BCB 5200 Introduction to Bioinformatics

Pairwise Sequence Alignment

Bioinformatics and Computational Biology
Saint Louis University

Biological question:

- How can we determine the similarity between sequences?

- Why is it important?

- Fundamental rules are:

  Similar sequence → Common ancestor ("*Homology*")

  Similar sequence → Similar structure → Similar function
  (The "*Sequence-to-Structure-to-Function Paradigm*")

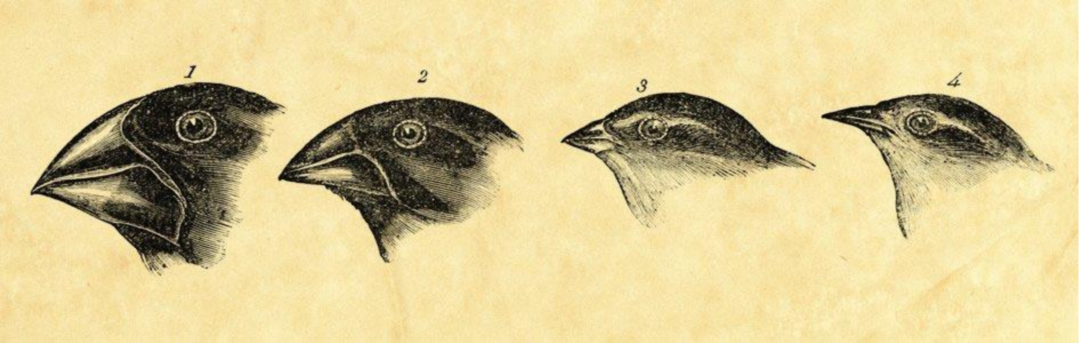
Learning objectives

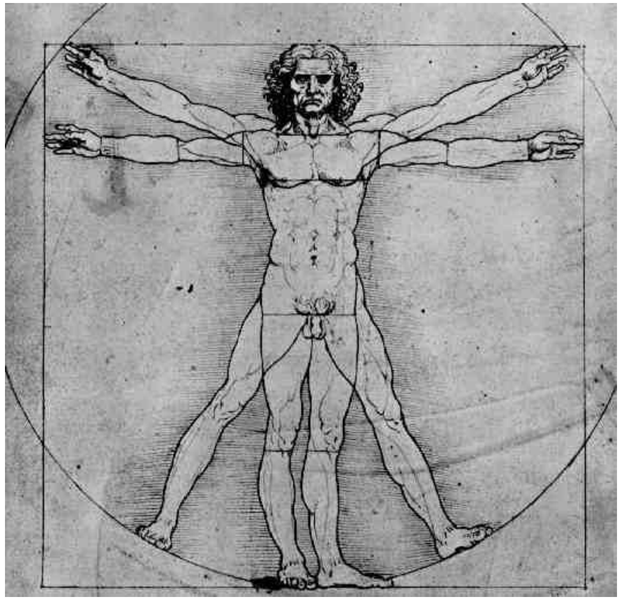
- Define homology as well as orthologs and paralogs
- Basic ideas about sequence alignment
- Contrast the utility of PAM and BLOSUM scoring matrices
- Uncover similar sequence regions using Dot matrix analysis
- Define dynamic programming (DP) and explain how global (Needleman–Wunsch) and local (Smith–Waterman) pairwise alignments are performed
- Perform pairwise alignment of protein or DNA sequences using DP-based methods

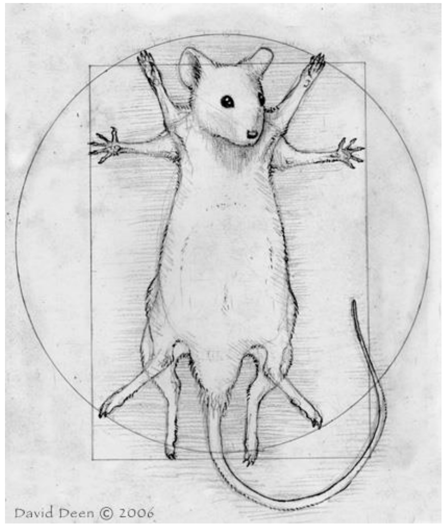
Outlines

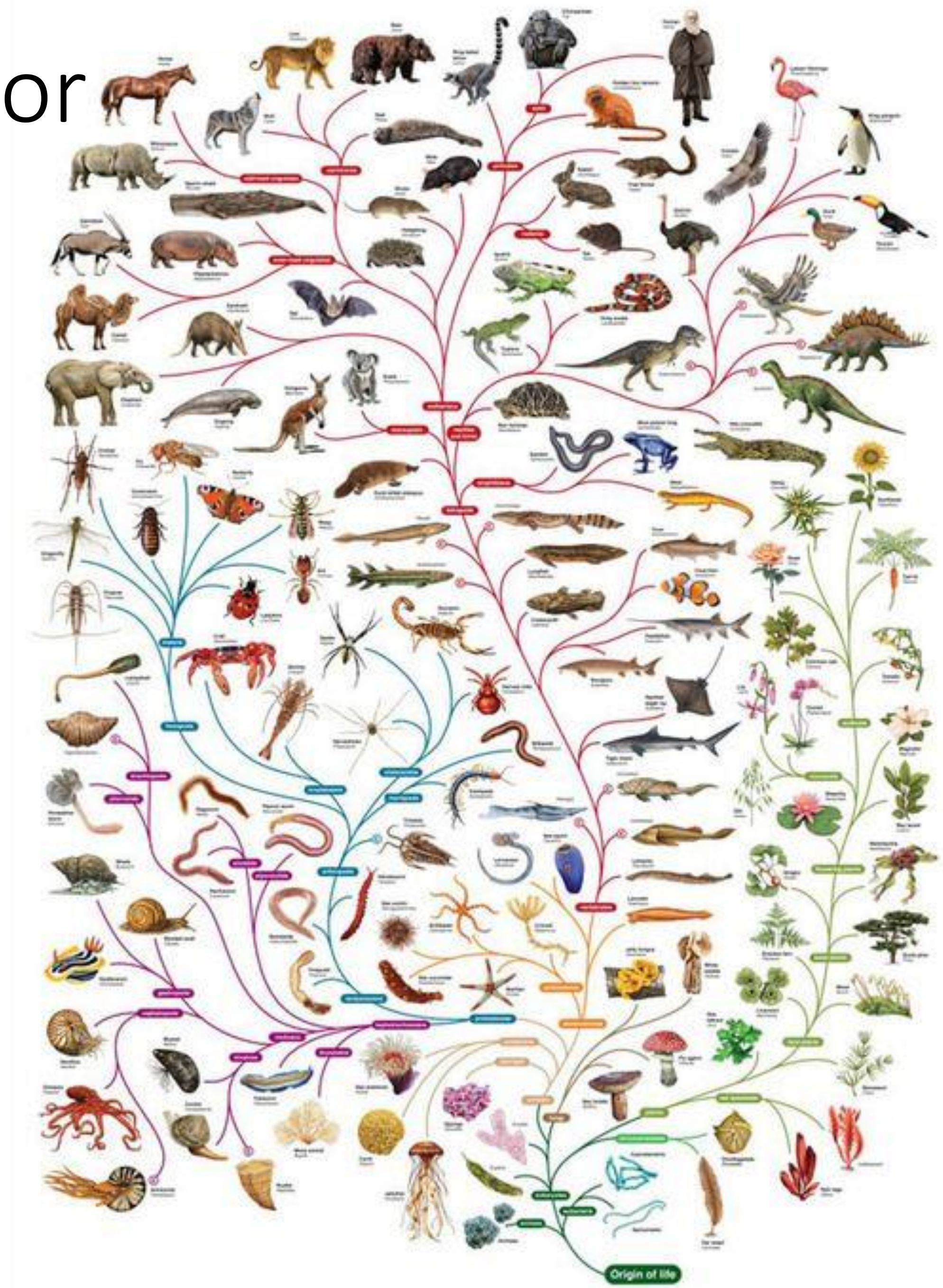
- Similarity & Homology
- Basic components of sequence alignment
  - Similarity or scoring Matrix
  - Gap penalties
- Dot matrix analysis
- Dynamic programming algorithm
  - Global sequence alignment: Needleman-Wunsch (NW) algorithm
  - Local sequence alignment: Smith-Waterman (SW) algorithm

Similarity is the primary indicator for evolutionary relationship

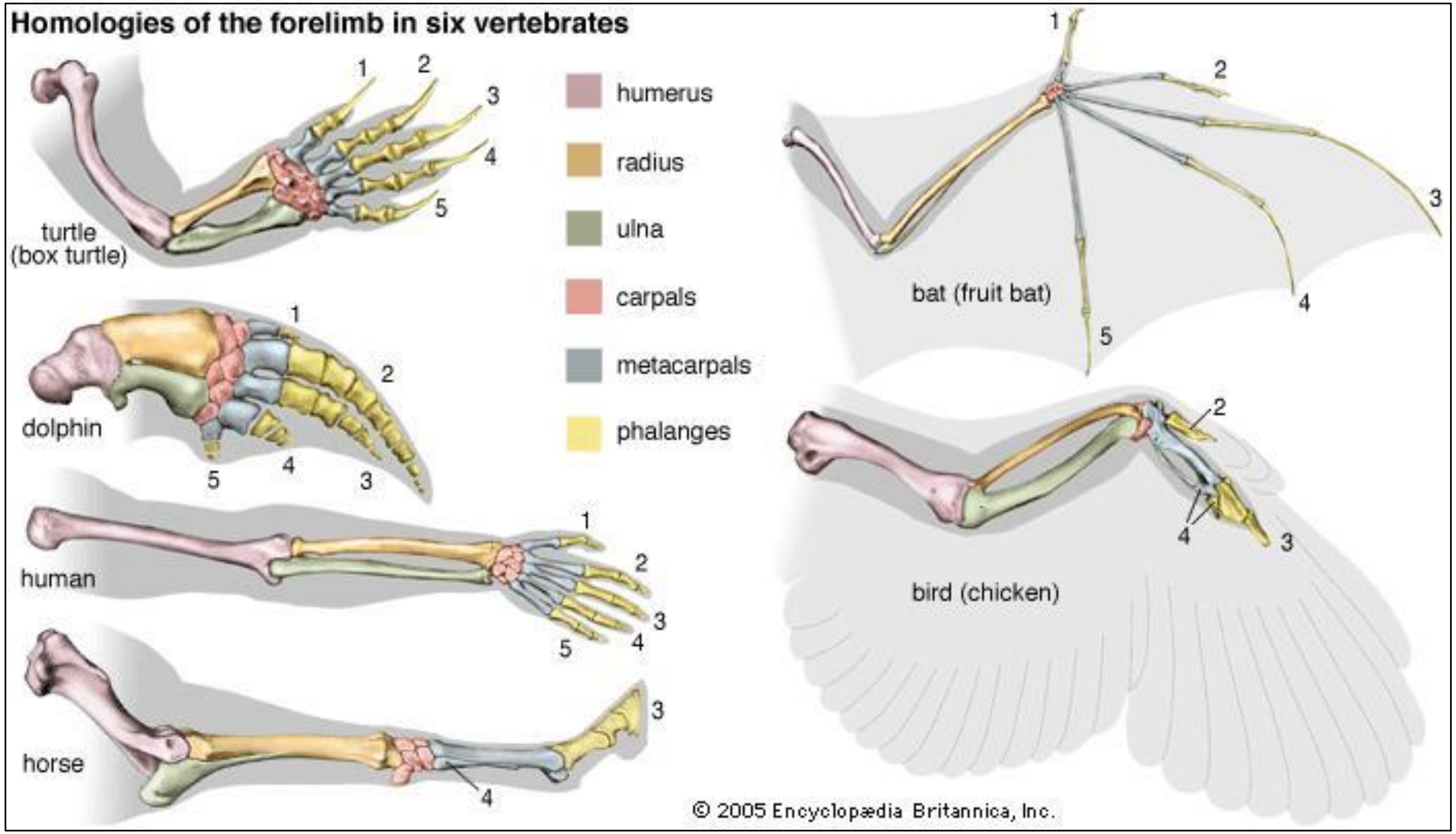





David Deen © 2006



Homologies: similar characteristics (position or structure) derived from a common ancestor

