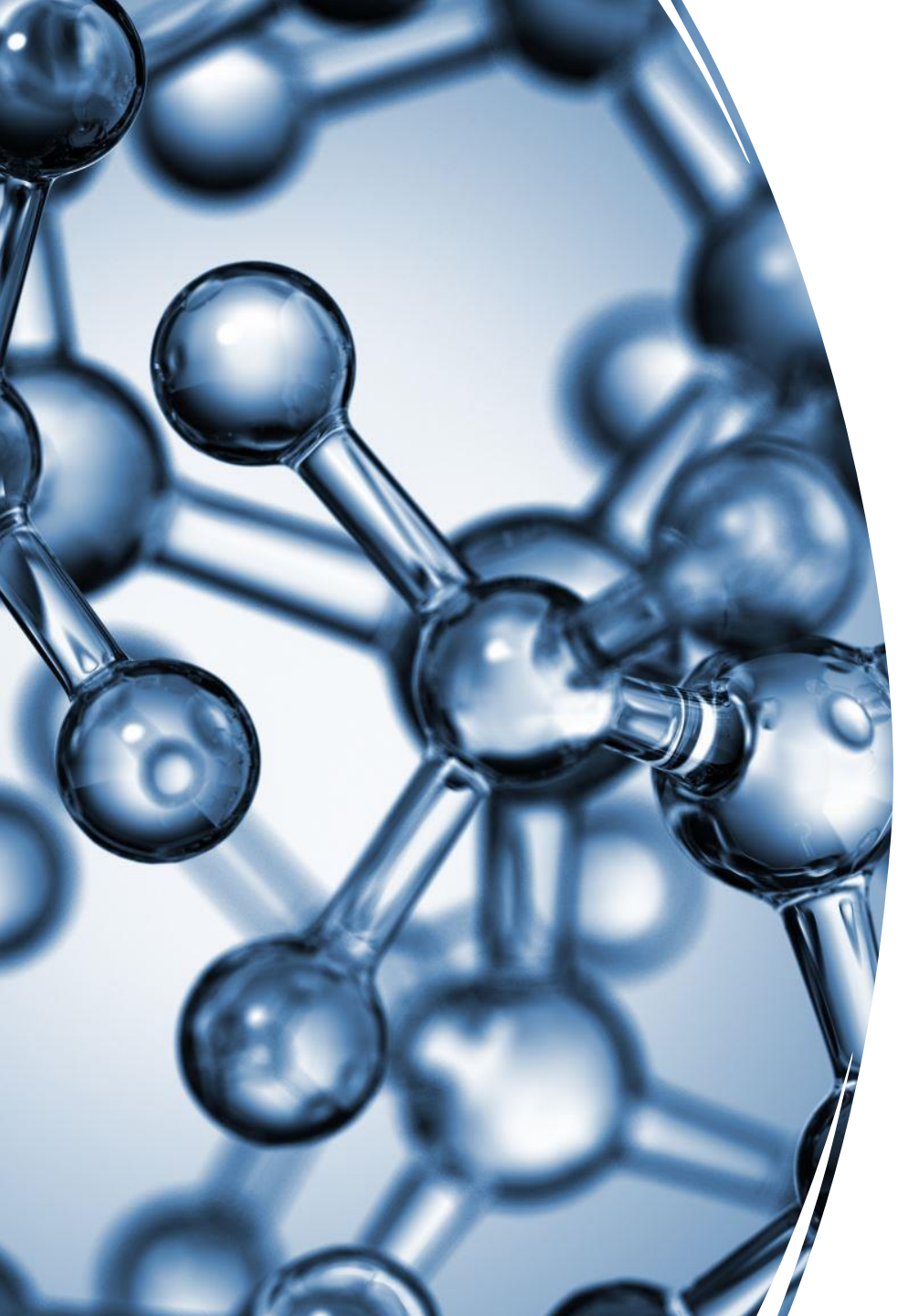


- basic amino acids (K, R, H)
- acidic amino acids (D, E)
- hydroxylated amino acids (S, T)
- hydrophobic amino acids (W, F, Y, L, I, V, M, A)

Globin homologs

In general, three-dimensional structures diverge much more slowly than amino acid sequence identity between two proteins.





Definitions: two types of homology

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.

Example of paralogs: HBA2 and HBA1 genes

HBA2 hemoglobin subunit alpha 2 [*Homo sapiens* (human)]

Gene ID: 3040, updated on 10-Oct-2020

Summary

Official Symbol	HBA2 provided by HGNC
Official Full Name	hemoglobin subunit alpha 2 provided by HGNC
Primary source	HGNC:HGNC:4824
See related	Ensembl:ENSG00000188536 MIM:141850
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	HBH; ECYT7; HBA-T2
Summary	The human alpha globin gene cluster located on chromosome 16 spans about 30 kb and includes seven loci: 5'- zeta - pseudozeta - mu - pseudoalpha-1 - alpha-2 - alpha-1 - theta - 3'. The alpha-2 (HBA2) and alpha-1 (HBA1) coding sequences are identical. These genes differ slightly over the 5' untranslated regions and the introns, but they differ significantly over the 3' untranslated regions. Two alpha chains plus two beta chains constitute HbA, which in normal adult life comprises about 97% of the total hemoglobin; alpha chains combine with delta chains to constitute HbA-2, which with HbF (fetal hemoglobin) makes up the remaining 3% of adult hemoglobin. Alpha thalassemias result from deletions of each of the alpha genes as well as deletions of both HBA2 and HBA1; some nondeletion alpha thalassemias have also been reported. [provided by RefSeq, Jul 2008]
Annotation information	Note: We are aware that there is inconsistency among GenBank, RefSeq, and other browsers about the gene order in the alpha globin cluster. According to review by staff at HbVar (http://globin.bx.psu.edu/hbvar/menu.html), the gene order, telomere to centromere, for these loci is ...HBZ(HBZ2)..HBZP(HBZ1)..HBA2..HBA1... [13 Feb 2013]
Orthologs	all

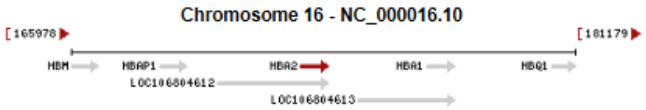
Genomic context

Location: 16p13.3

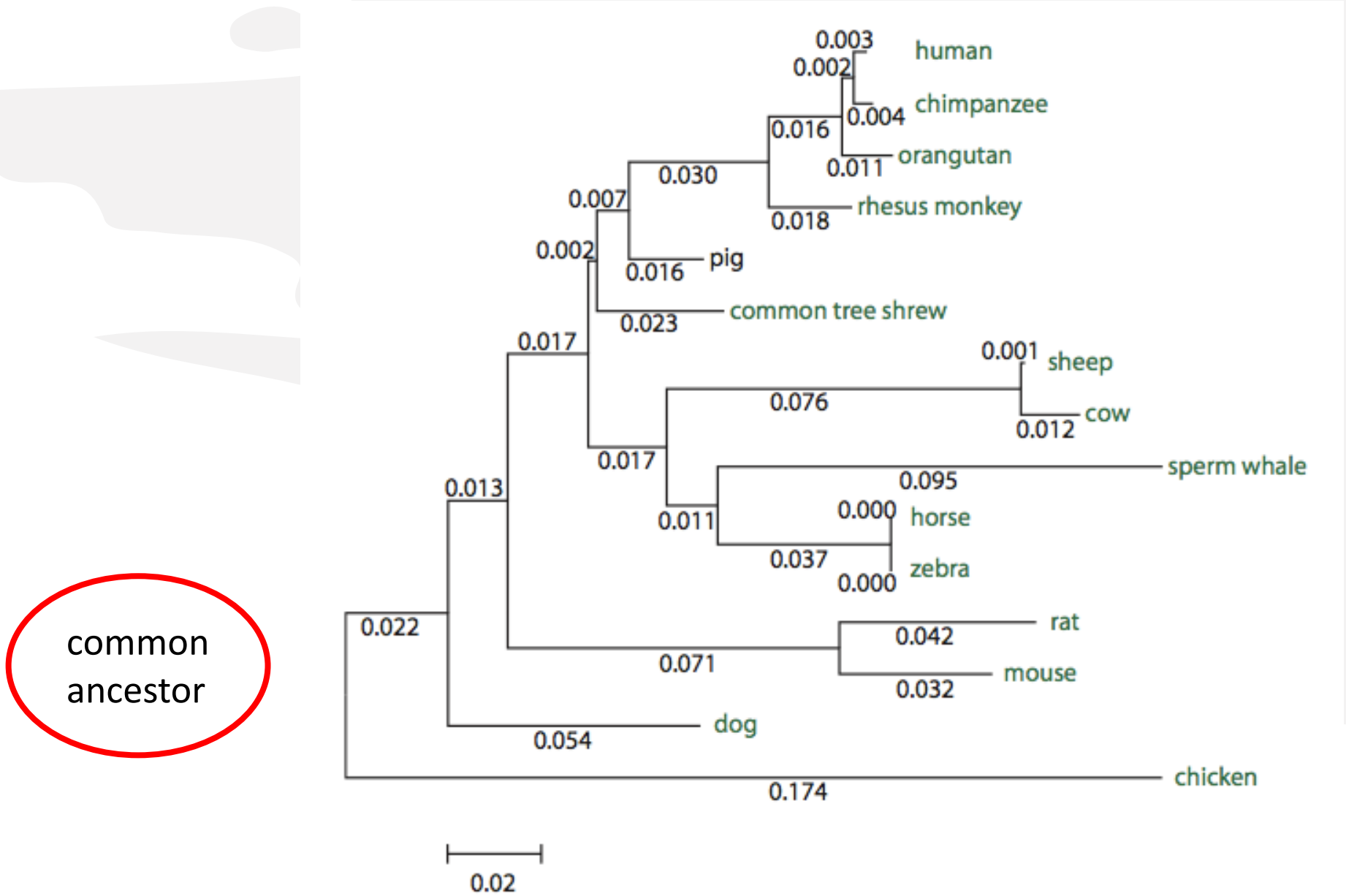
See HBA2 in Genome Data Viewer

Exon count: 3

Annotation release	Status	Assembly	Chr	Location
109.20200815	current	GRCh38.p13 (GCF_000001405.39)	16	NC_000016.10 (172876..173710)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	16	NC_000016.9 (222846..223709)

