Tutorials (by Dominik Hörtnagel)

- Immediately after the lecture, starting today
- Duration ~ 1h
- Location: PC room L.01.125, except:
  - 7th & 14th November
  - 16th, 23rd and 30th January
  - On those days, try to bring a laptop

Chapter 3

Pairwise Sequence Alignment

Overview

1. Introduction
2. Introduction to Biological Databases
3. Pairwise Sequence Alignment
4. Database Similarity Searching
5. Multiple Sequence Alignment
6. Profiles and Hidden Markov Models
7. Protein Motifs and Domain Prediction
8. Gene Prediction
9. Promoter and Regulatory Element Prediction
10. Phylogenetics Basics
11. Phylogenetic Tree Construction Methods and Programs
12. Protein Structure Basics
13. Protein Structure Visualization, Comparison and Classification
14. Protein Secondary Structure Prediction
15. Protein Tertiary Structure Prediction
16. RNA Structure Prediction
17. Genome Mapping, Assembly and Comparison
18. Functional Genomics
19. Proteomics

Evolutionary basis of sequence variation

- Genomes undergo random changes, or *mutations*, due to the limited fidelity of DNA polymerases (typically 1 error for every 10^7 – 10^8 nucleotides copied)

- Most mutations are detrimental and therefore rapidly eliminated from the population

- Some mutations are “neutral” and therefore tolerated

- Other mutations (particularly: beneficial ones) are positively selected during evolution

Amino acid variation in protein sequences

- Variations in sequence due to evolution can be in the form of substitutions, insertions and deletions
- Residues that perform key functional and structural roles are strictly maintained (or *conserved*) by natural selection (*e.g.* catalytic residues in an enzyme)
- Residues that have less important roles show more variation

What can we learn from sequence variation?

- Sequence comparison can identify patterns of strong conservation, such as sequence motifs or domains that are required for certain functions (*e.g.* catalytic activity)
- Conversely, it can also identify areas of high variation (*e.g.* "linkers" that merely connect functionally important protein domains)
- Significant sequence similarity shared by a group of proteins suggests that these proteins belong to the same family and have a similar structure as well as a similar function

What can we learn from sequence variation?

- Traces of evolution in sequences allow identification of common ancestry, both of genes/proteins and the species they occur in
- In other words: if similarity is high, this suggests that the sequences must have derived from a common evolutionary origin (as it is very unlikely that extensive sequence similarity is acquired merely by chance)

Example of a sequence comparison (pairwise alignment)

```
H.sapiens   MPKSKELVSSSSSGSDSDSEVDKKLKRKKQVAPEKPVKKQKTG------ETSRALSSSKQ
C.elegans   M----------SSSSSSEDELEKKVTKEQKKKETKSKKRQSEAVEEEKQEVKKAKNEEEV
            *           **.*.*:.*::**:.::::    *. *:*. .      *..:* ...:

H.sapiens   SSSSRDDN---MFQIGKMRYVSVRDFKGKVLIDIREYWMDPEGE-MKPGRKGISLNPEQW
C.elegans   SGRLKDSDGNEMFEIGNLRYATVSKFKGKEYVNIREYYIDRDSQKMMPSRKGISLSKAQW
            *.  :*.:   **:**::**.:* .****  ::******:* :.: * *.******.  **

H.sapiens   SQLKEQISDIDDAVRKL
C.elegans   ANLKDLIPEID---KKF
            ::**: *.:**   :*:
```

- Identical residues (*)
- Non-identical residues: *substitutions*
- Residues with no counterpart in the other sequence: *insertions* or *deletions* (-)

Sequence homology *vs* identity and similarity

Homology ≡ having a common evolutionary origin

- Sequences can be either homologous or non-homologous

Sequence identity ≡ percentage of aligned residues that are exactly the same in both sequences

Sequence similarity ≡ percentage of aligned residues that are similar in physiochemical properties such as size, charge and hydrophobicity