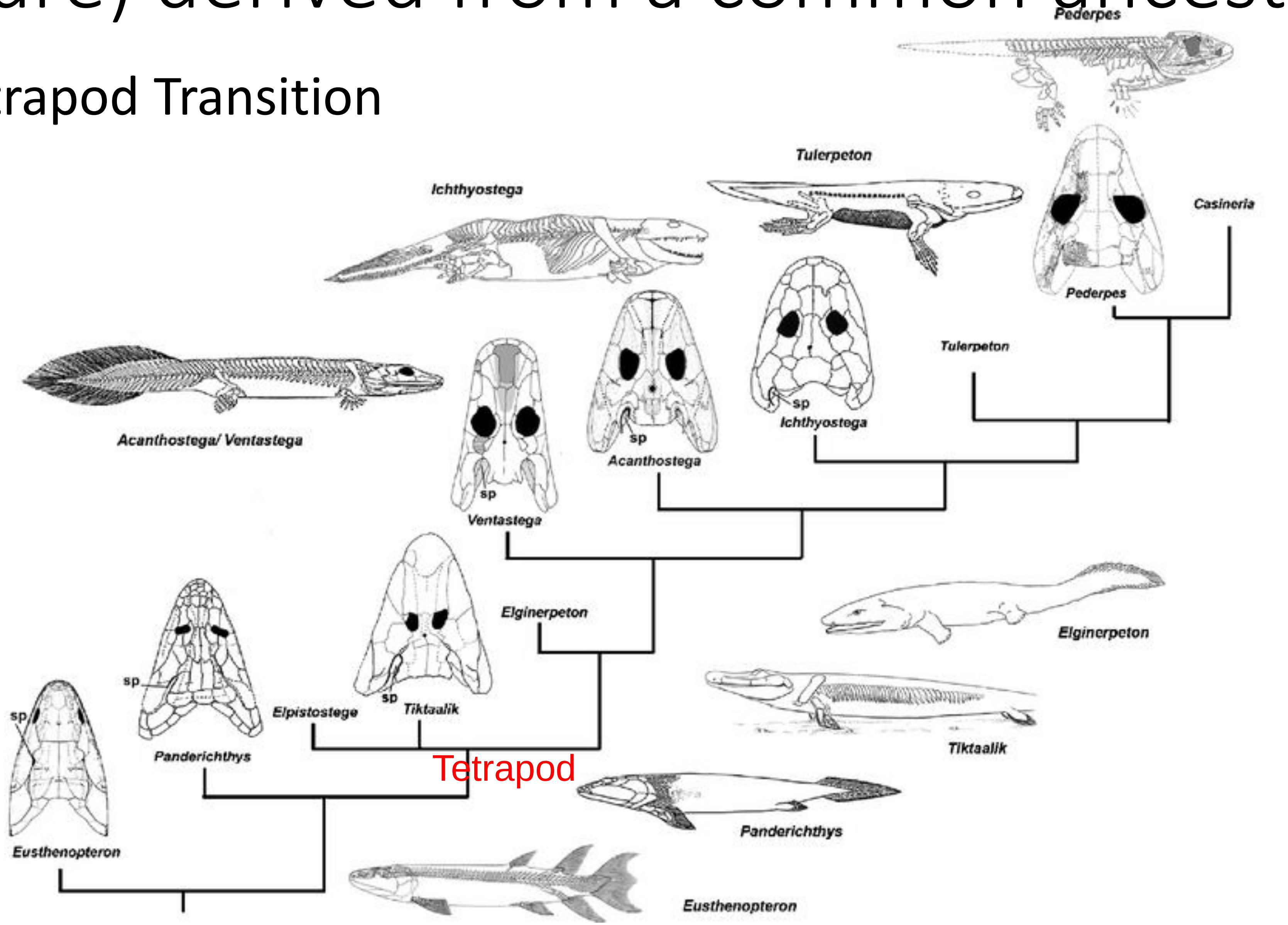


- Homologous structures are derived from a structure present in a common ancestor.

- The common ancestor of the tetrapods (mammals, birds, reptiles, etc.) had a forelimb with similar components.

Homologies: similar characteristics (position or structure) derived from a common ancestor

Fish-Tetrapod Transition



Fish

Tetrapod

Sequence

- Definition: A sequence S is an ordered set of n characters (si) representing nucleotides or amino acids. S = {s1, s2,…,sn-1 , sn}
    - DNA is composed of four **nucleotides** or **bases**: si = {A, C, G, T}
    - RNA is composed of four nucleotides: si = {A, C, G, U}(T is transcribed as U)
    - Proteins are composed of twenty **amino acids**

Biomolecular sequences

- DNA: **5’-ACGATCGACTGGTATATCGATGCT-3’**
- RNA: **5’-ACGAUCGACUGGUAUAUCGAUGCU-3’**
- Protein: **MFINRWLFSTNHKDIGTLYLLFGAWCS**

Sequence alignment in Biology

- The purpose of a sequence alignment is to line up all residues in the inputted sequence(s) for maximal level of similarity, in the sense of their functional or evolutionary relationship.

```
Snail conotoxin: GVVEHCCHRPCSNAEFKKYC-
                 | || ||||||      |||
Human insulin:   GIVEQCCHRPCNIFDLEKYCN
```

Homology of sequences

- Two genes are homologous if they originated from a common ancestral gene

- Homology: a qualitative inference; yes or no!

- Note on terminology !!!! ：“X is 30% homologous to Y”.
    - This language is incorrect because, by definition above, a pair of sequences is or is not homologous, that is, it has or does not have a common ancestor.

Identity vs Similarity

- **Identity and Similarity** are quantities that describe the relatedness of sequences
- Identity: proportion of bases or amino acid residues that are identical
  - Refers to the percentage of identities between two nucleotide or protein sequences
- Similarity - measurable (quantitative)
  - Percentage of identities + similar residues (relatively conserved throughout evolution) between two sequences

```
                   |: identical residues (12/21 percentage of identity)

Snail conotoxin: GVVEHCCHRPCSNAEFKKYC-
                 |.||.||||||.  .  |||
Human insulin:   GIVEQCCHRPCNIFDLEKYCN

                  .: similar residues (12+4)/21 percentage of similarity)
```

Definitions: two types of homology

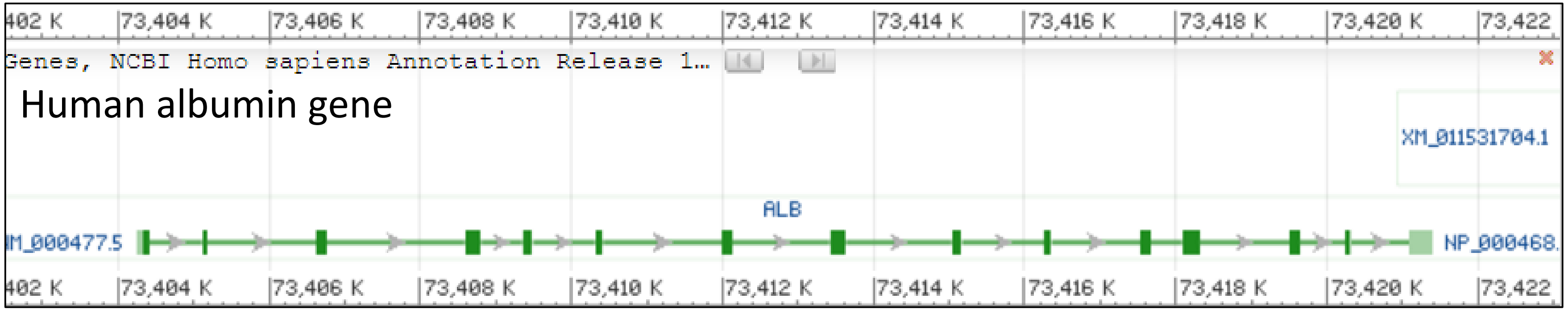
Orthologs

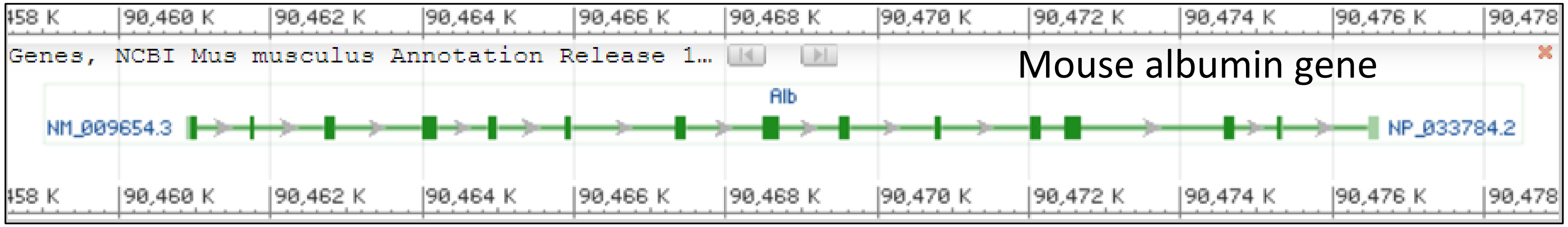
Homologous sequences in different species that arose from a common ancestral gene during **speciation**; *may or may not be responsible for a similar function.*

Paralogs

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

Homologous Genes

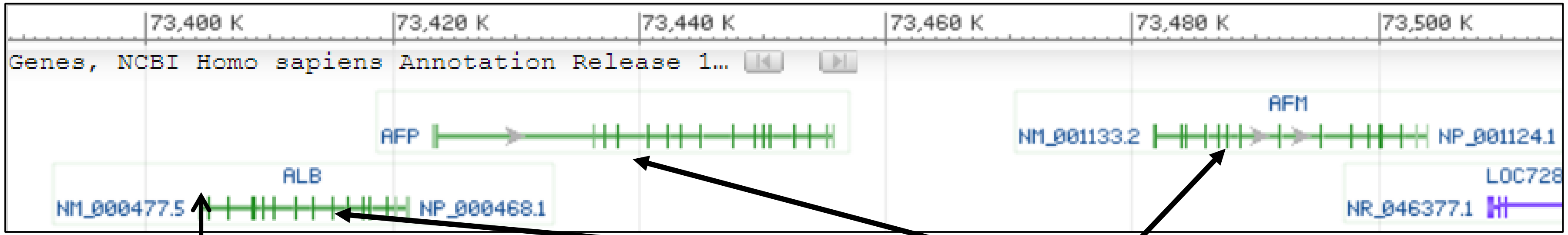

Human albumin gene


Mouse albumin gene

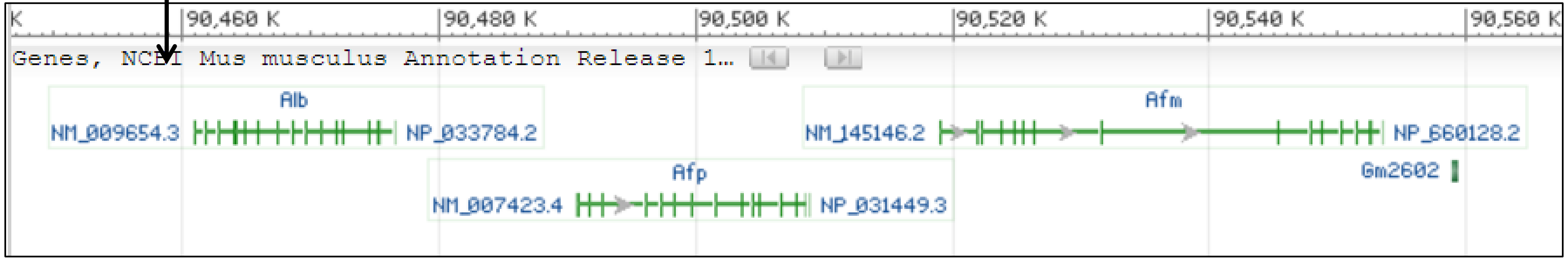
- Homologous genes constitute a gene family, for example serum albumin gene family
- Both human and mouse has albumin (ALB) gene
- It descended from a gene present in a common ancestor of the two species.