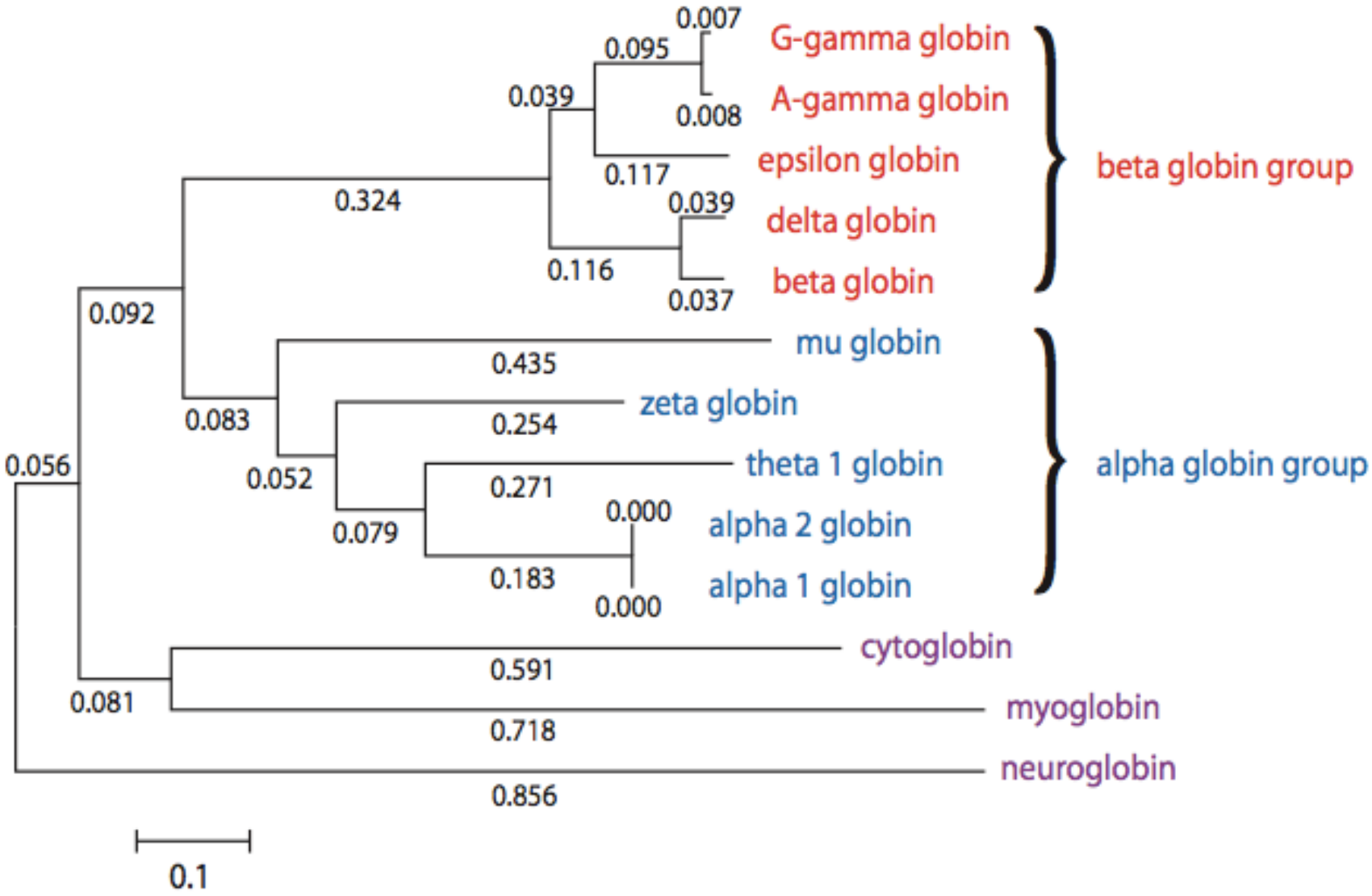


Myoglobin proteins: examples of orthologs



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

This tree shows human globin paralogs.

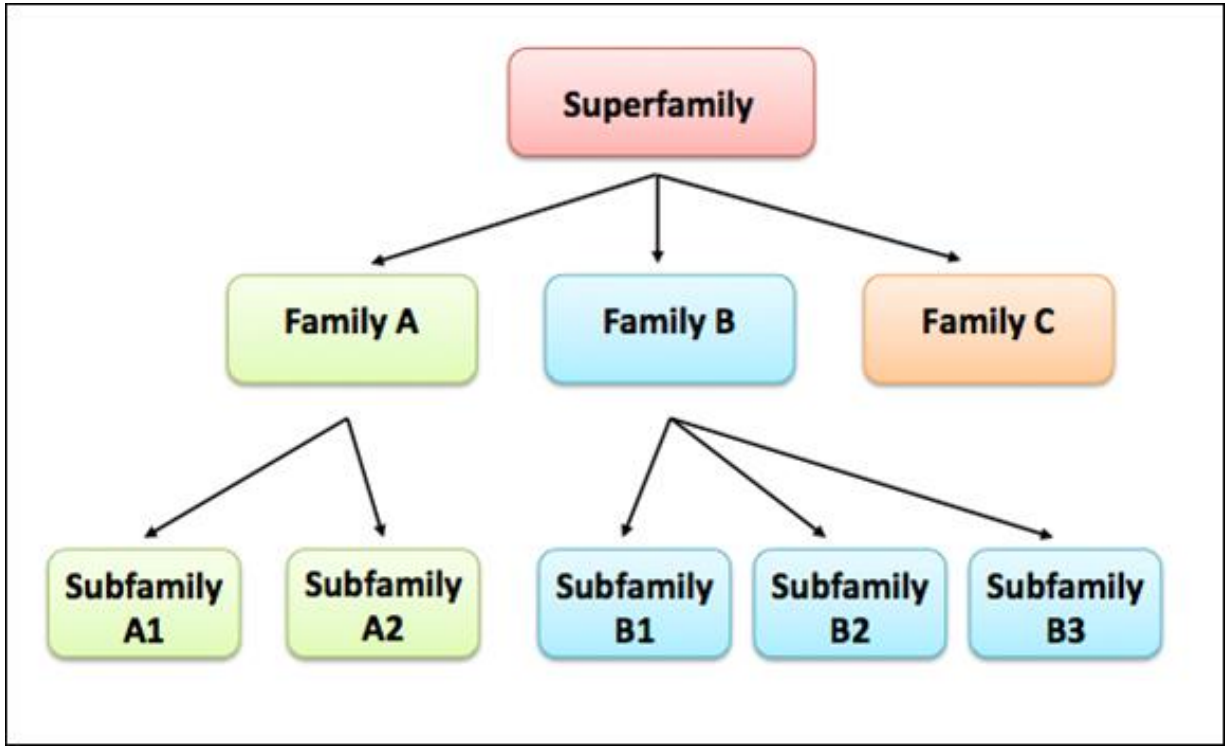


Why classify proteins?

- Proteins can be classified into groups according to sequence or structural similarity.
- These groups often contain well characterized proteins whose function is known.
- When a novel protein is identified, its functional properties can be proposed based on the group to which it is predicted to belong.

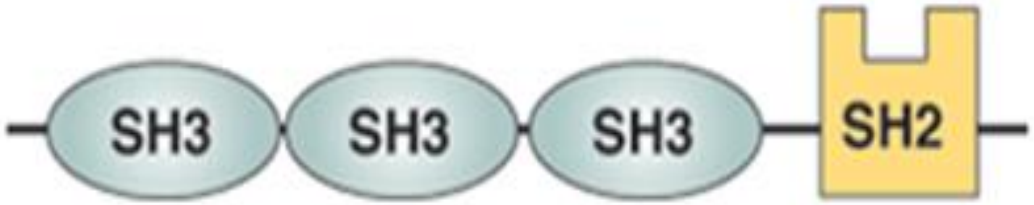
What are protein families?

A group of proteins that share a common evolutionary origin, reflected by their related functions and similarities in sequence or structure.