- Sequence identity and similarity are quantitative (percentages)
- High sequence identity and/or similarity suggest homology

Inferring homology from sequence identity/similarity: an important caveat

The shorter the sequence the higher the chance that high identity/similarity is attributable to random chance:

- Shorter sequences require higher cutoffs for inferring homologous relationships

Sequence homology *vs* sequence identity/similarity

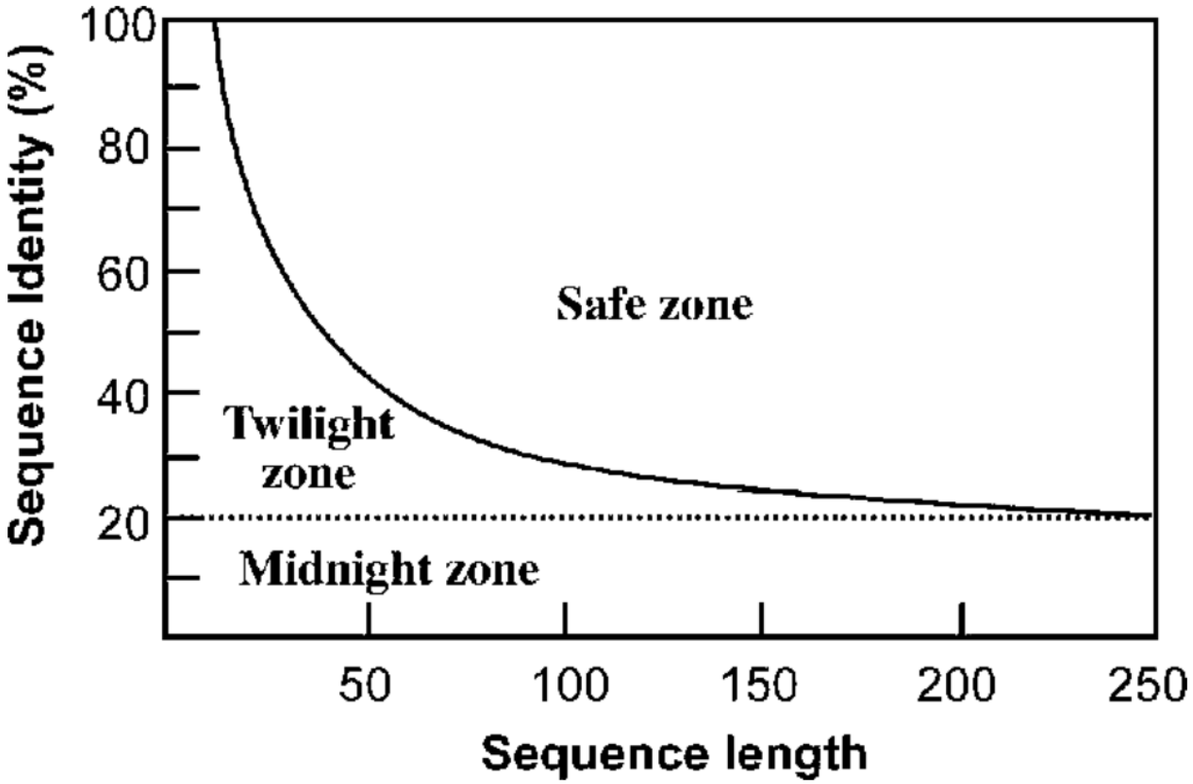


Figure 3.1: The three zones of protein sequence alignments. Two protein sequences can be regarded as homologous if the percentage sequence identity falls in the safe zone. Sequence identity values below the zone boundary, but above 20%, are considered to be in the twilight zone, where homologous relationships are less certain. The region below 20% is the midnight zone, where homologous relationships cannot be reliably determined. (*Source:* Modified from Rost 1999).