Orthologs and Paralogs

Many gene families have multiple members; Serum albumin gene family: albumin (Alb), alpha-fetoprotein (AFP) and afamin (Afm).

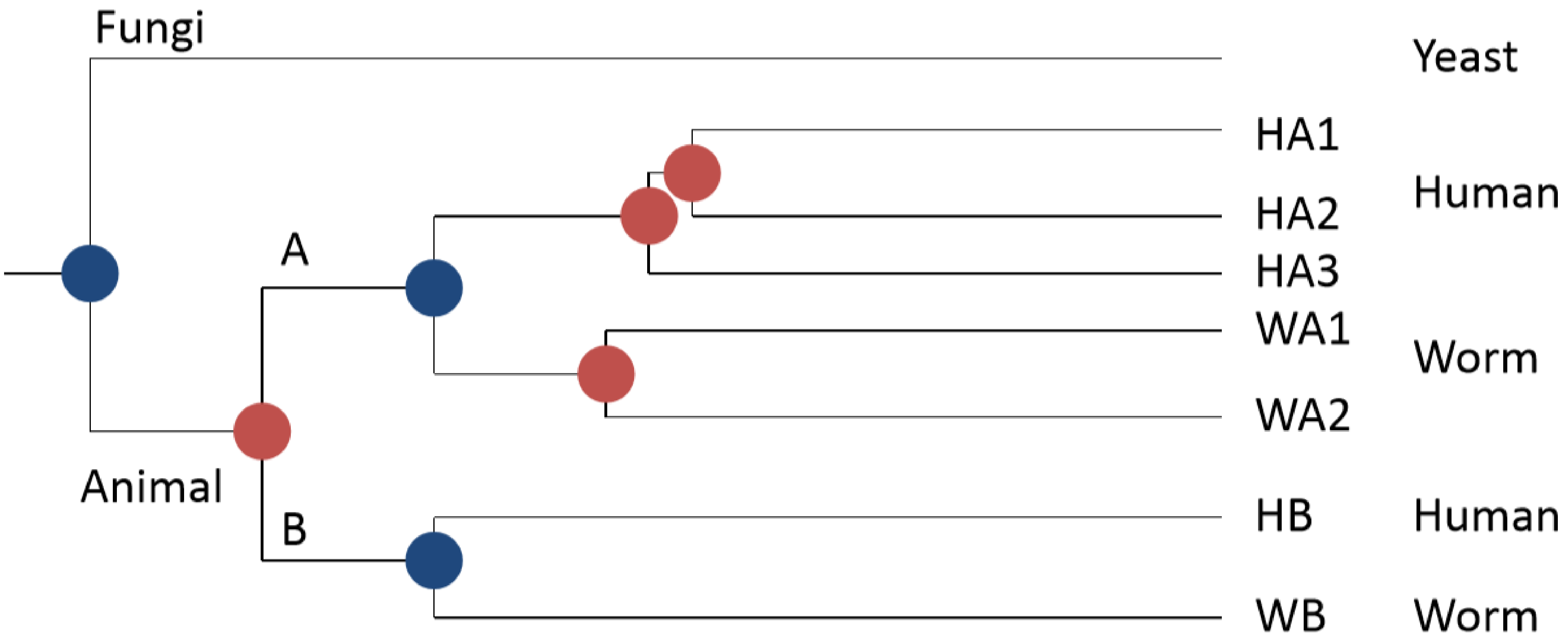


- Human ALB, AFP and AFM are **paralogous genes**.
- **Paralogs are derived by gene duplication within a species (or ancestral species).**



- ALB (human) and Alb (mouse) are **orthologous** genes.
- **Orthologs are derived by speciation events (homologs between species).**

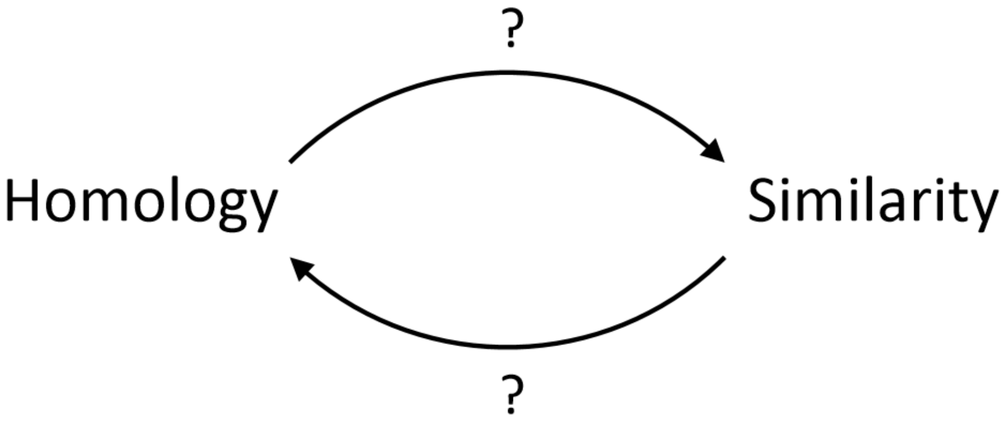
Orthologs and Paralogs



Ortholog comes with speciation Paralog comes with duplication

Revised based on Sonnhammer, E.L., and Koonin, E.V. (2002). Orthology, paralogy an
d proposed classification for paralog subtypes. TRENDS in Genetics18, 619–620.

Homology vs Similarity



- Homology is usually inferred based on the similarity of two or more sequences; however
  - Alignment with statistically significant score might not necessarily indicate they are homologous
  - On the other hand, two sequences might be homologous without sharing statistically significant identity.

- What is the golden standard?
  - Structure, functional identity and/or evolutionary history of proteins

Outlines

- Similarity & Homology

- Basic components of sequence alignment
    - Similarity or scoring Matrix
    - Gap penalties

- Dot matrix analysis

- Dynamic programming algorithm
    - Global sequence alignment: Needleman-Wunsch (NW) algorithm
    - Local sequence alignment: Smith-Waterman (SW) algorithm

Sequence alignment

A: TCAGACGATTG

L_A = 11

B: TCGGAGCTG

L_B = 9

```
TCAGACGATTG
|| || |  ||
TCGGA-GC-TG
```

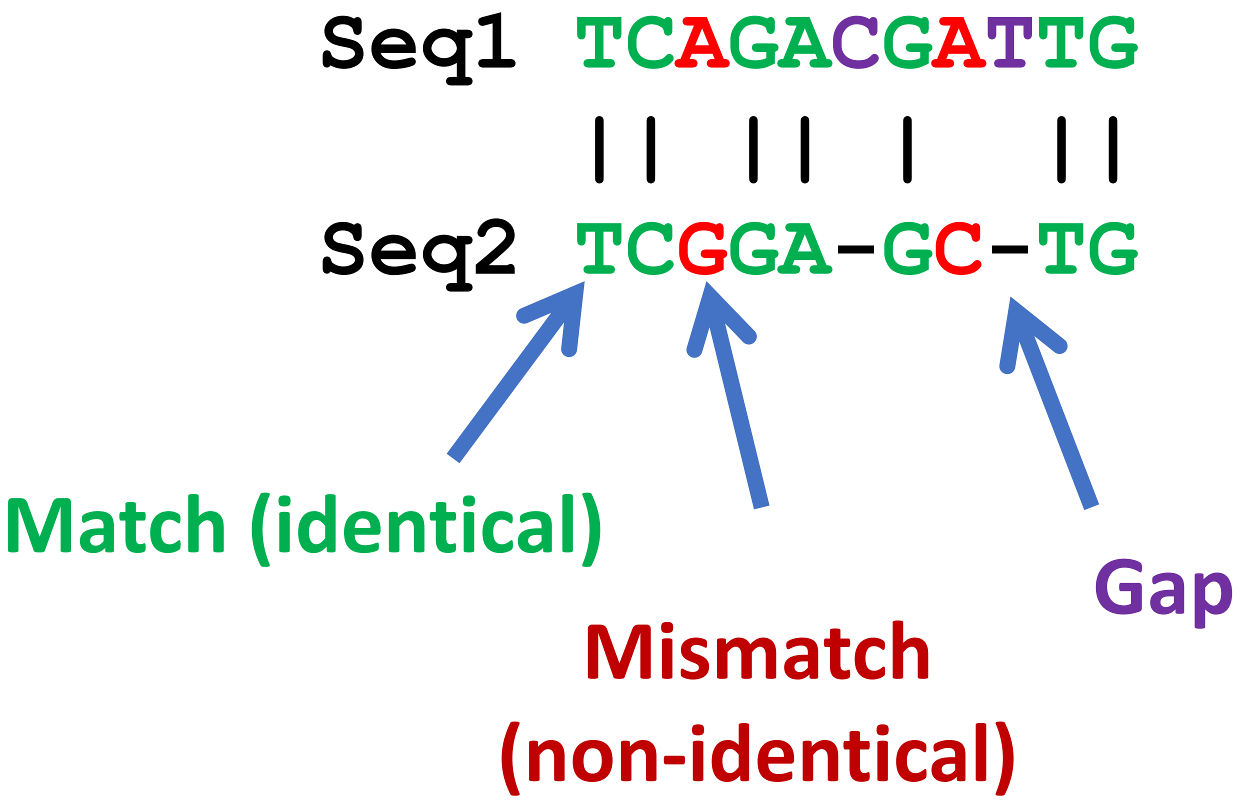
```
TCAG-ACG-ATTG
|| | | |  | |
TC-GGA-GC-T-G
```

```
TCAGACGATTG
|| ||
TCGGAGCTG--
```

```
TCAG-ACGATTG
|| | | | |
TC-GGA-GCTG-
```

How to define the similarity?

Sequence alignment



```
Seq1 TCAGACGATTG
     || || |  ||
Seq2 TCGGA-GC-TG
```

- We assign scores based on matches, mismatches, gap opening penalty, and gap extension penalty.
- These scores add up to the total raw score, which reflects degree of similarity

Match and Mismatch score matrix

- A concise way to express the character(residue) substitution costs can be achieved with a N×N matrix (N is 4 for DNA and 20 for protein)
- The substitution matrix for the simple scoring scheme:

	C	T	A	G
C	1	-1	-1	-1
T	-1	1	-1	-1
A	-1	-1	1	-1
G	-1	-1	-1	1

```
Seq1 TCAGACGATTG
     || || |  ||
Seq2 TCGGA-GC-TG
```

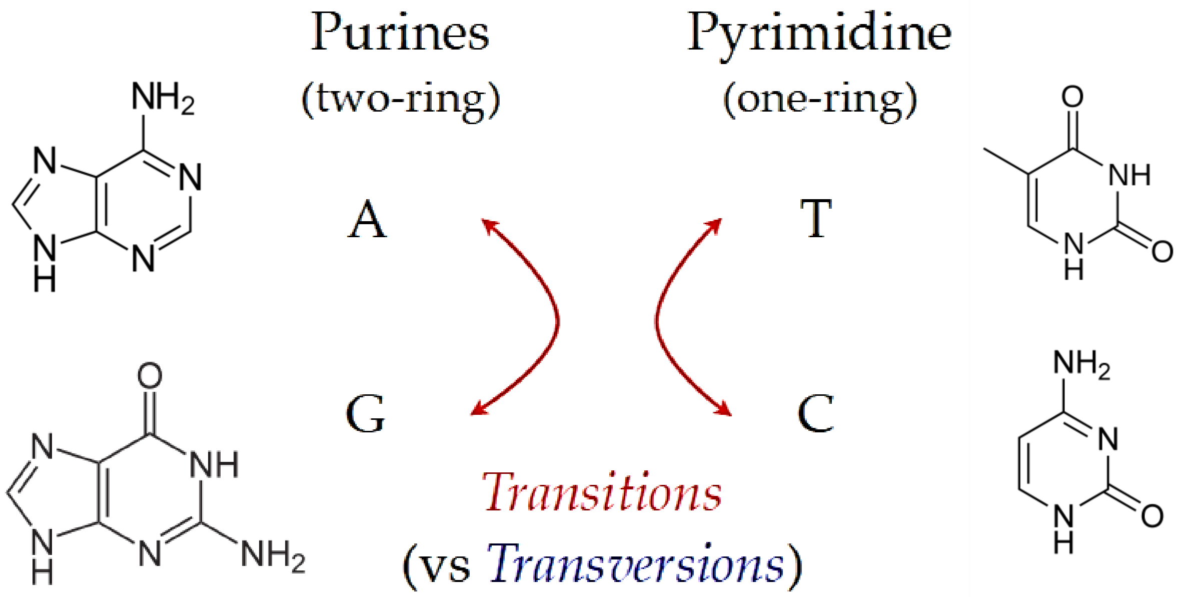
Score= 7x1 + 2x(-1) + 2x(-1) =3

where gap opening penalty is -1

Nucleotide Substitution Matrices

- A, G are purines, T, C are pyrimidines.
- Transversion are less likely to occur compared to transition