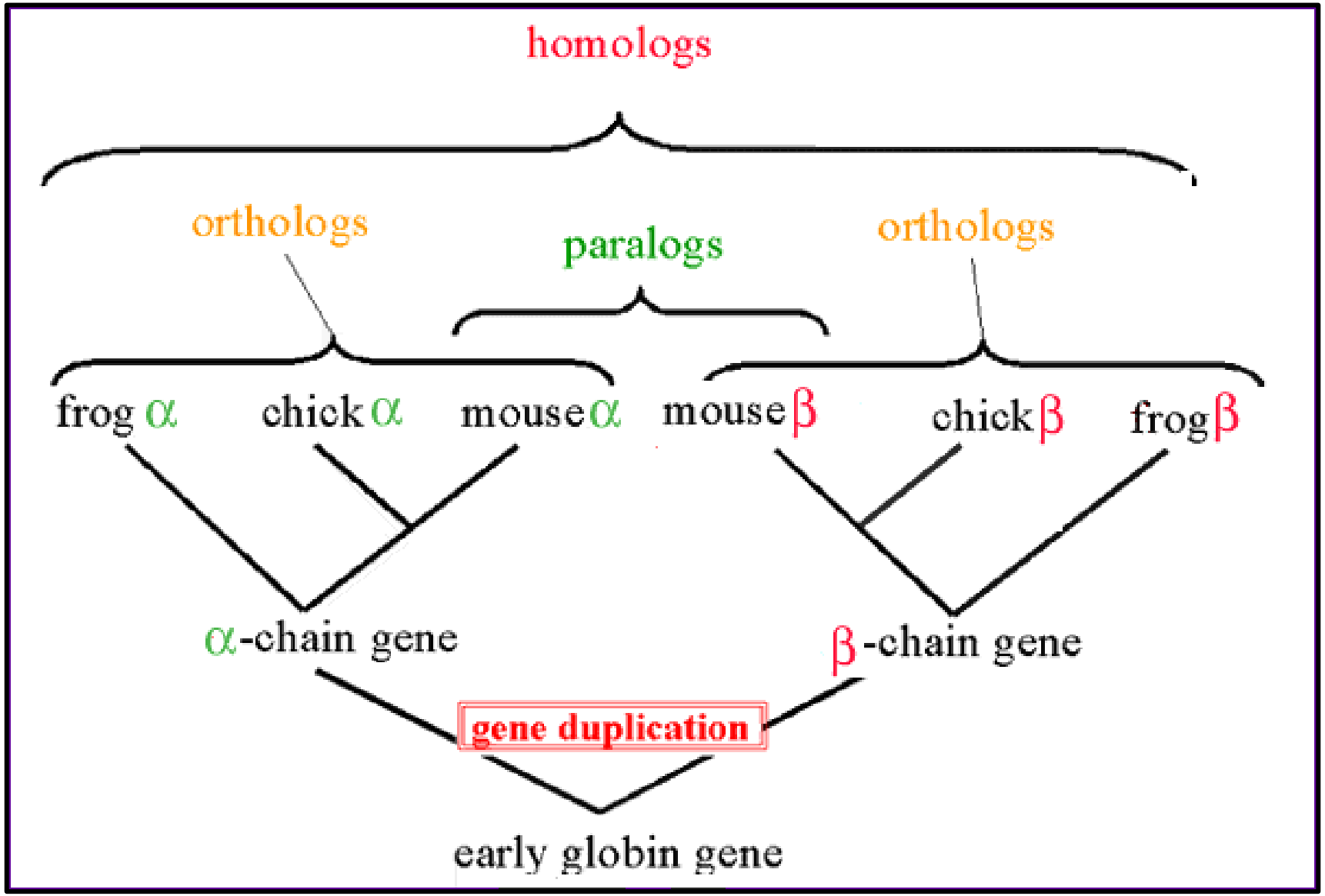
What are protein domains?

- Domains are distinct functional and/or structural units in a protein.
- They are responsible for a particular function or interaction, contributing to the overall role of a protein.
- Domains may exist in a variety of biological contexts, where similar domains can be found in proteins with different functions.



Domain composition of Nck

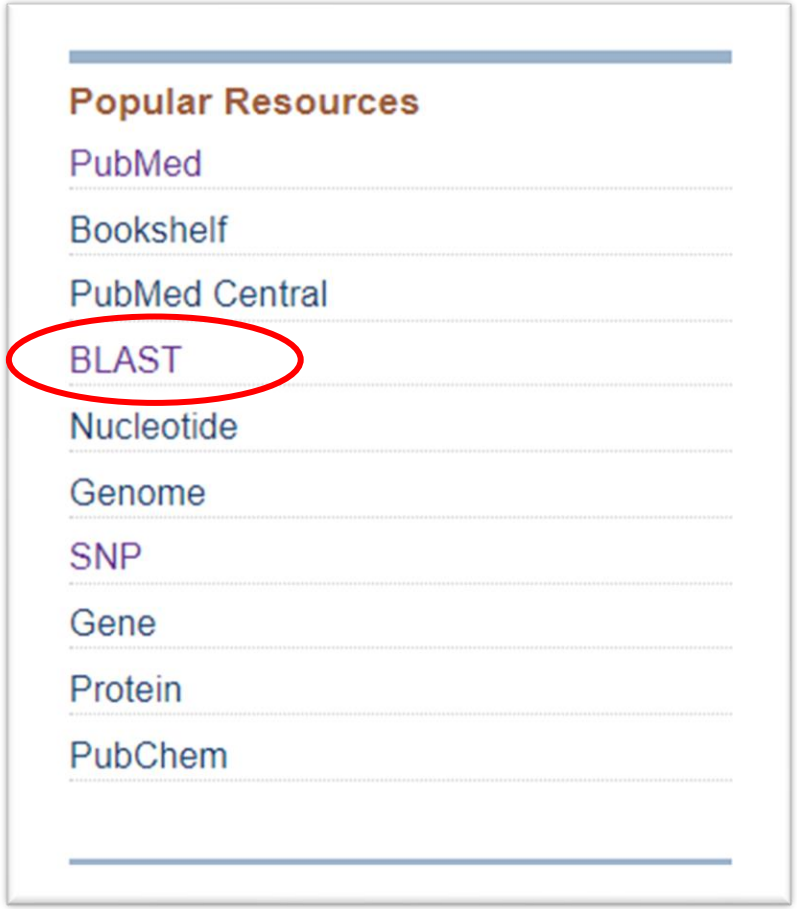
Orthologs and paralogs are often viewed in a single tree

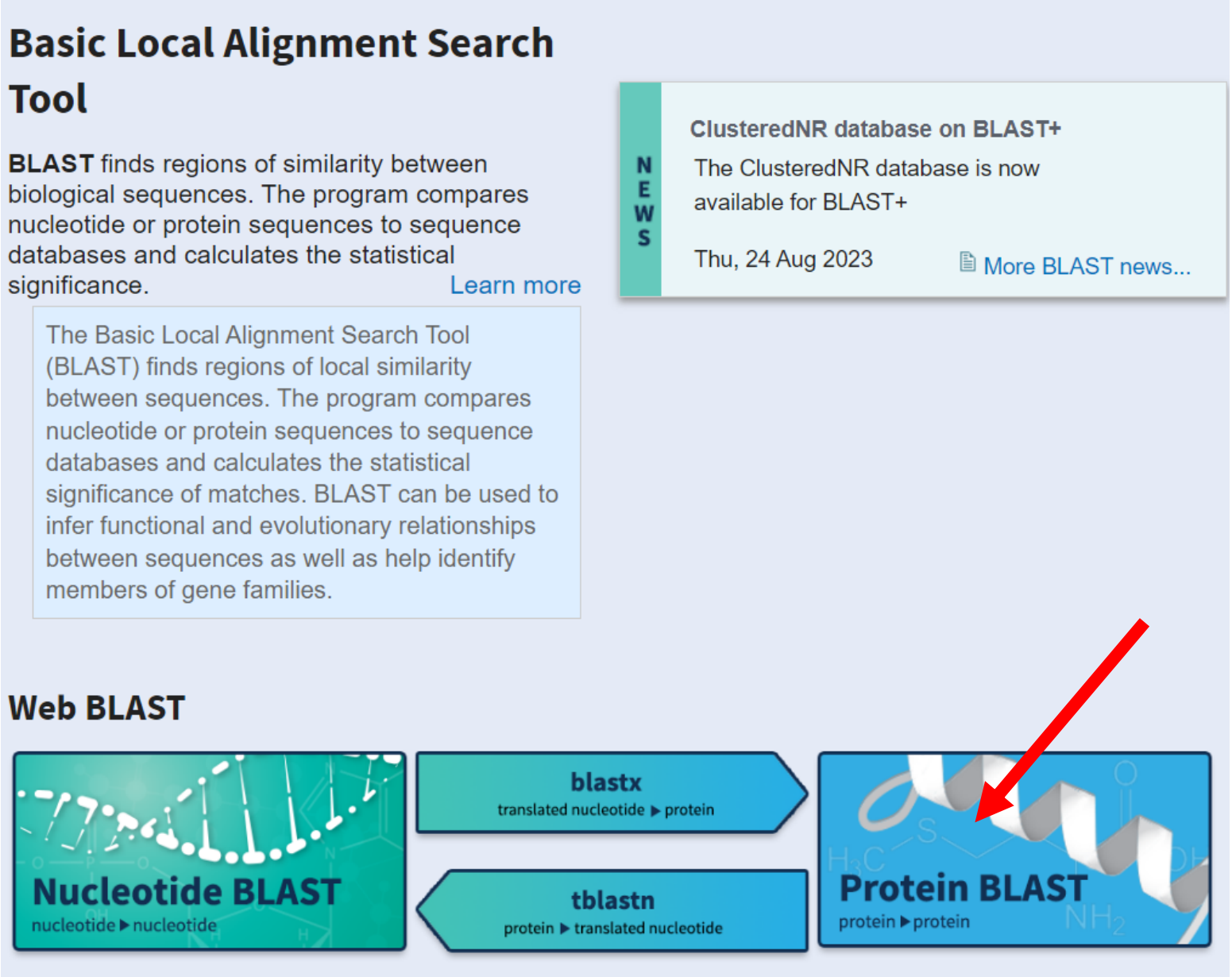


General approach to pairwise alignment

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

Find BLAST from the home page of NCBI and select protein BLAST…





- Choose align two or more sequences…

Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Databases: Standard databases (nr etc.): New Experimental databases

Try experimental clustered nr database
For more info see What is clustered nr?