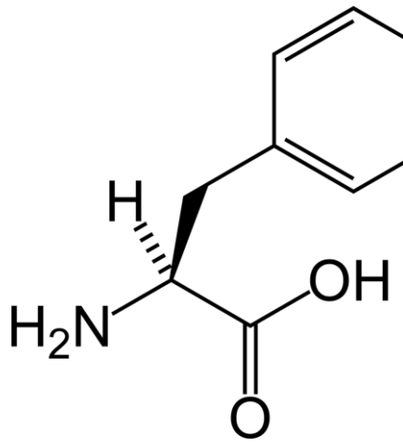
Sequence similarity *vs* sequence identity

For nucleotides sequence similarity and sequence identity are synonymous

For proteins sequence similarity and sequence identity are very different:

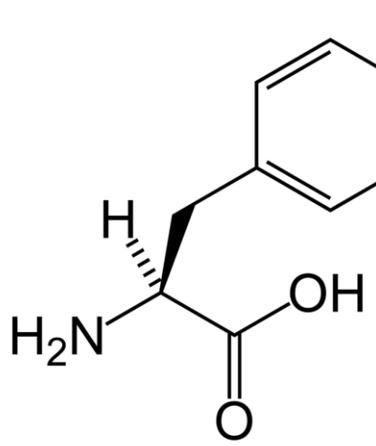
- Sequence identity ≡ percentage of exact matches between two aligned sequences
- Sequence similarity ≡ percentage of aligned residues that have similar physicochemical characteristics and can often be substituted for each other without affecting the structure and functionality of the protein much (*e.g.* Phe, Tyr, Trp)

Amino acids with aromatic rings in the side chain



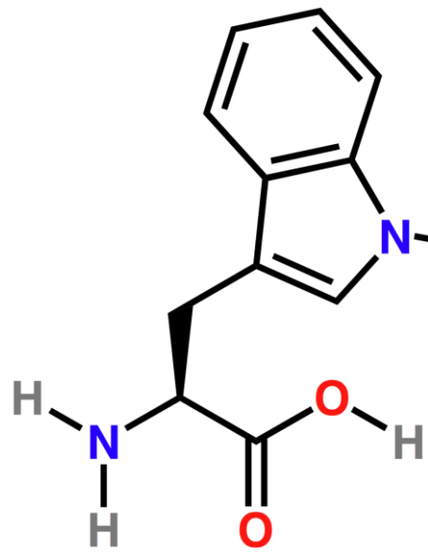
F

Phe



Y

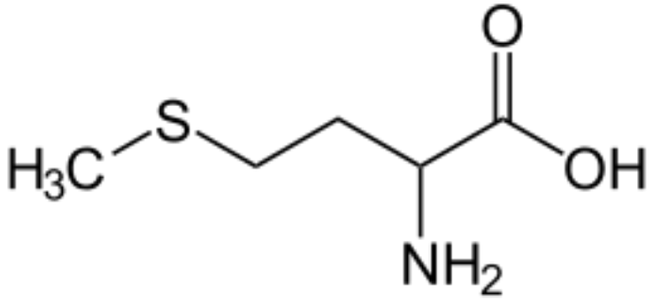
Tyr



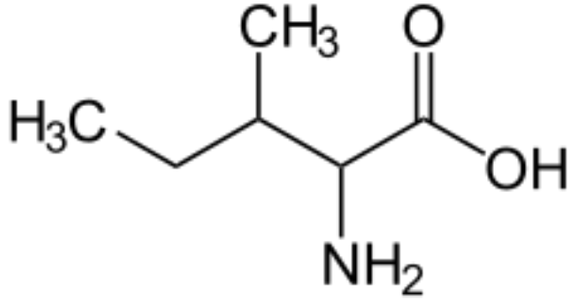
W

Trp

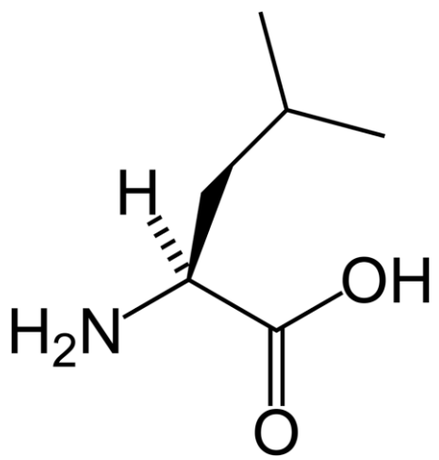
Amino acids with large hydrophobic side chains