	C	T	A	G
C	2	1	-1	-1
T	1	2	-1	-1
A	-1	-1	2	1
G	-1	-1	1	2



Nucleotide Substitution Matrices

- A, G are purines, T, C are pyrimidines.
- Transversion are less likely to occur compared to transition

	C	T	A	G
C	2	1	-1	-1
T	1	2	-1	-1
A	-1	-1	2	1
G	-1	-1	1	2

```
Seq1 TCAGACGATTG
     || || |  ||
Seq2 TCGGA-GC-TG
```

Score= 2+2+1+2+2-1+2-1-1+2+2 =12

where gap opening penalty is -1

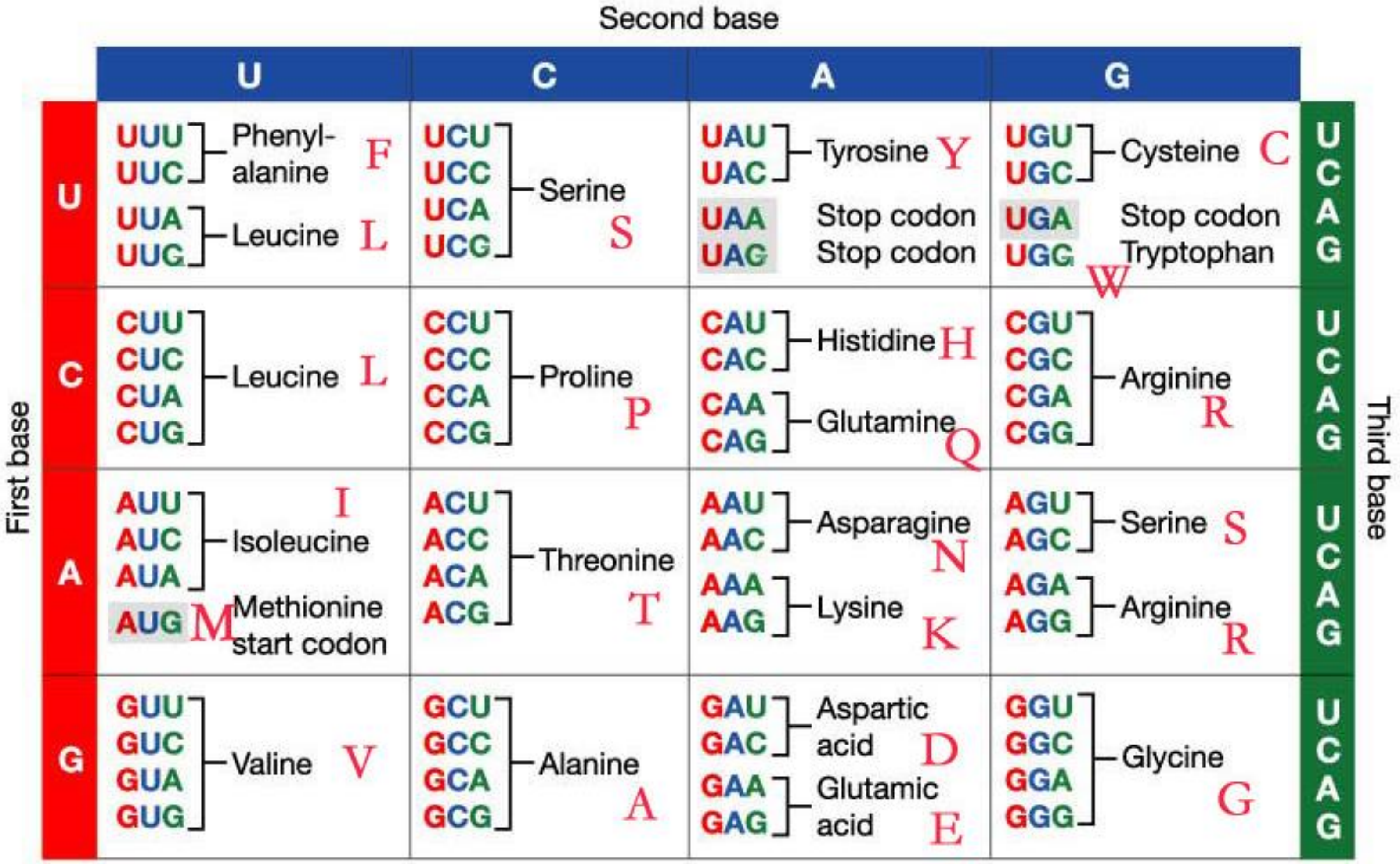
Scoring amino acid substitutions

- 20 amino acids (20x20 matrix)

- Codon usage of amino acids are different
- Amino acids share similarity based on chemical and physical properties; therefore not all substitutions are equally likely due to physical/chemical constraints
- Amino acids are NOT distributed evenly

```
Snail conotoxin: GVVEHCCHRPCSNAEFKKYC-
                 |.||.||||||.  .  |||
Human insulin:   GIVEQCCHRPCNIFDLEKYCN
```

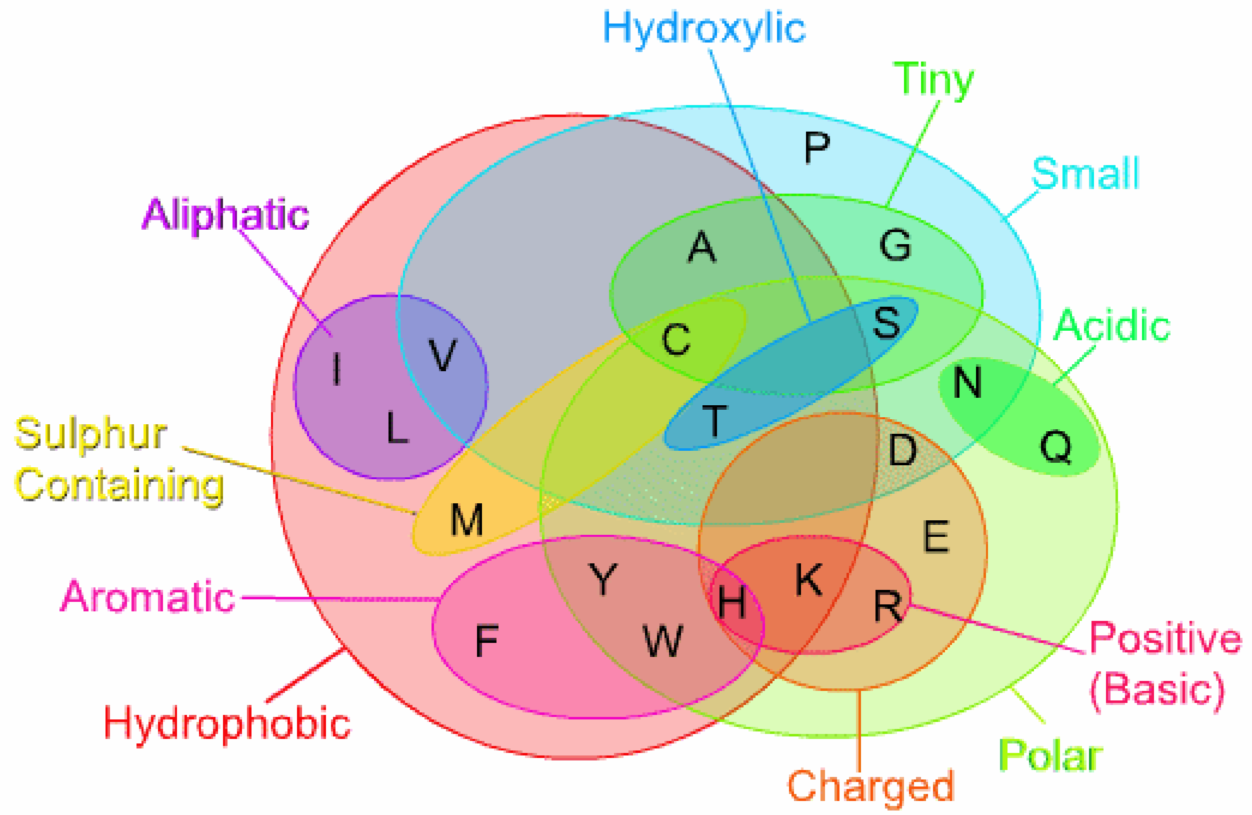
Codons per amino acid are different



	Second base				
First base	U	C	A	G	Third base
U	UUU UUC Phenyl-alanine F UUA UUG Leucine L	UCU UCC UCA UCG Serine S	UAU UAC Tyrosine Y UAA Stop codon UAG Stop codon	UGU UGC Cysteine C UGA Stop codon UGG Tryptophan W	U C A G
C	CUU CUC CUA CUG Leucine L	CCU CCC CCA CCG Proline P	CAU CAC Histidine H CAA CAG Glutamine Q	CGU CGC CGA CGG Arginine R	U C A G
A	AUU AUC AUA Isoleucine I AUG M Methionine start codon	ACU ACC ACA ACG Threonine T	AAU AAC Asparagine N AAA AAG Lysine K	AGU AGC Serine S AGA AGG Arginine R	U C A G
G	GUU GUC GUA GUG Valine V	GCU GCC GCA GCG Alanine A	GAU GAC Aspartic acid D GAA GAG Glutamic acid E	GGU GGC GGA GGG Glycine G	U C A G

The codons of some residues only differ in one base, which makes substitution of residues much easier during evolution, such as D/E, S/P/T/A, or F/L/I/V.