Compare: Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection

Algorithm:
- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

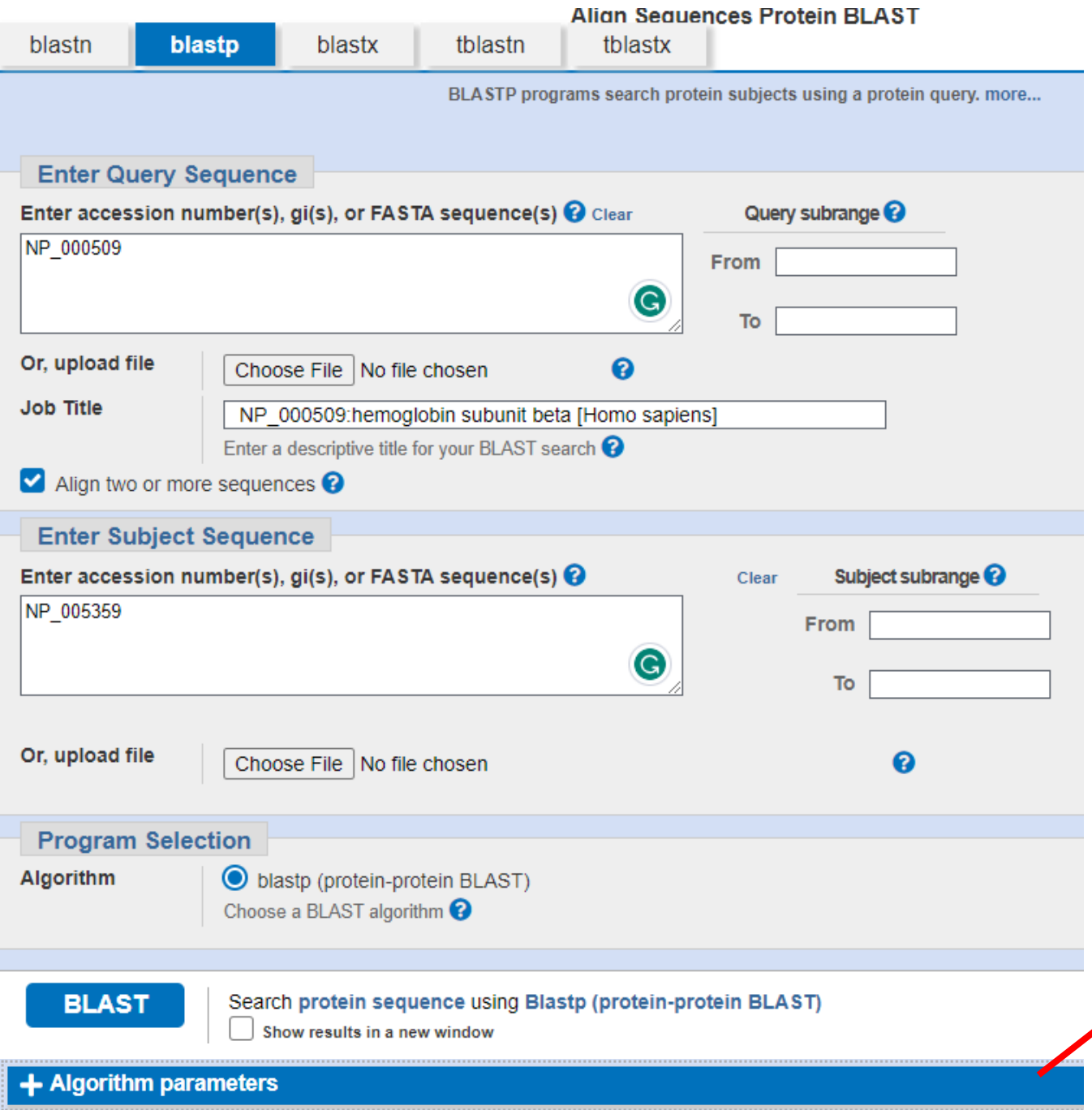
Feedback

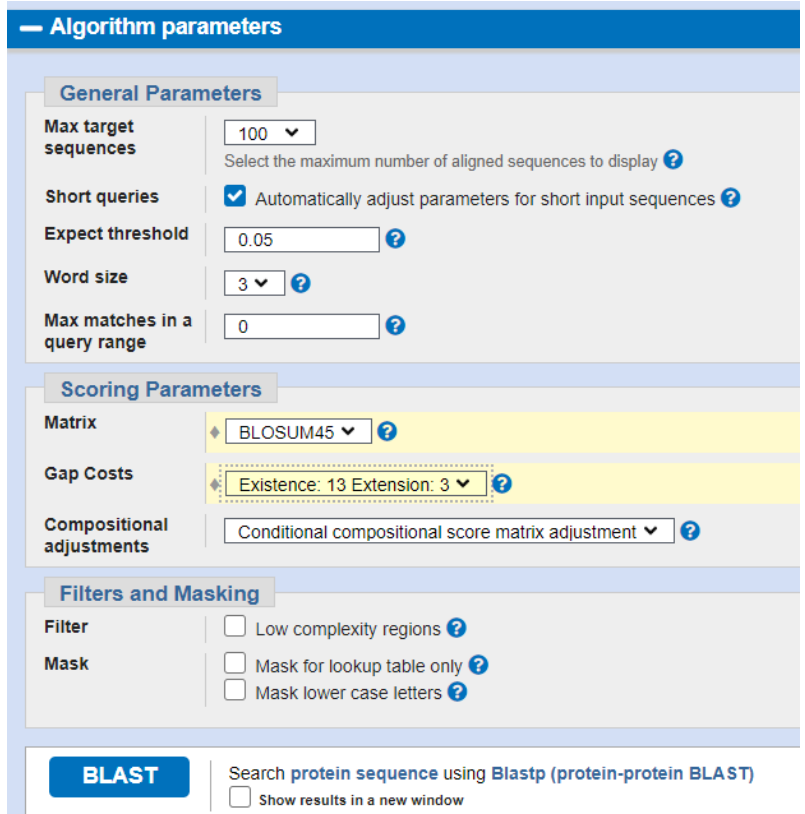
+ Algorithm parameters

Enter the two sequences (as accession numbers or in the fasta format) and click **BLAST**.

NP_000509
NP_005359

Optionally select "Algorithm parameters" and note the matrix option.





Align Sequences Protein BLAST

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein subjects using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

>NP_000509.1 hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST
PDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKE

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

NP_005359

Subject subrange

From

To

Or, upload file Choose File No file chosen

Program Selection

Algorithm ◉ blastp (protein-protein BLAST)

Choose a BLAST algorithm

BLAST Search **protein sequence** using **Blastp (protein-protein BLAST)**

☐ Show results in a new window

+ Algorithm parameters

Pairwise alignment of human beta globin (the “query”) and myoglobin (the “subject”)

myoglobin isoform 1 [Homo sapiens]

Sequence ID: NP_001349775.1 Length: 154 Number of Matches: 1

See 12 more title(s) See all Identical Proteins(IPG)

Range 1: 3 to 147 GenPept Graphics ▼ Next Match ▲ Previous Match

```
 Score = 43.9 bits (102),  Expect = 1e-09, Method: Composition-based stats.
 Identities = 37/145 (25%), Positives = 57/145 (39%), Gaps = 2/145 (1%)

                     ▼                          ▼
Query  4    LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
       ──►  L+  E   V  +WGKV  D    G E L RL   +P T   F+ F  L + D +  +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
       ──►  K HG  VL A    L    + +     L++ H  K  +  +    +   ++ VL
Sbjct  63   KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
       ──►  +F    Q A  K +      +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASNY  147
```

We’ll examine the highlighted green region of the alignment in more detail.