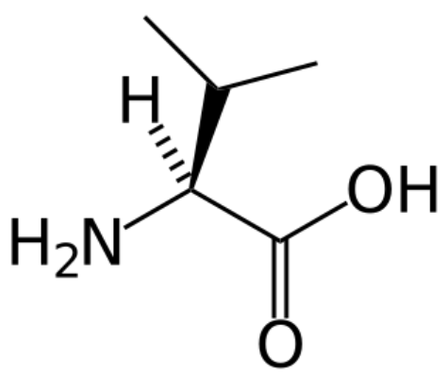
M
Met



I
Ile

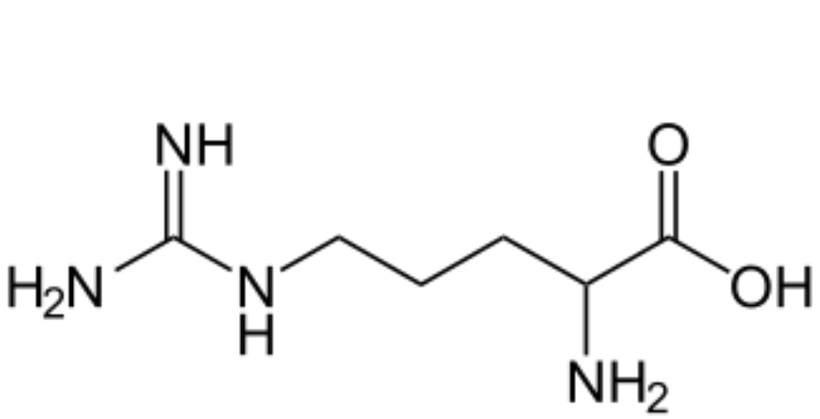


L
Leu

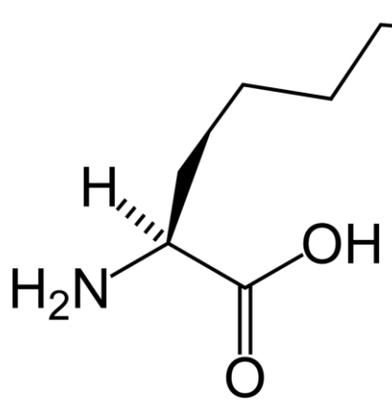


V
Val

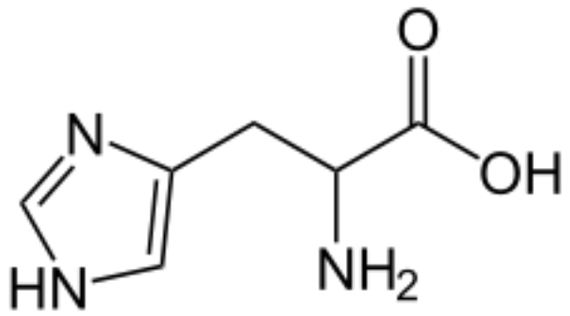
Amino acids with basic side chains (charged +)