Amino acids share similarity based on chemical and physical properties

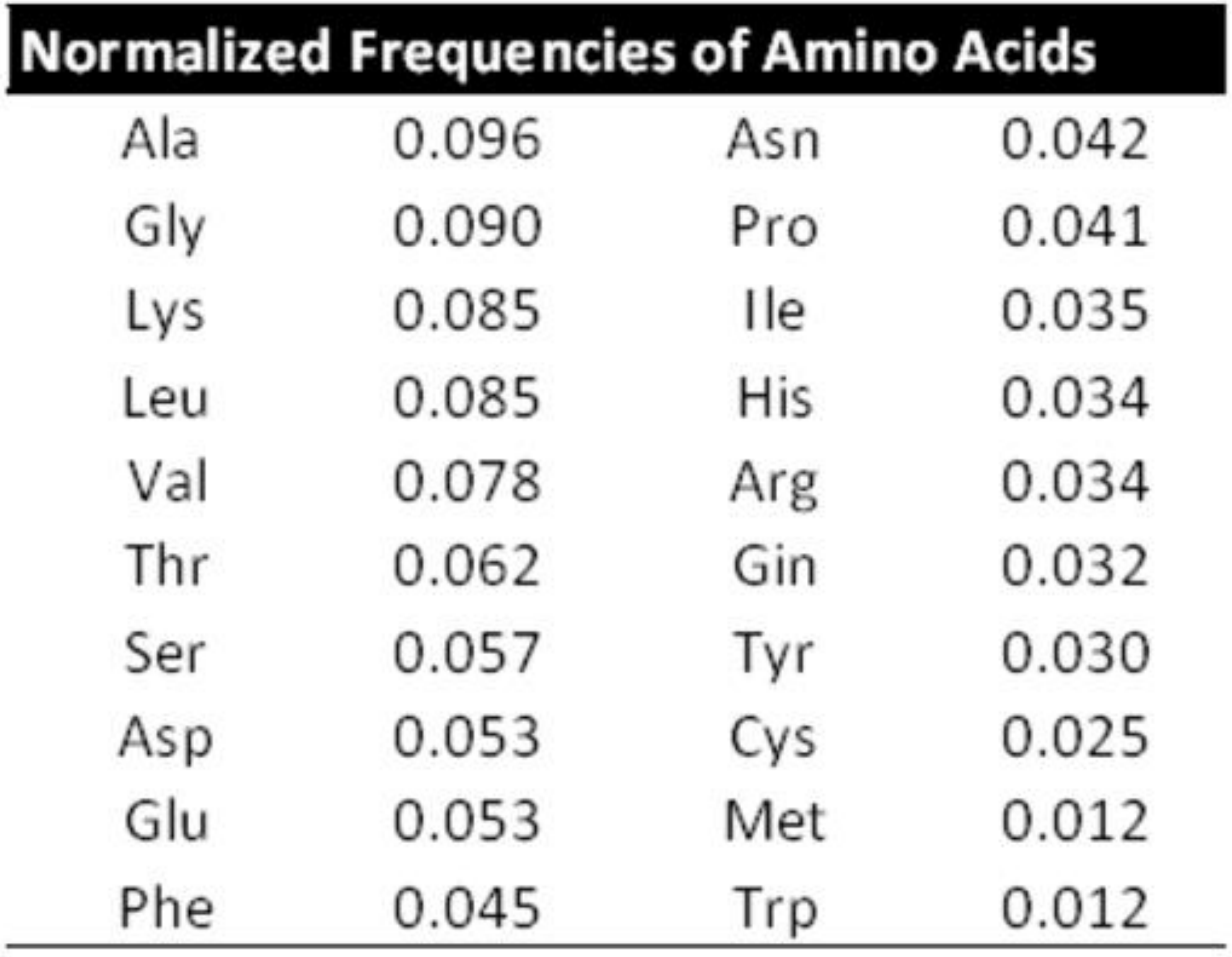


Amino Acids

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

- Residue substitution will happen more easily between amino acids with similar property: size, polarity, charge, hydrophobicity during evolution

Frequencies of amino acid are different



Normalized Frequencies of Amino Acids

Ala	0.096	Asn	0.042
Gly	0.090	Pro	0.041
Lys	0.085	Ile	0.035
Leu	0.085	His	0.034
Val	0.078	Arg	0.034
Thr	0.062	Gin	0.032
Ser	0.057	Tyr	0.030
Asp	0.053	Cys	0.025
Glu	0.053	Met	0.012
Phe	0.045	Trp	0.012

- How often a given amino acid appears in the protein (determined by an empirical analysis)

- Codon usage factors:
  - Met and Tryp have only 1 codon
  - Leu, Ser and Arg have 6 codons

  - But can not explain all

Protein Substitution Matrices to measure similarity of amino acids?

- Need--- scoring systems to model sequence change over evolutionary time

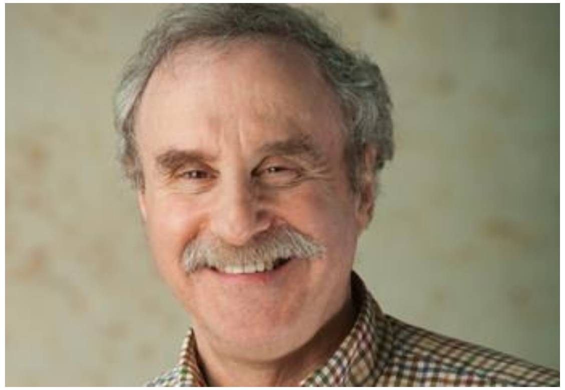
- Favor matching identical or related amino acids

- Penalize poorly matched or unrelated amino acids

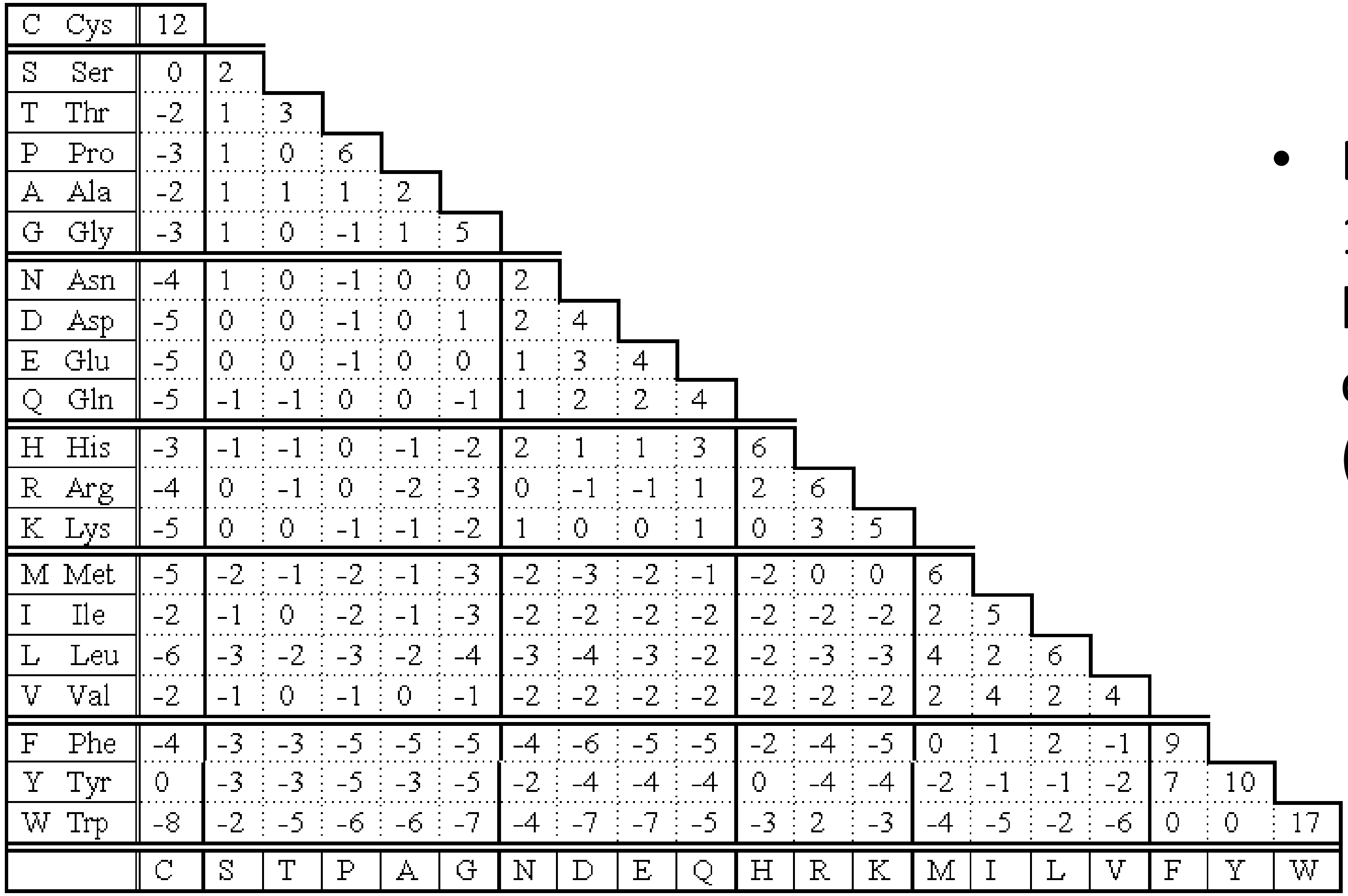
- Take into considerations the relative abundance of amino acids in proteins

Protein substitution matrices

- PAM (“Point Accepted Mutation") family, Dayhoff matrix (1978)
  - Derived from trusted alignments between closely related proteins
  - PAM250, PAM120, etc.
- BLOSUM ("BLOcks SUbstitution Matrix") family, Henikoff matrix (1992)
  - Derived from the BLOCKS database (http:blocks.fhcrc.org) ungapped multiple alignments of conserved segments (3-60 aa in length) of related proteins
  - Directly estimated from sequences with different degrees of divergence
  - BLOSUM62, BLOSUM50, etc.



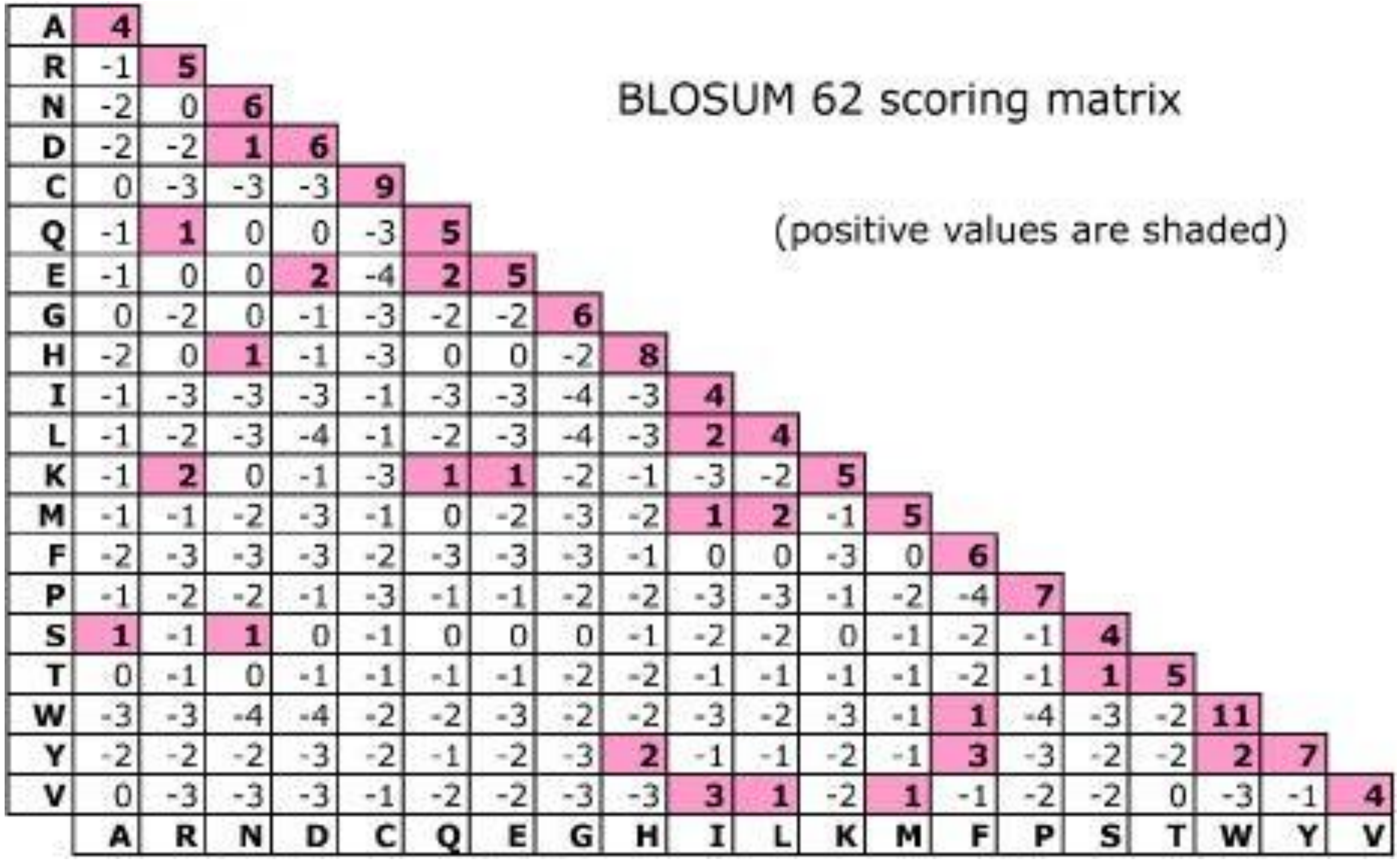
Dayhoff matrix (PAM 250)



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

- Based on the observation of 1572 accepted mutations between 34 superfamilies of closely related sequences (Dayhoff et al 1978)

BLOSUM62 substitution matrix



BLOSUM 62 scoring matrix

(positive values are shaded)

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	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-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