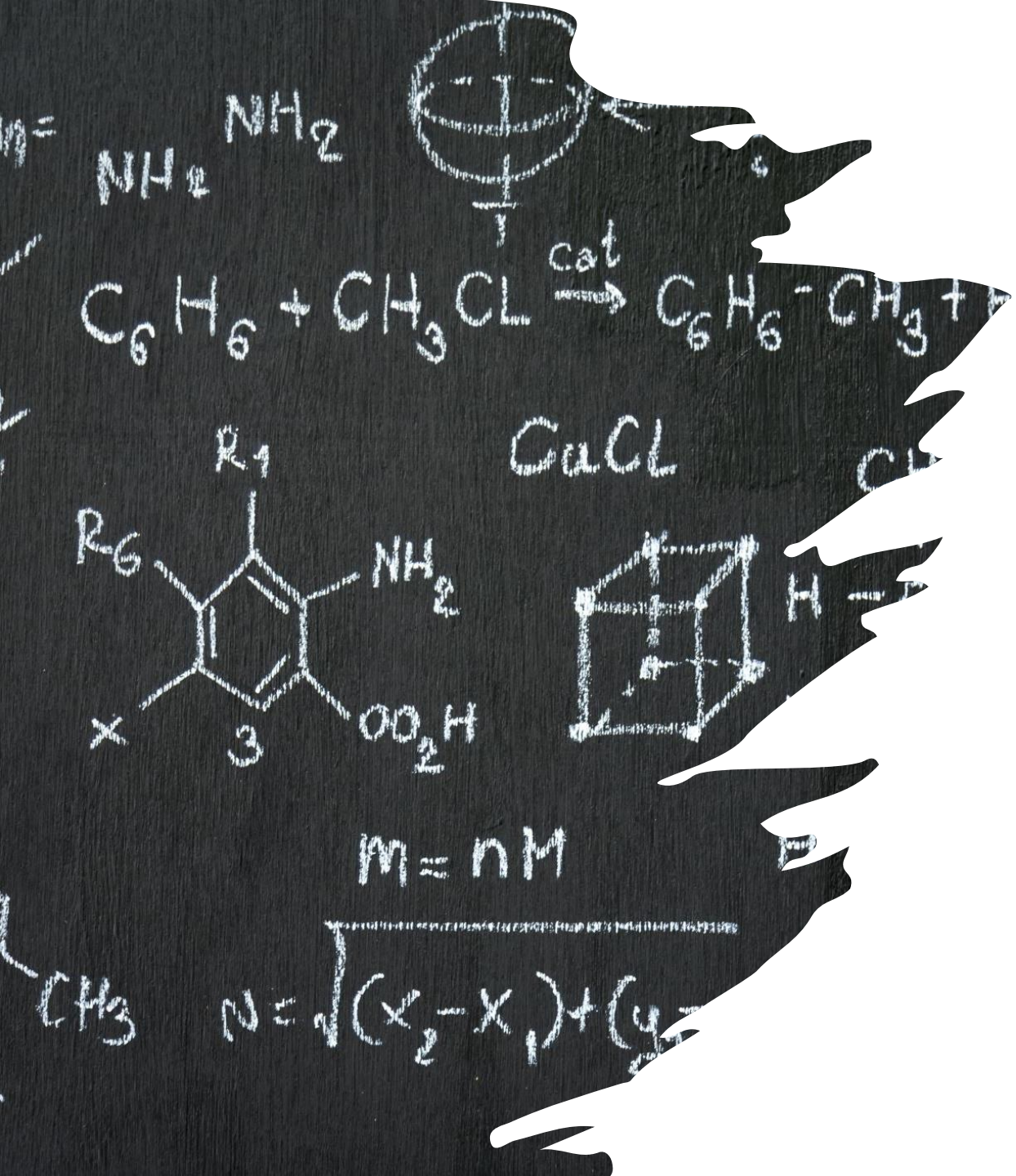
How raw scores are calculated: an example

```
 Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34

match      4  11 5   6    6 5 4 5    sum of matches: +60 (round up to +61)
                6 4               4
mismatch   -1 1    0   -2 -2  -4  0  sum of mismatches: -13
            -2      0   -3   0
gap open             -11             sum of gap penalties: -13
gap extend            -2
                                     total raw score: 61 - 13 - 13 = 35
```

For a set of aligned residues, we assign scores based on matches, mismatches, gap open penalties, and gap extension penalties. These scores add up to the total raw score.



Where do scores come from?

We'll examine scoring matrices. These are related to the properties of the 20 common amino acids.

Gap

- Positions at which a letter is paired with a null are called gaps.

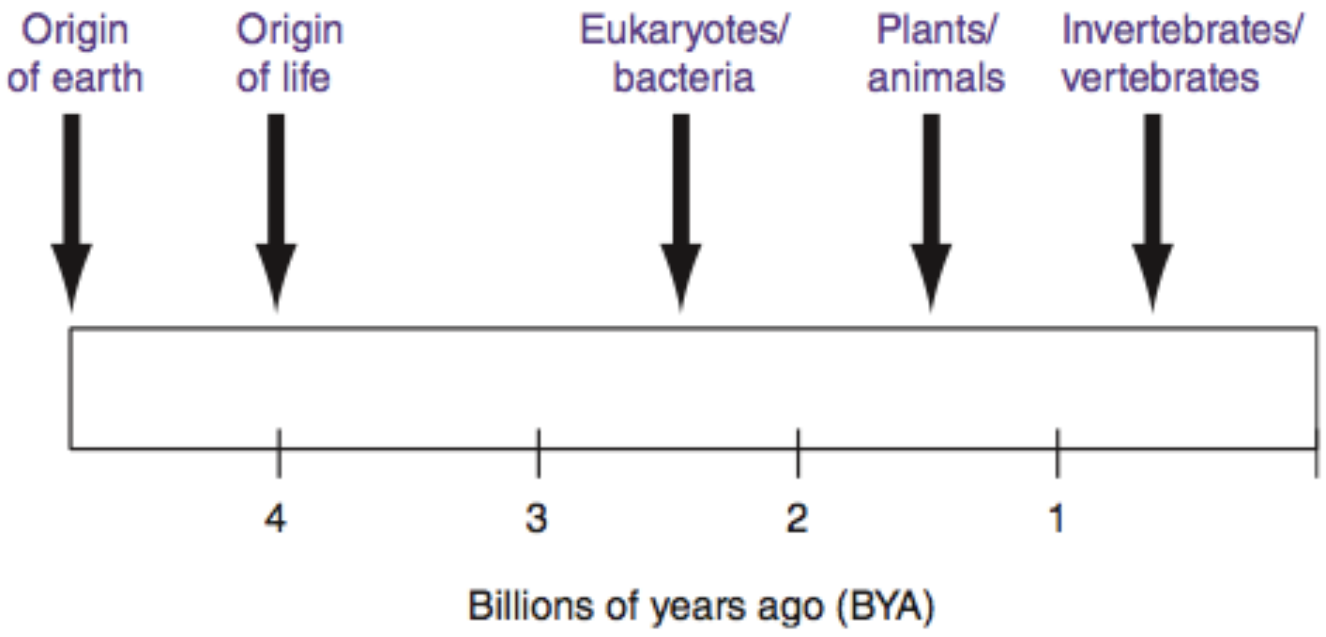
- Gap scores are typically negative.

- 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. Thus, there are separate penalties for gap creation and gap extension.

- In BLAST, it is rarely necessary to change gap values from the default.

Pairwise alignment and the evolution of life

- When two proteins (or DNA sequences) are homologous they share a common ancestor.
- We can infer the sequence of that ancestor.
- When we align globins from human and a plant we can imagine their common ancestor, a single celled organism that lived 1.5 billion years ago, and we can infer that ancient globin sequence.
- Through pairwise alignment we can look back in time at sequence evolution.

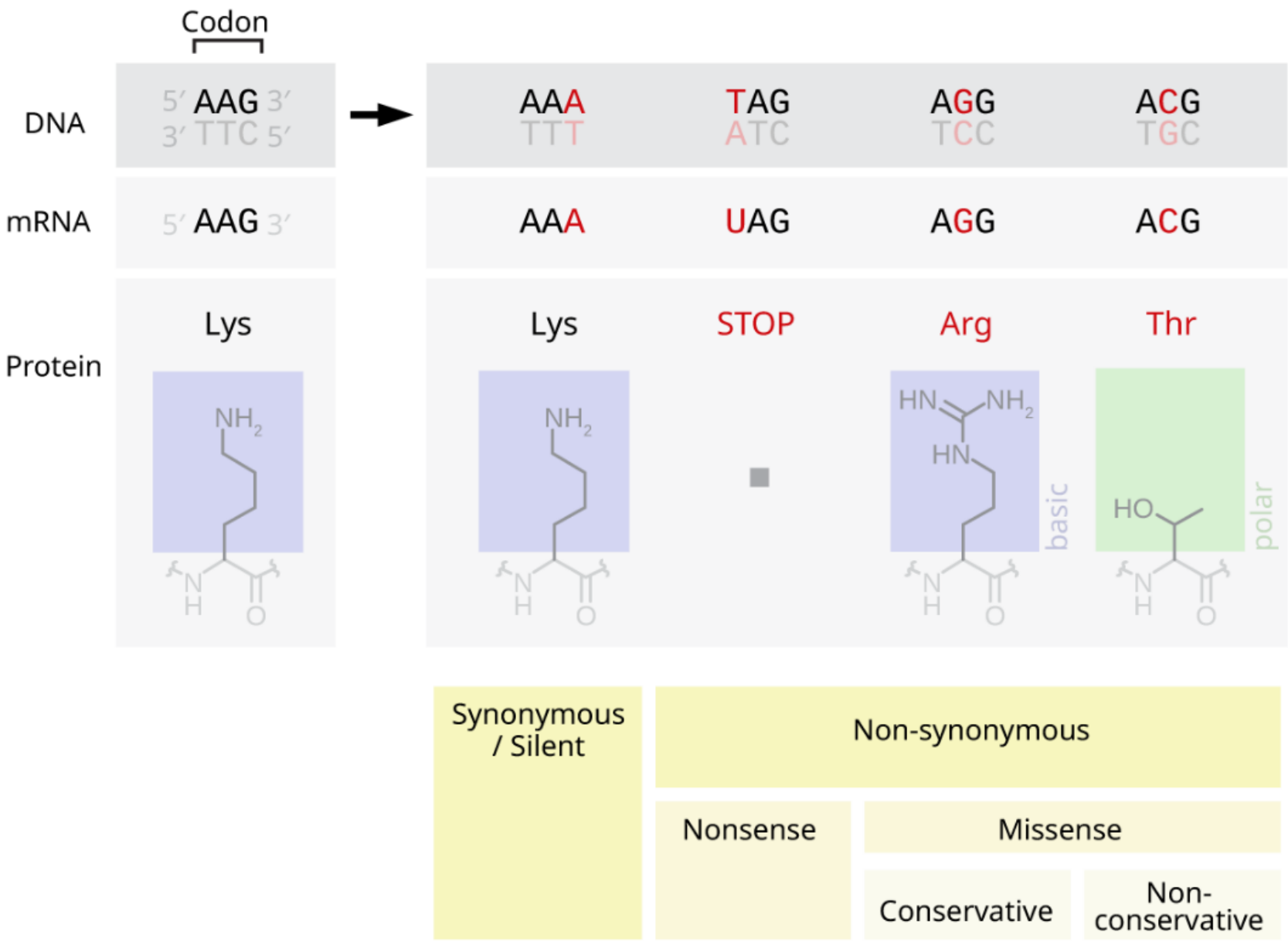


Scoring matrices -> Dayhoff model: 7 steps

Step 1: Accepted point mutations (PAMs) in protein families

PROTEIN	PAMS PER 100 MILLION YEARS
Immunoglobulin (Ig) kappa chain C region	37
Kappa casein	33
Epidermal growth factor	26
Serum albumin	19
Hemoglobin alpha chain	12
Myoglobin	8.9
Nerve growth factor	8.5
Trypsin	5.9
Insulin	4.4
Cytochrome c	2.2
Glutamate dehydrogenase	0.9
Histone H3	0.14
Histone H4	0.10

Margaret Dayhoff and colleagues developed scoring matrices in the 1960s and 1970s. They defined PAMs as “accepted point mutations.” Some protein families evolve very slowly (e.g. histones change little over 100 million years); others (such as kappa casein) change very rapidly.



An example of point mutations at an amino acid site coding for lysine. The missense mutations may be classed as **point accepted mutations** if the **mutated protein is not rejected by natural selection** (Source: Wikipedia).

Dayhoff's 34 protein superfamilies

Protein	PAMs per 100 million years
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 a	12
Hemoglobin b	12

Dayhoff's 34 protein superfamilies

Protein	PAMs per 100 million years
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	
carbonic anhydrase C	16
Hemoglobin a	12
Hemoglobin b	12

human (NP_005203) versus mouse (NP_031812) kappa casein

```
 Score = 57.8 bits (138),  Expect = 3e-07
 Identities = 39/118 (33%), Positives = 61/118 (51%), Gaps = 2/118 (1%)

Query  1    MKSFLLVVNALALTLPFLAVEVQNQKQPACHENDERPFYQKTAPYVPMYYVPNSYPYYGT  60
            M++F++V+N LALTLPFLA E+QN       E ++  + ++   Y P+  V N +  Y
Sbjct  2    MRNFIVVMNILALTLPFLAAEIQNPDSNCRGEKNDIVYDEQRVLYTPVRSVLN-FNQYEP  60

Query  61   NLYQRRPAI-AINNPYVPRTYYANPAVVRPHAQIPQRQYLPNSHPPTVVRLPNLHPSF  117
            N Y  RP++ A  +PY+         ++R  A I + Q +PN      V     +PSF
Sbjct  61   NYYHYRPSLPATASPYMYYPLVVRLLLLRSPAPISKWQSMPNFPQSAGVPYAIPNPSF  118
```

Dayhoff's 34 protein superfamilies

Protein	PAMs per 100 million years
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
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

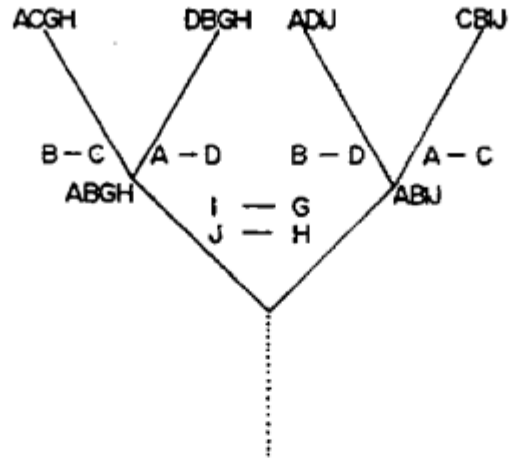


Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Figure 79. Matrix of accepted point mutations derived from the tree of Figure 78.

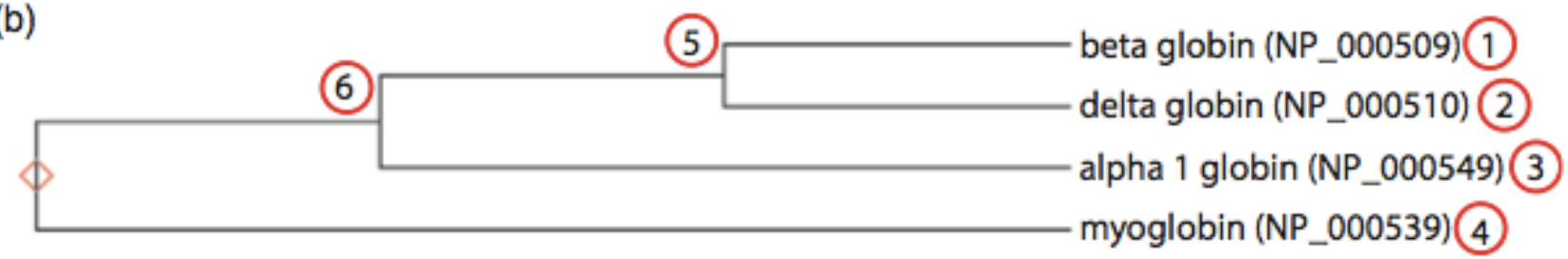
https://chagall.med.cornell.edu/BioinfoCourse/PDFs/Lecture2/Dayhoff1978.pdf

Step 1: accepted point mutations are defined not by the pairwise alignment but with respect to the common ancestor

(a)

```
beta globin      MVHLTPEEKSAVTALWGKV
delta globin     MVHLTPEEKTAVNALWGKV
alpha 1 globin   MV.LSPADKTNVKAAWGKV
myoglobin        .MGLSDGEWQLVLNVWGKV
5                MVHLSPEEKTAVNALWGKV
6                MVHLTPEEKTAVNALWGKV
```

(b)



Dayhoff et al. evaluated amino acid changes. They applied an evolutionary model to compare changes such as 1 versus 2 not to each other but to an inferred common ancestor at position 5.

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
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												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	y	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

From a survey of 1572 observed substitutions, the original amino acid (columns) are compared to the changes (rows).

Dayhoff model step 2 (of 7): Frequency of amino acids

TABLE 3.1 Normalized frequencies of amino acid. These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence.

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

If 20 amino acids occurred in nature at equal frequencies, each would be observed 5% of the time. However, some are more common (G, A, L, K) and some rare (C, Y, M, W).

Normalized frequencies of amino acids: we need these values to calculate denominator $p_i p_j$

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 in the genetic code
- These frequencies f_i sum to 1

Dayhoff model step 3: amino acid substitutions

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
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												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	y	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

From a survey of 1572 observed substitutions, the original amino acid (columns) are compared to the changes (rows).

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	Ala																				
R	Arg	30																			
N	Asn	109	17																		
D	Asp	154	0	532																	
C	Cys	33	10	0	0																
Q	Gln	93	120	50	76	0															
E	Glu	266	0	94	831	0	422														
G	Gly	579	10	156	162	10	30	112													
H	His	21	100	226	43	10	243	23	10												
I	Ile	66	30	36	13	17	8	35	0	3											
L	Leu	95	17	37	0	0	75	15	17	40	253										
K	Lys	57	477	322	85	0	147	104	60	23	43	39									
M	Met	29	17	0	0	0	20	7	7	0	57	207	90								
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

Figure 80. Numbers of accepted point mutations (X 10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

https://chagall.med.cornell.edu/BioinfoCourse/PDFs/Lecture2/Dayhoff1978.pdf

Dayhoff model step 3: amino acid substitutions

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G 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

Zooming in on the previous table, note that substitutions are very common (e.g. D → E, A → G) while others are rare (e.g. C → Q, C → E). The scoring system we use for pairwise alignments should reflect these trends.

Dayhoff model step 3: Relative mutability of amino acids

TABLE 3.2 Relative mutabilities of amino acids. The value of alanine is arbitrarily set to 100.

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

Dayhoff et al. used (1) data on the frequency of amino acids and (2) data on observed and inferred numbers of substitutions to determine the relative mutability of amino acids. In a scoring system alignment of two tryptophans will be weighted more heavily than two asparagines.

Dayhoff step 4 (of 7): Mutation probability matrix for the evolutionary distance of 1 PAM

		Original amino acid																			
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
Replacement amino acid	A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2
	R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
	N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0
	D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1
	H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.3
	L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	99.5	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.2
	K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	99.3	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
	M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	99.5	0.0	0.0	0.0	0.0	0.3	0.0
	P	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	99.3	0.1	0.0	0.0	0.0	0.0
	S	0.3	0.1	0.3	0.1	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.2	98.4	0.4	0.1	0.0	0.0
	T	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.3	98.7	0.0	0.0	0.1
	W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	0.0
	Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	99.5	0.0
	V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	99.0

This mutation probability matrix includes original amino acids (columns) and replacements (rows). The diagonals show that at a distance of 1 PAM most residues remain the same about 99% of the time (see shaded entries). Note how cysteine (C) and tryptophan (W) undergo few substitutions, and asparagine (N) many.