- Based on over 500 groups of local multiple alignments (blocks) of related protein sequences.
- The matrix number (eg BLOSUM62) represents the percentage threshold of similarity between the sequences used in the construction of the matrix
- For example, BLOSUM62 is derived from alignment of sequences that share 62% similarity, BLOSUM45 is based on 45% similarity in aligned sequences

The alignment score with substitution matrix

- Consider two alternative alignments of ANRGDFS and ANREFS with the gap opening penalty of 10:

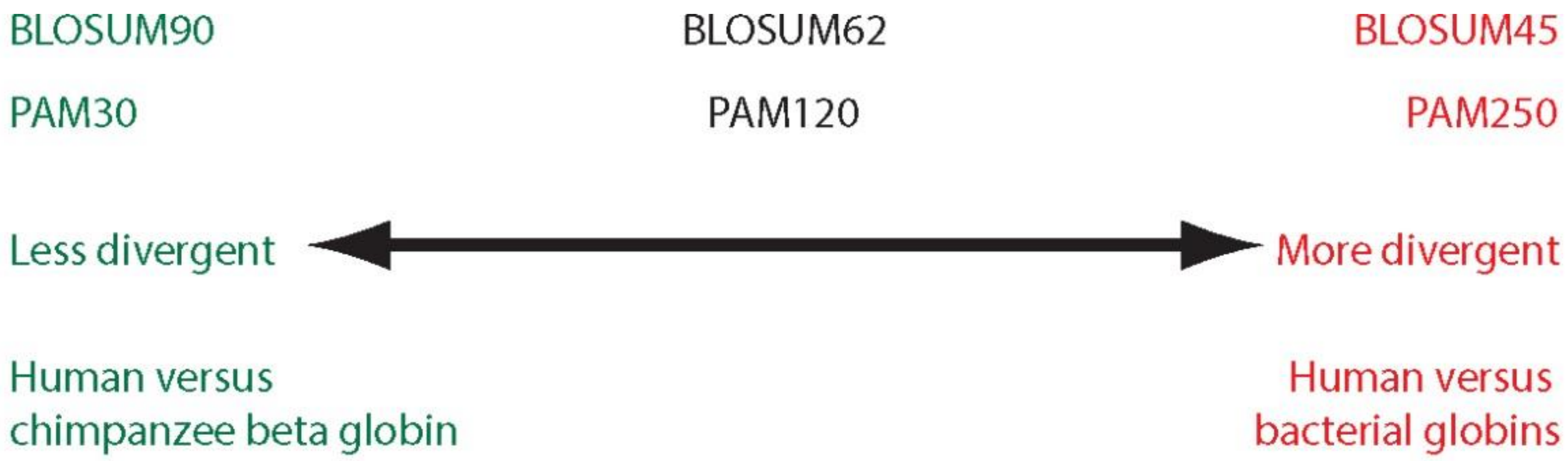
```
ANRGDFS
ANR-EFS
```

score: 4+6+5-10+2+6+4 = 17

```
ANRGDFS
ANRE-FS
```

score: 4+6+5-2-10+6+4 = 13

Relationships of PAM and BLOSUM matrices



BLOSUM90 BLOSUM62 BLOSUM45

PAM30 PAM120 PAM250

Less divergent ↔ More divergent

Human versus chimpanzee beta globin

Human versus bacterial globins

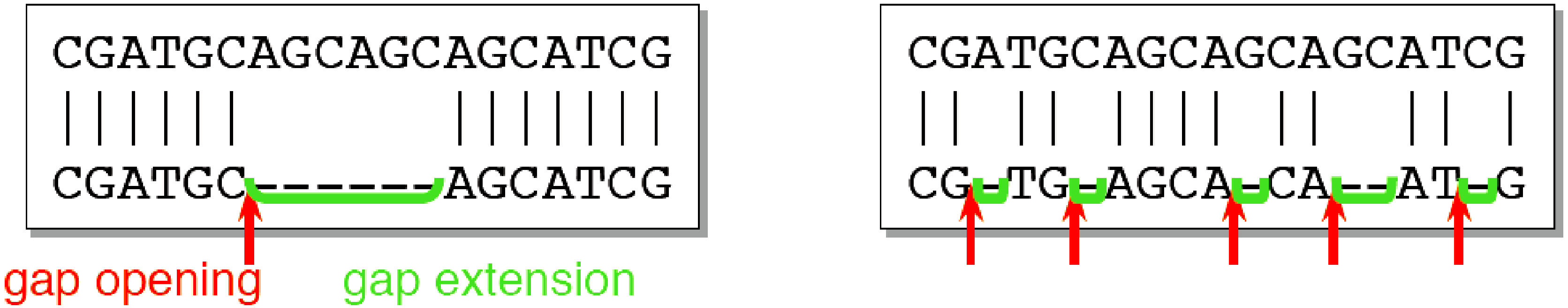
- Lower PAM matrices (eg PAM30) and higher BLOSUMs (eg BLOSUM90) are best suited for finding blocks of short local alignments of highly similar sequences.
- Higher PAM matrices (eg PAM250) and lower BLOSUMs (eg BLOSUM45 or BLOSUM30) tend to find weaker alignment blocks (less similarity) and long aligned blocks to find more distant sequences.
- No single matrix answers all the questions!

Gaps, opening and extension penalties

- Positions at which a letter is paired with a null are called gaps which correspond to an insertion or a deletion of residues

- Gap scores are typically negative.

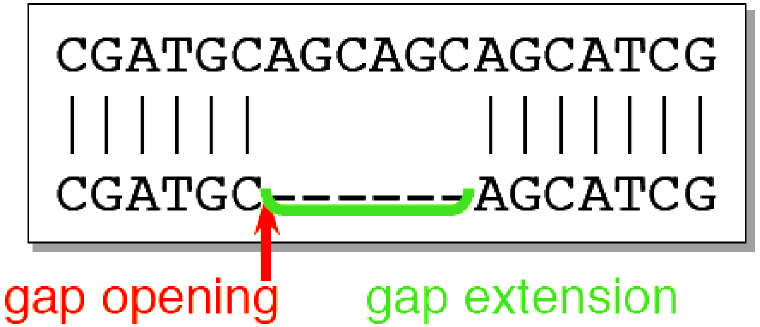
- There are separate penalties for <u>gap opening</u> and <u>gap extension</u> as the presence of a gap is more significant than the length of the gap.



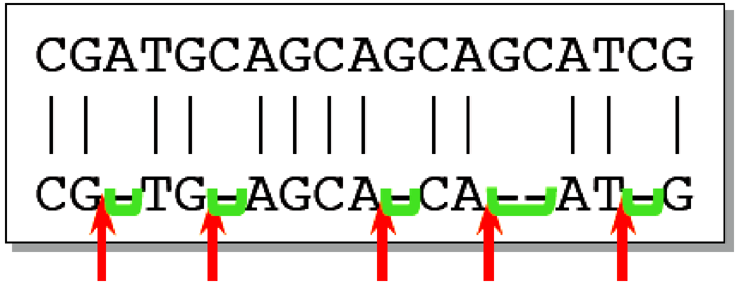
Two alignments with identical number of gaps but very different gap distribution.
We may prefer one large gap to several small ones.

Gap opening and extension penalties

- Gap opening penalty: Counted each time a gap is opened in an alignment
- Gap extension penalty: Counted for each extension of a gap in an alignment
- Match = 1, mismatch = 0, gap opening = -10. gap extension = -1



13 x 1 - 10 - 6 x 1 = -3



13 x 1 - 5 x 10 - 6 x 1 = -43

Outlines

- Similarity & Homology
- Basic components of sequence alignment
  - Similarity or scoring Matrix
  - Gap penalties
- Dot matrix analysis
- Dynamic programming algorithm
  - Global sequence alignment: Needleman-Wunsch (NW) algorithm
  - Local sequence alignment: Smith-Waterman (SW) algorithm

Principle of sequence alignment

```
Seq1 TCAGACGATTG
     || || |  ||
Seq2 TCGGA-GC-TG
```

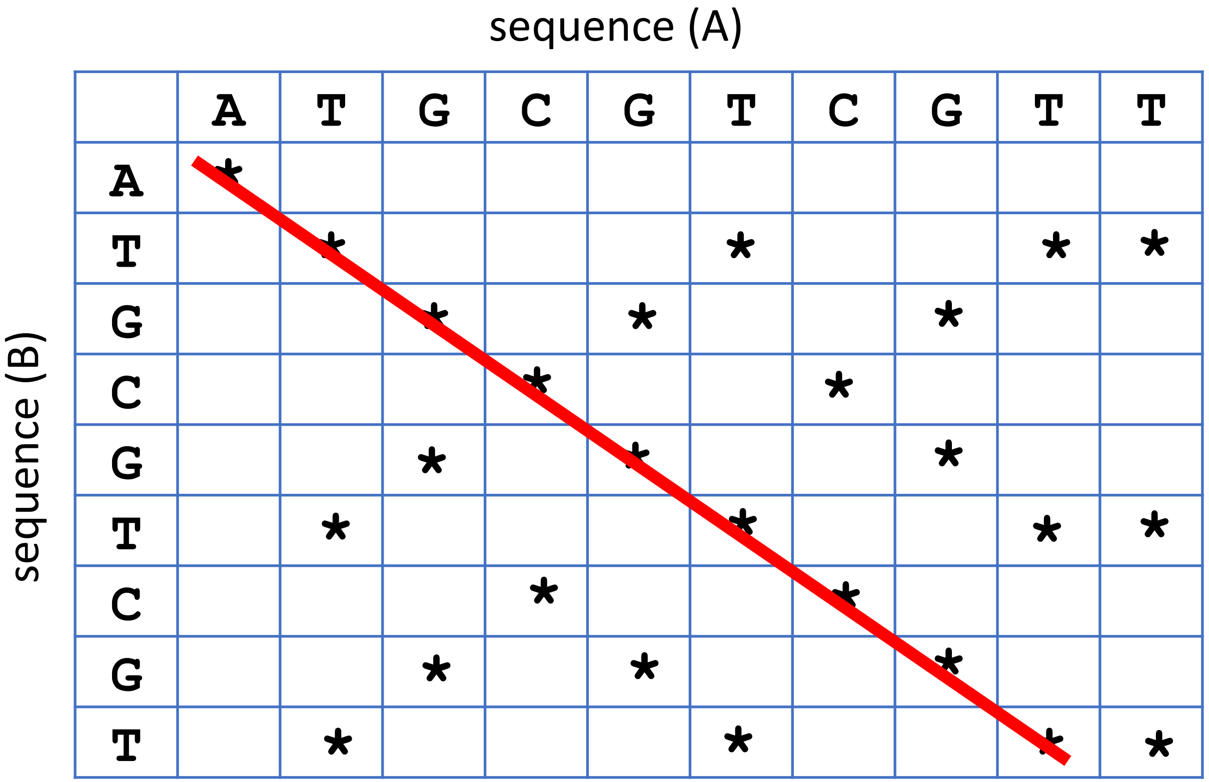
- The purpose of an alignment is to assess the degree of similarity and the possibility of homology between sequences
  - To maximize the number of matches or to minimize the number of mismatches
  - To minimize the number of gaps.