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.** At an evolutionary interval of PAM1, one change has occurred over a length of 100 amino acids.

- Other PAM matrices are extrapolated from PAM1. For PAM250, 250 changes have occurred for two proteins over a length of 100 amino acids.

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

Dayhoff step 4 (of 7): Mutation probability matrix for the evolutionary distance of 1 PAM

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His
A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0
R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1
N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2
D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0
C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0
Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2
E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0
G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0
H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1

At this evolutionary distance of 1 PAM, 1% of the amino acids have diverged between each pair of sequences. The columns are percentages that sum to 100%.

Dayhoff step 5 (of 7): PAM250 and other PAM matrices

```
                                 ▼ ▼ ▼ ▼ ▼       ▼▼              ▼ ▼   ▼   ▼ ▼▼
NP_002037.2          164   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII   207
XP_001162057.1       164   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII   207
NP_001003142.1       162   IHDHFGIVEGLMTTVHAITATQKTVDGPSGKMWRDGRGAAQNII   205
XP_893121.1          168   IHDNFGIMEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   211
XP_576394.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_058704.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
XP_001070653.1       162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
XP_001062726.1       162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_989636.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_525091.1          161   INDNFEIVEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII   204
XP_318655.2          161   INDNFGILEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII   204
NP_508535.1          170   INDNFGIIEGLMTTVHAVTATQKTVDGPSGKLWRDGRGAGQNII   213
NP_595236.1          164   INDTFGIEEGLMTTVHATTATQKTVDGPSKKDWRGGRGASANII   207
NP_011708.1          162   INDAFGIEEGLMTTVHSLTATQKTVDGPSHKDWRGGRTASGNII   205
XP_456022.1          161   INDEFGIDEALMTTVHSITATQKTVDGPSHKDWRGGRTASGNII   204
NP_001060897.1       166   IHDNFGIIEGLMTTVHAITATQKTVDGPSSKDWRGGRAASFNII   209
```

Consider a multiple alignment of glyceraldehyde 3-phosphate protein sequences. Some substitutions are observed in columns (arrowheads). These give us insight into changes tolerated by natural selection.

Dayhoff step 5 (of 7): PAM250 and other PAM matrices

```
mouse     AIPNPSFLAMPTNENQDNTAIPTIDPITPIVST--PVPTM------ESIVNTVANPEAST
rabbit    S--HPFFMAILPNKMQDKAVTPTTNTIAAVEPT--PIPTT------EPVVSTEVIAEASP
sheep     PHPHLSFMAIPPKKDQDKTEIPAINTIASAEPTVHSTPTT------EAVVNAVDNPEASS
cattle    PHPHLSFMAIPPKKNQDKTEIPTINTIASGEPT--STPTT------EAVESTVATLEDSP
pig       PRPHASFIAIPPKKNQDKTAIPAINSIATVEPT--IVPATEPIVNAEPIVNAVVTPEASS
human     PNLHPSFIAIPPKKIQDKIIIPTINTIATVEPT--PAPAT------EPTVDSVVTPEAFS
horse     PCPHPSFIAIPPKKLQEITVIPKINTIATVEPT--PIPTP------EPTVNNAVIPDASS
          .  :  *:*: .:: *:    *  :.*:.  .*    *:       *.  .     :  .
```

Now consider the alignment of distantly related kappa caseins. There are few conserved column positions, and many some columns (double arrowheads) have five different residues among the 7 proteins.

<u>We want to design a scoring system that is tolerant of distantly related proteins: if the scoring system is too strict then the divergent sequences may be penalized so heavily that authentic homologs are not identified or aligned.</u>

Dayhoff step 5 (of 7): PAM250 and other PAM matrices

original amino acid

replacement amino acid replacement amino acid

PAM0	A	R	N	D	C	Q	E	G
A	100	0	0	0	0	0	0	0
R	0	100	0	0	0	0	0	0
N	0	0	100	0	0	0	0	0
D	0	0	0	100	0	0	0	0
C	0	0	0	0	100	0	0	0
Q	0	0	0	0	0	100	0	0
E	0	0	0	0	0	0	100	0
G	0	0	0	0	0	0	0	100

original amino acid

PAM∞	A	R	N	D	C	Q	E	G
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

At the extreme of perfectly conserved proteins (PAM0) there are no amino acid replacements. At the extreme of completely diverged proteins (PAM∞) the matrix converges on the background frequencies of the amino acids.

PAM250 matrix: for proteins that share ~20% identity

		Original amino acid																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Replacement amino acid	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

Compare this to a PAM1 matrix, and note the diagonal still has high scores but much information content is lost.

Table 23

Correspondence between Observed Differences and the Evolutionary Distance

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

https://chagall.med.cornell.edu/BioinfoCourse/PDFs/Lecture2/Dayhoff1978.pdf

Dayhoff step 6 (of 7): from a mutation probability matrix to a relatedness odds matrix

$$R_{ij} = \frac{M_{ij}}{f_i}$$

1. A **relatedness odds matrix** reports the probability that amino acid ***j*** will change to ***i*** in a homologous sequence.

2. The numerator models the observed change. The denominator *fi* is the probability of amino acid residue *i* occurring in the second sequence by chance.

3. A positive value indicates a replacement happens more often than expected by chance. A **negative value** indicates the **replacement is not favored**.

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.

- 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,

$$s_{i,j} = 10 \times \log\left(\frac{q_{i,j}}{p_{i,j}}\right)$$

$S(trp,trp) = 10 \log_{10} (0.55/0.010) = 17.4$

What do the numbers mean in a log odds matrix?

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

A score of 0 is neutral.

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.

Dayhoff step 7 (of 7): log odds scoring matrix

$$s_{ij} = 10 \times \log_{10}\left(\frac{M_{ij}}{f_i}\right).$$

- A log odds matrix is the logarithmic form of the relatedness odds matrix.
- ***sij*** <u>is the score for aligning any two residues in a pairwise alignment.</u> (There is also a score for aligning a residue with itself.)
- ***Mij*** <u>is of the observed frequency of substitutions for each pair of amino acids.</u> These values ("target frequencies") are derived from a mutation probability matrix.

Example of a score for aligning cysteine and leucine using the values in a PAM250 scoring matrix

$$s_{(\text{cysteine, leucine})} = 10 \times \log_{10}\left(\frac{0.02}{0.085}\right) = -6.3$$

Log-odds matrix for PAM250

6									
-3	5								
4	0	6							
2	-5	0	9						
-3	-1	-2	-5	6					
-3	0	-2	-3	1	2				
-2	0	-1	-3	0	1	3			
-2	-3	-4	0	-6	-2	-5	17		
-1	-4	-2	7	-5	-3	-3	0	10	
2	-2	2	-1	-1	-1	0	-6	-2	4
L	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

This is a useful matrix for comparing **distantly related proteins**. Note that an alignment of two tryptophan (W) residues earns +17 and a W to T mismatch is -5.

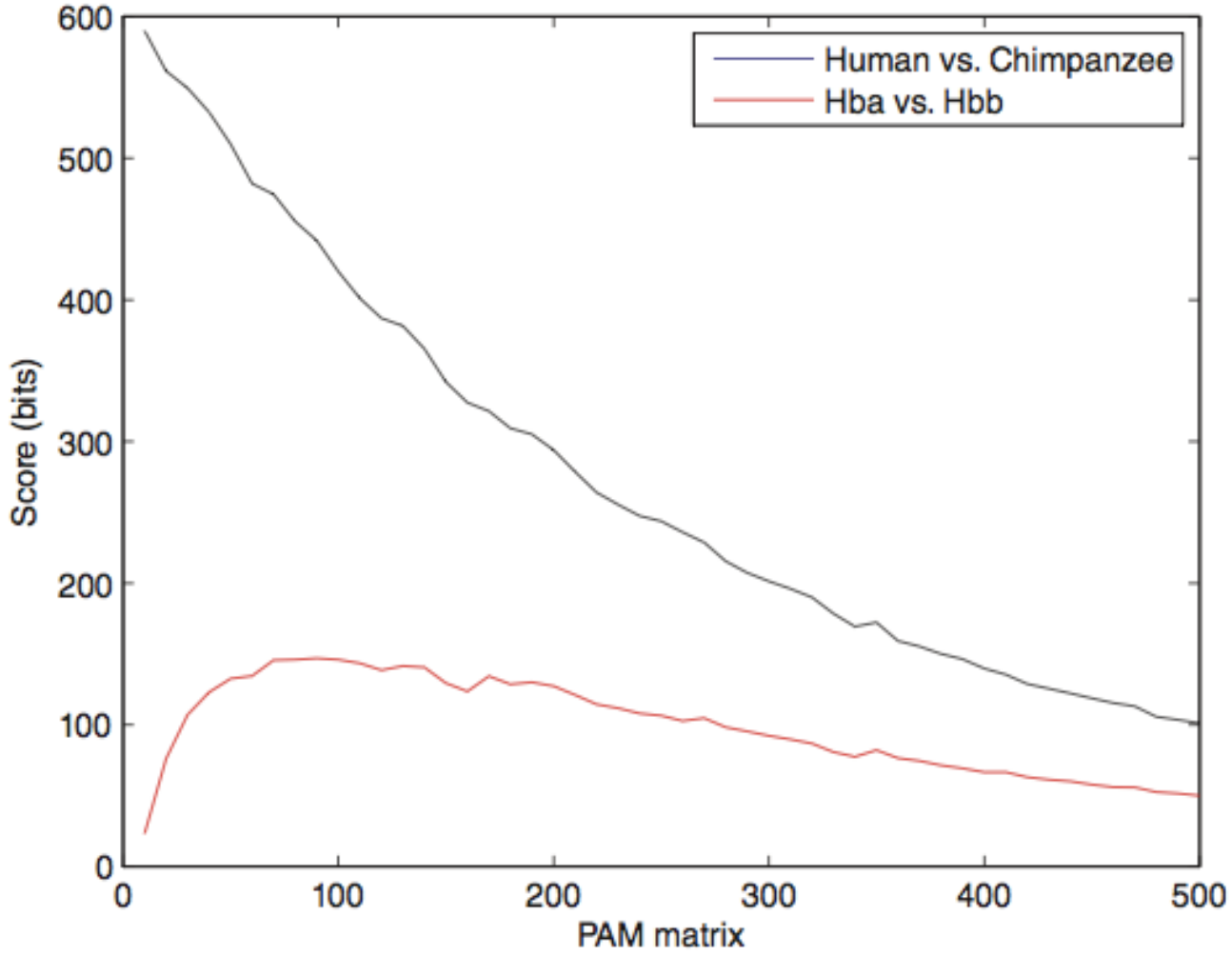
Log-odds matrix for PAM10

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	-2	-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

More closely related proteins

This is an example of a scoring matrix **with "severe" penalties**. A match of W to W earns +13, but a mismatch (e.g. W aligned to T) has a score of -19, far lower than in PAM250.

Effect of scoring matrix on bit scores



Look at score for distantly related proteins (e.g. beta globin versus alpha globin) and note that PAM10 or similar matrices assign very low scores. This effect is not seen for very closely related proteins (e.g. a chimp vs. human globin).

PAM matrices are based on data from the alignment of closely related protein families, and they involve the assumption that substitution probabilities for highly related proteins (e.g., PAM40) can be extrapolated to probabilities for distantly related proteins (e.g., PAM250).

BLOSUM matrices are based on empirical observations of more distantly related protein alignments.

https://www.youtube.com/watch?v=0_66UK-439M&t=13s

BLOSUM62 scoring matrix

blocks substitution matrix

A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

BL62 is the default scoring matrix at the NCBI BLAST site.

Derivation of BLOSUM matrices

BLOSUM matrices are derived from comparisons of blocks of sequences from the Blocks database.

What are blocks and what is the blocks database?

A *block* is an ungapped multiple alignments of highly conserved, short regions. Here is what a sample block looks like:

```
AABCDA...BBCDA
DABCDA.A.BBCBB
BBBCDABA.BCCAA
AAACDAC.DCBCDB
CCBADAB.DBBDCC
AAACAA...BBCCC
```

Conserved blocks in alignment

The blocks database contains multiple alignments of conserved regions in protein families.

https://snipcademy.com/pairwise-alignment#blosum---blocks-substitution-matrix

BLOSUM Matrices

BLOSUM matrices are based on local alignments.

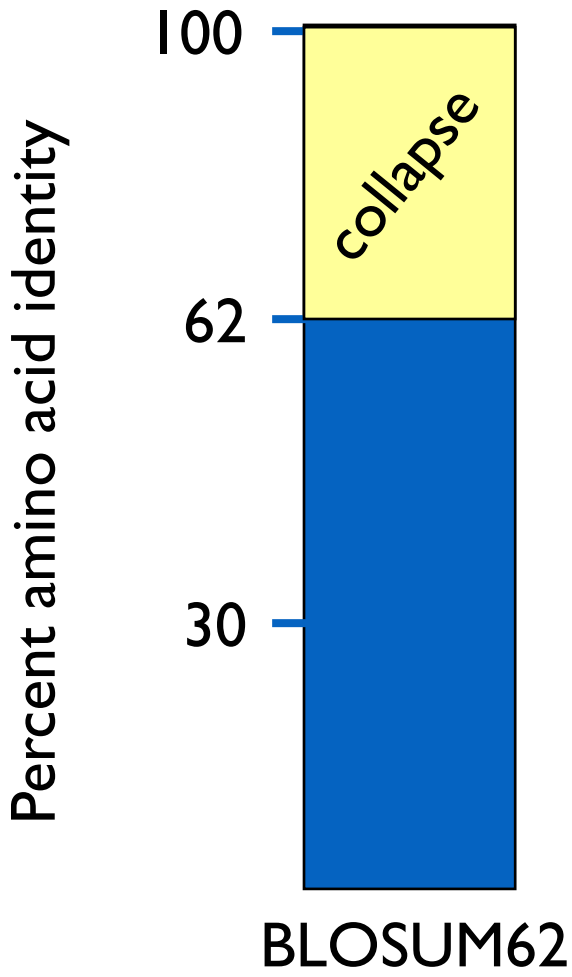
All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins.

BLOSUM stands for blocks substitution matrix.

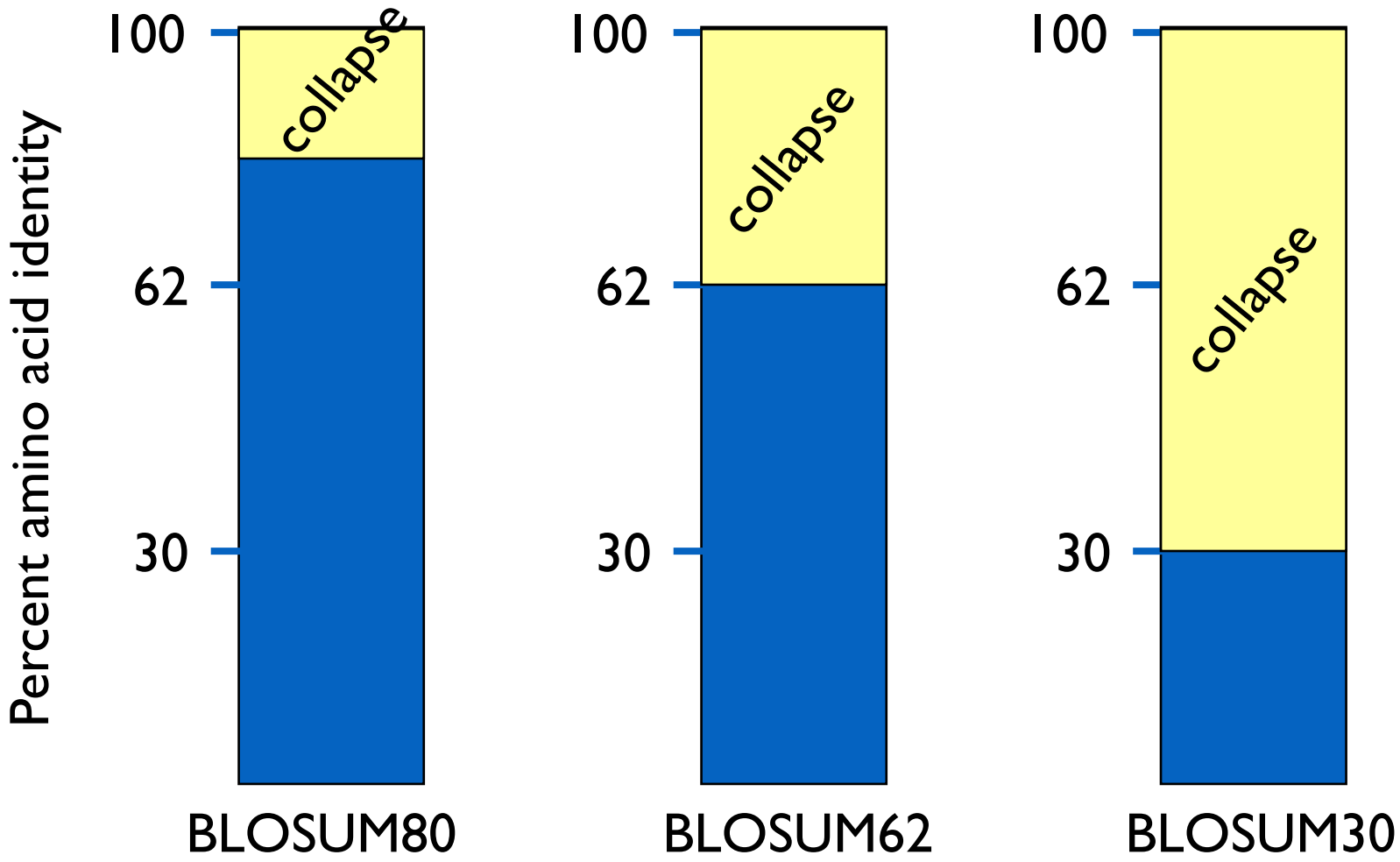
BLOSUM62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.

BLOSUM62 is the default matrix in BLAST 2.0.

BLOSUM Matrices



BLOSUM Matrices



PAM matrices with similar BLOSUM matrices (relative entropy of each PAM matrix is from Altschul 1991)[18]

PAM matrix	Equivalent BLOSUM matrix	Relative entropy (bits)
PAM100	Blosum90	1.18
PAM120	Blosum89	0.98
PAM160	Blosum60	0.70
PAM200	Blosum52	0.51
PAM250	Blosum45	0.36

Source: Wikipedia

Summary of PAM and BLOSUM matrices

BLOSUM90	BLOSUM62	BLOSUM45
PAM30	PAM120	PAM250
Less divergent	← →	More divergent
Human versus chimpanzee beta globin		Human versus bacterial globins

A higher PAM number, and a lower BLOSUM number, tends to correspond to a matrix tuned to more divergent proteins.