Algorithms for sequence alignment

- Dot Matrix Method (Gibbs and McIntyre 1970)
- Dynamic programming

- Common steps:
    1. Setting up a two-dimensional matrix
    2. matching or scoring the matrix
    3. Identifying the optimal alignment

Dot matrix analysis



	A	T	G	C	G	T	C	G	T	T
A	*									
T		*				*			*	*
G			*		*			*		
C				*			*			
G			*		*			*		
T		*				*			*	*
C				*			*			
G			*		*			*		
T		*				*			*	*

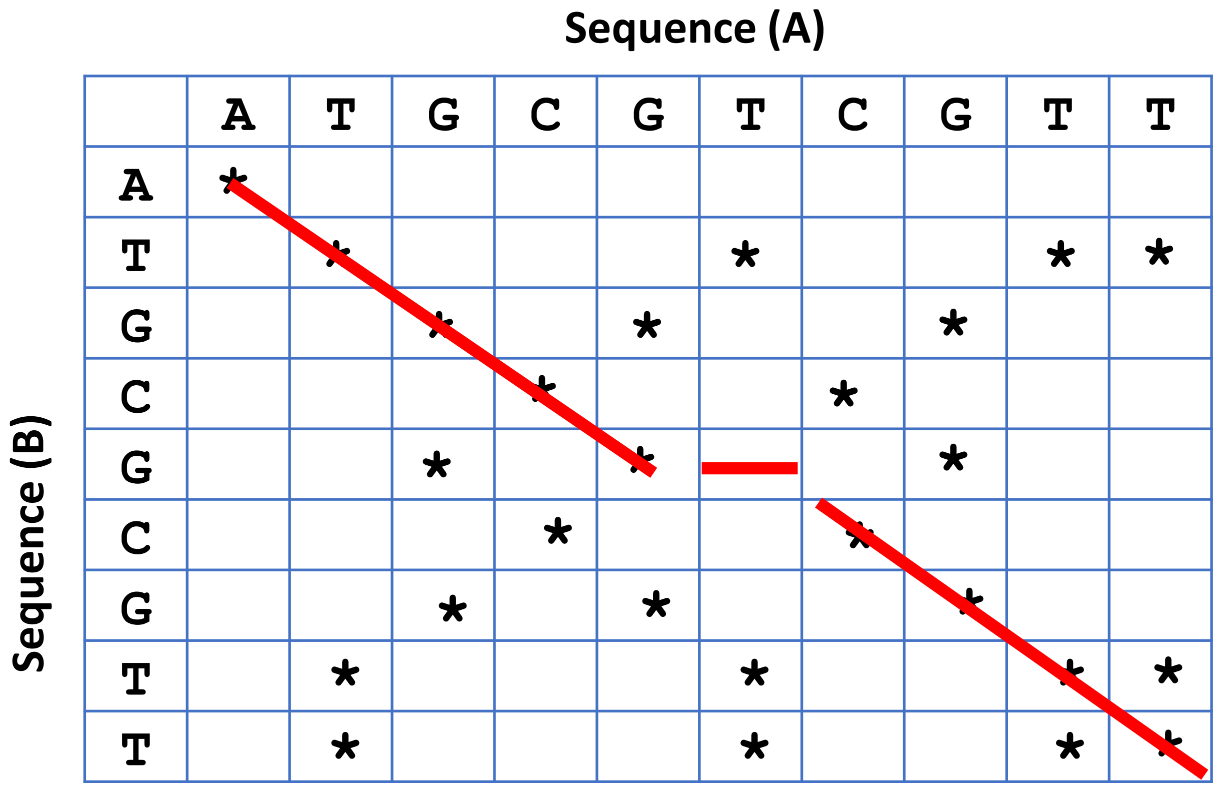
```
ATGCGTCGTT
|||||||||
ATGCGTCGT-
```

1. **Setup of a matrix**

2. **Matching**: starting from the first character in B, one moves across the page keeping in the first row and placing a dot in the column where the character in A is the same. The process is continued until all possible comparisons between A and B are made (**identical matrix**)

3. **Identify the optimal alignment**: Any region of similarity is revealed by a diagonal line of dots (Isolated lines not on diagonal represent random matches)

Dot matrix analysis



Sequence (A)

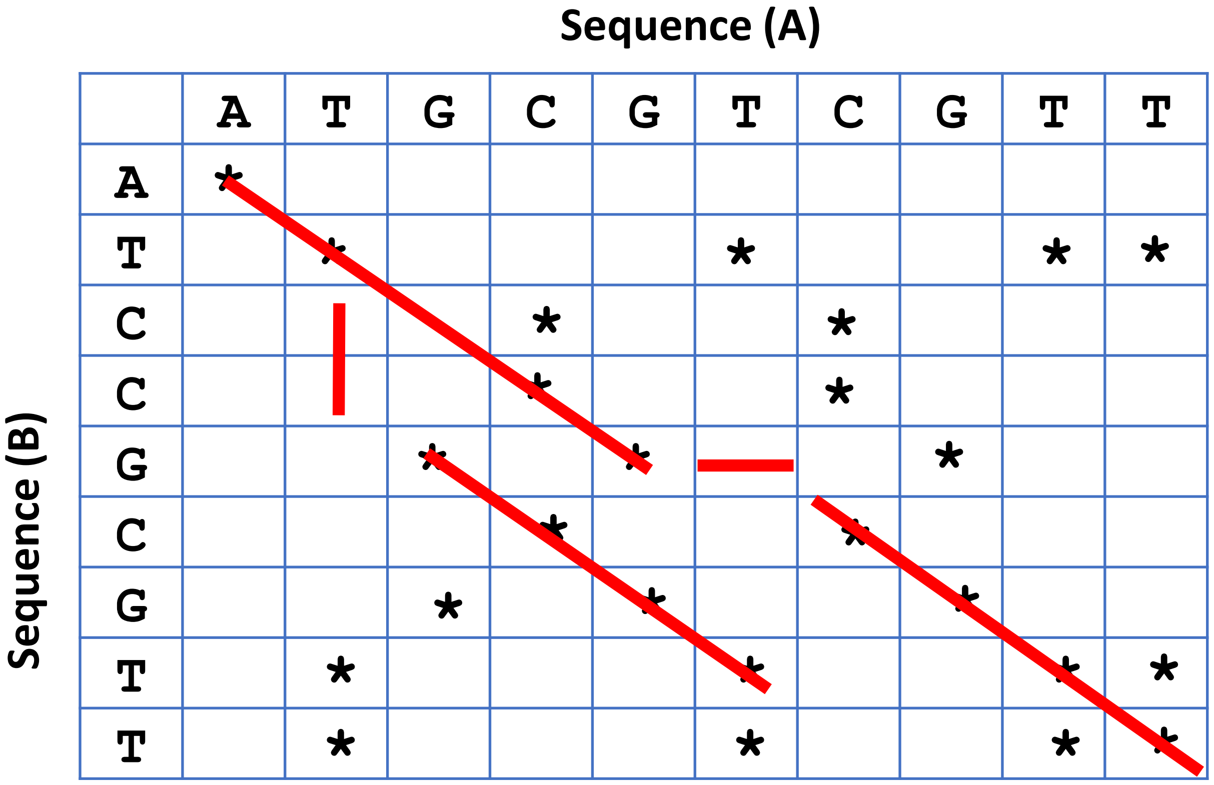
Sequence (B)

	A	T	G	C	G	T	C	G	T	T
A	*									
T		*				*			*	*
G			*		*			*		
C				*			*			
G			*		*			*		
C				*			*			
G			*		*			*		
T		*				*			*	*
T		*				*			*	*

```
ATGCGTCGTT
||||| ||||
ATGCG-CGTT
```

A gap is introduced by each vertical or horizontal skip

Dot matrix analysis

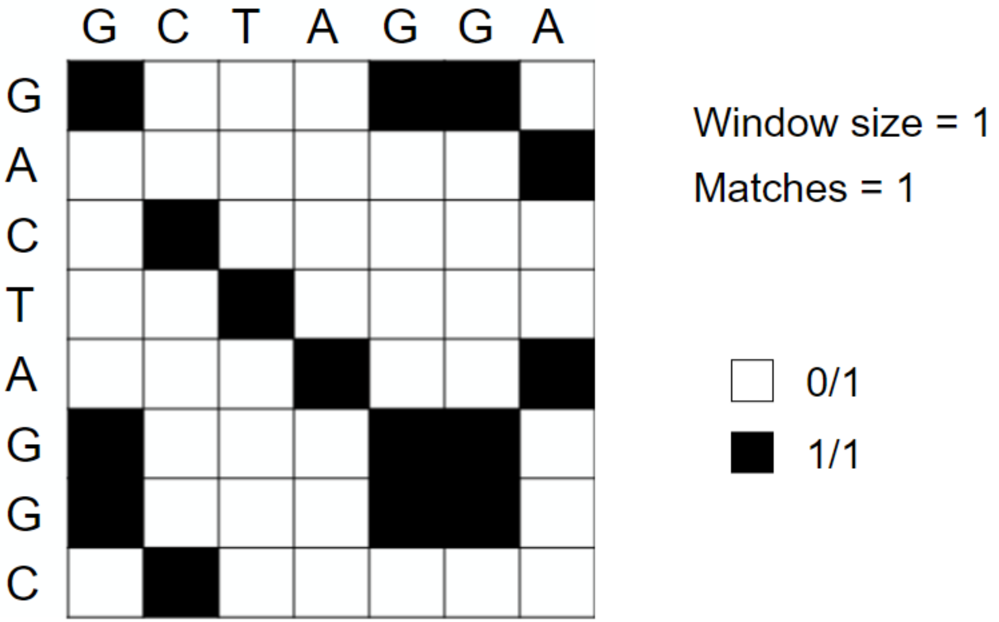


	A	T	G	C	G	T	C	G	T	T
A	*									
T		*				*			*	*
C				*			*			
C				*			*			
G			*		*			*		
C				*			*			
G			*		*			*		
T		*				*			*	*
T		*				*			*	*

```
ATGCGTCGTT        AT--GCGTCGTT
|| || ||||        ||  ||||
ATCCG-CGTT        ATCCGCGTT---
```

Dot matrix analysis using a sliding window

- Detection of matching regions can be improved by filtering out random matches by using a sliding window
- It means that instead of comparing a single sequence position, more positions is compared at the same time and dot is printed only if a certain minimal number of matches occur



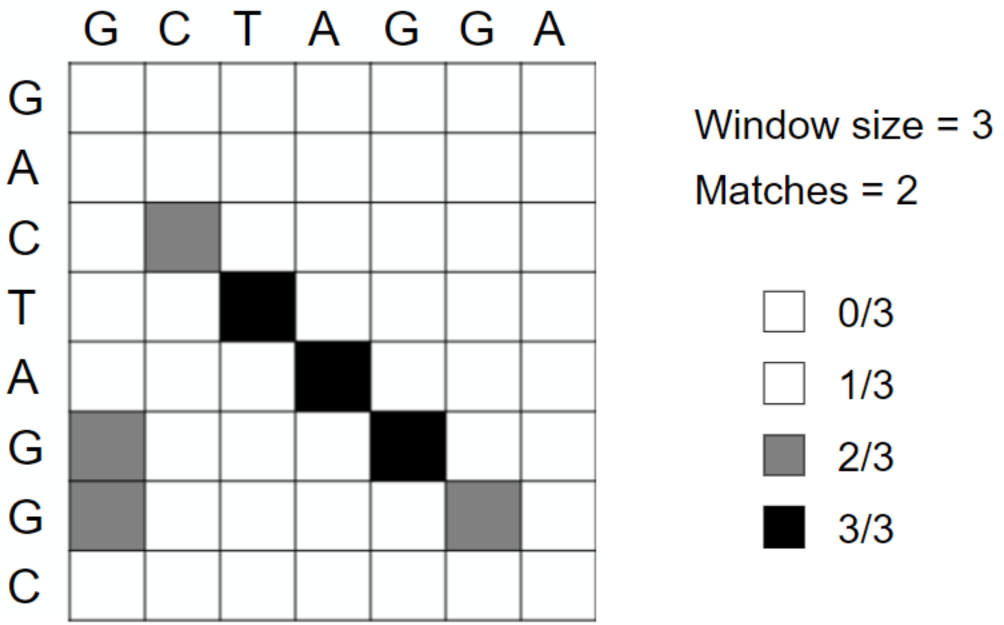
	G	C	T	A	G	G	A
G	■				■	■	
A							■
C		■					
T			■				
A				■			■
G	■				■	■	
G	■				■	■	
C		■					

Window size = 1

Matches = 1

□ 0/1

■ 1/1



	G	C	T	A	G	G	A
G							
A							
C		▨					
T			■				
A				■			
G	▨				■		
G	▨					▨	
C							

Window size = 3

Matches = 2

□ 0/3

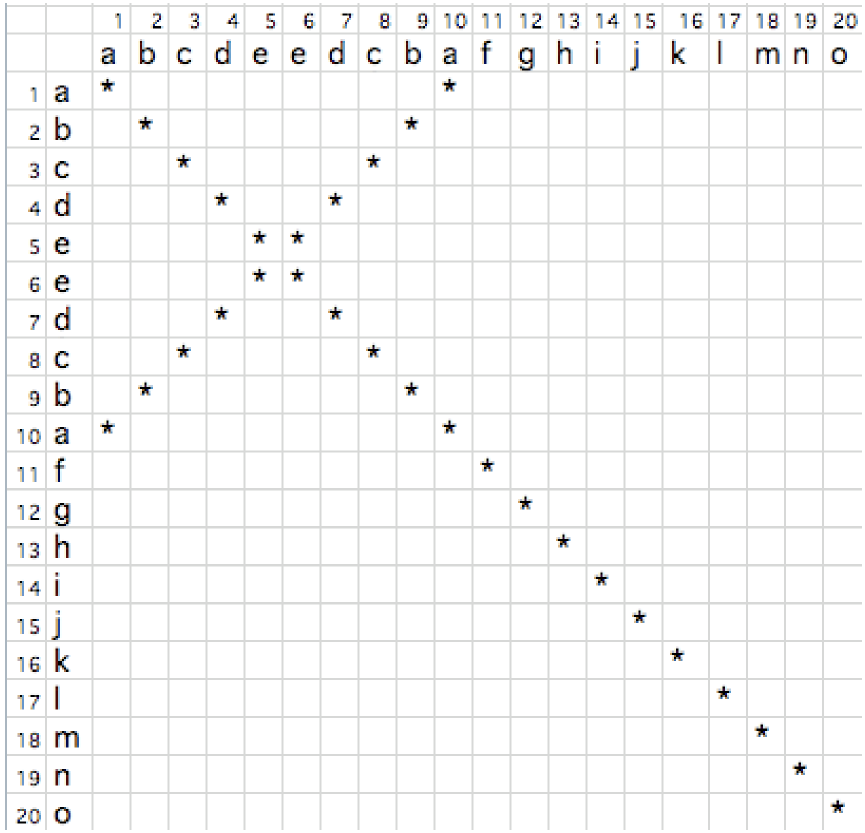
□ 1/3

▨ 2/3

■ 3/3

Dot matrix analysis to find sequence repeats

- Dot matrix analysis can be used to find **direct** and **inverted** repeats within the sequences
  - Reverse diagonals (perpendicular to diagonal) indicate inversions
  - Reverse diagonals crossing diagonals (Xs) indicate palindromes

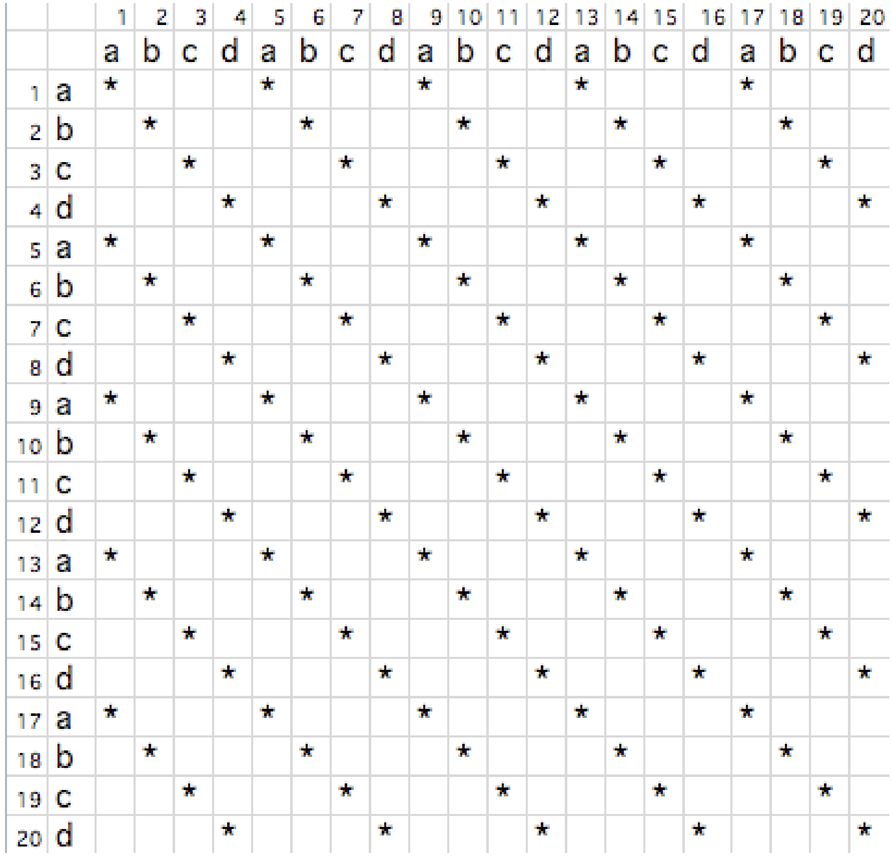


		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		a	b	c	d	e	e	d	c	b	a	f	g	h	i	j	k	l	m	n	o
1	a	*									*										
2	b		*							*											
3	c			*					*												
4	d				*			*													
5	e					*	*														
6	e					*	*														
7	d				*			*													
8	c			*					*												
9	b		*							*											
10	a	*									*										
11	f											*									
12	g												*								
13	h													*							
14	i														*						
15	j															*					
16	k																*				
17	l																	*			
18	m																		*		
19	n																			*	
20	o																				*

Dot matrix analysis to find sequence repeats

- Can use to find amino acid direct repeats within a protein by comparing a protein sequence to itself

- Repeats appear as a set of diagonal runs stacked vertically



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
1	a	*				*				*				*				*			
2	b		*				*				*				*				*		
3	c			*				*				*				*				*	
4	d				*				*				*				*				*
5	a	*				*				*				*				*			
6	b		*				*				*				*				*		
7	c			*				*				*				*				*	
8	d				*				*				*				*				*
9	a	*				*				*				*				*			
10	b		*				*				*				*				*		
11	c			*				*				*				*				*	
12	d				*				*				*				*				*
13	a	*				*				*				*				*			
14	b		*				*				*				*				*		
15	c			*				*				*				*				*	
16	d				*				*				*				*				*
17	a	*				*				*				*				*			
18	b		*				*				*				*				*		
19	c			*				*				*				*				*	
20	d				*				*				*				*				*

Advantages

- Fairly easy to Implement.
- Easy to understand visually.
- It shows all possible alignment of pairs.
- It can be used in combination of other methods.
- Readily reveals the presence of insertions/deletions and direct and inverted repeats that are more difficult to find by the other, more automated methods

Limitations

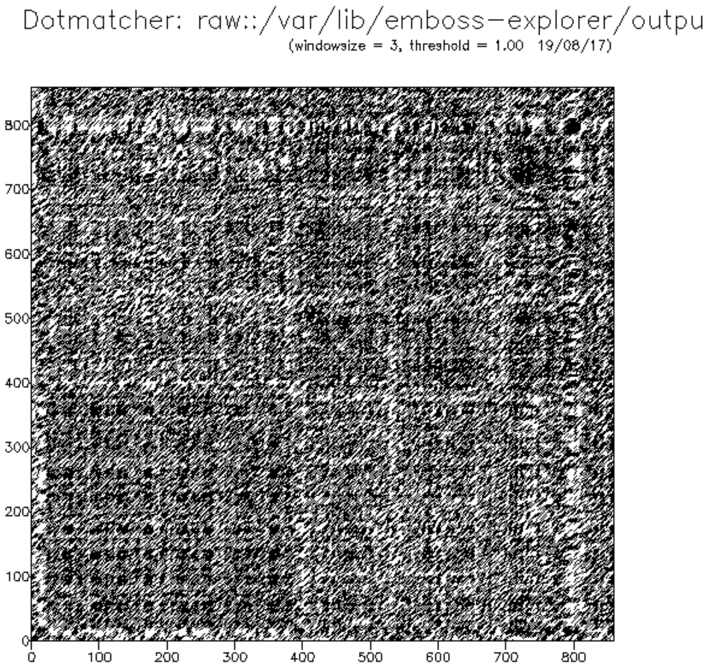
- Most dot matrix computer programs do not show an actual alignment.
- Does not return a score to indicate how ‘optimal’ a given alignment is (no statistical significance that could be tested).
- Might not be able to align two divergent sequences

Dot Plot analysis programs

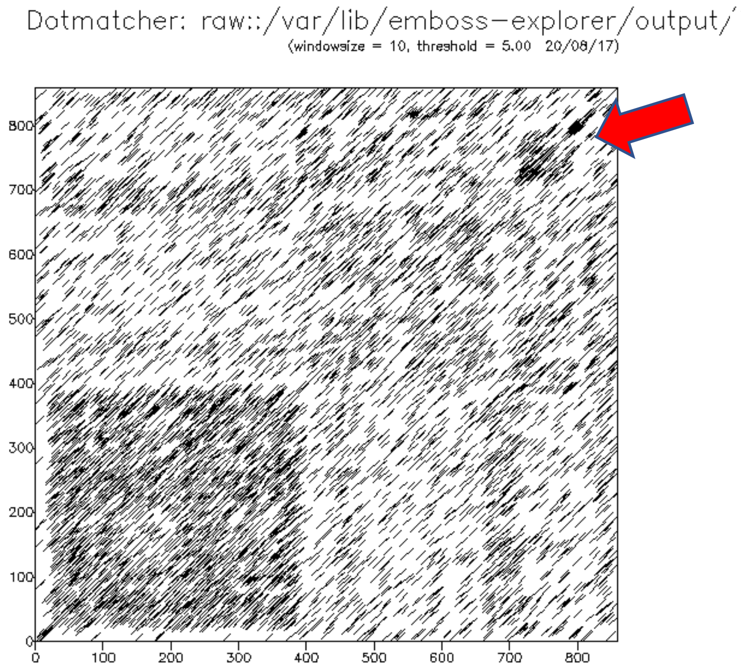
- Dotmatcher: http://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher
- Dotlet: http://myhits.isb-sib.ch/cgi-bin/dotlet
- DOTTER: http://sonnhammer.sbc.su.se/Dotter.html

Dot matrix analysis of LDL receptor (P01130.1)

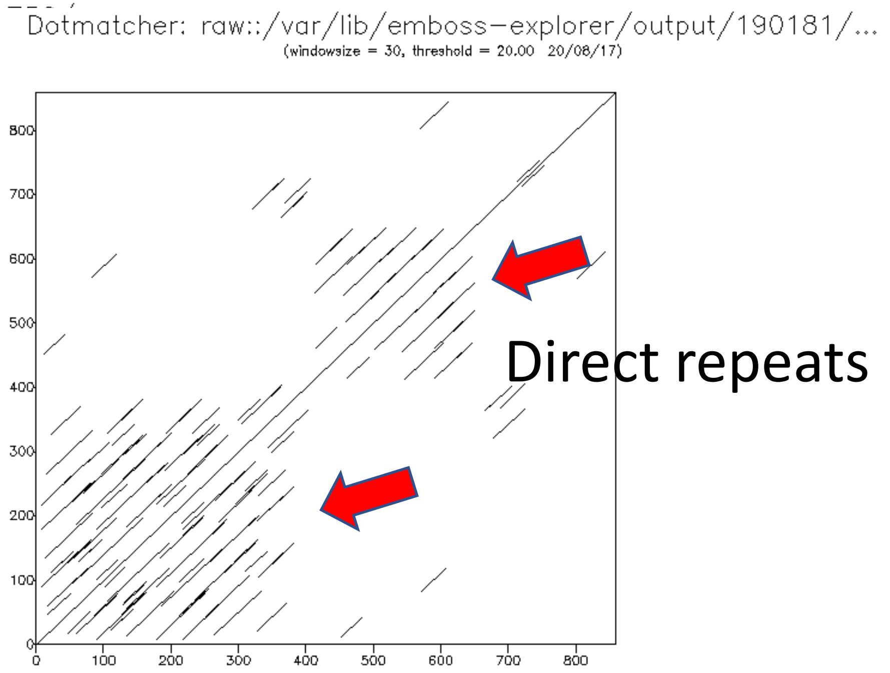
Low complexity region



Windows 1, stringency 1



Windows 10, stringency 5



Direct repeats

Windows 30, stringency 20