R
Arg

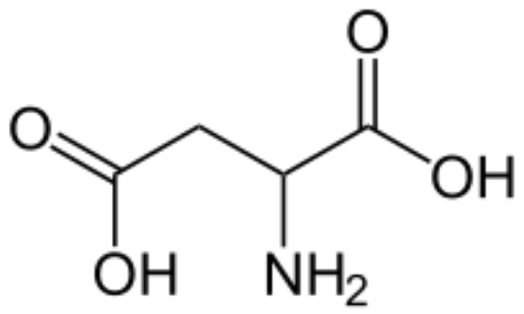


K
Lys



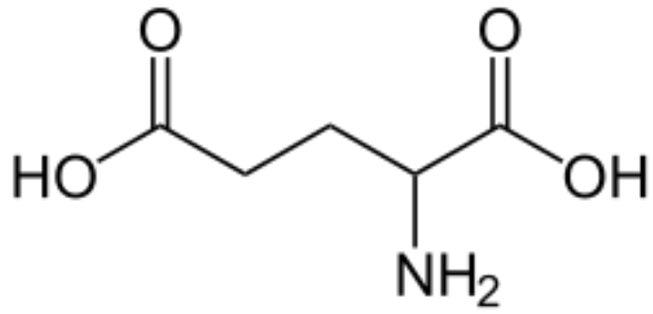
H
His

Amino acids with acidic side chains (charged -)



D

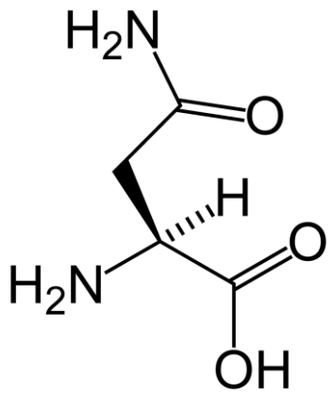
Asp



E

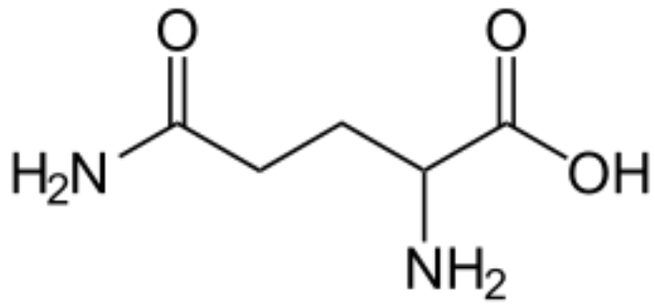
Glu

Amino acids with large polar side chains



N

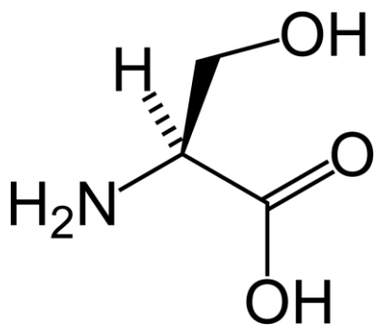
Asn



Q

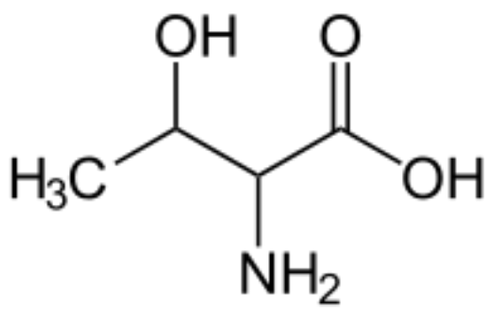
Gln

Amino acids with small polar side chains



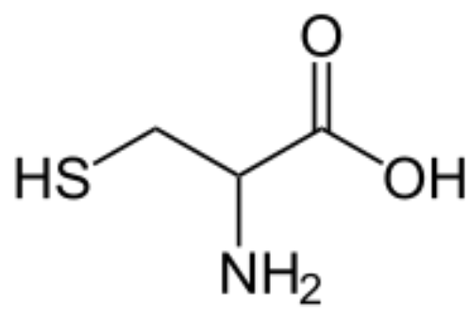
S

Ser



T

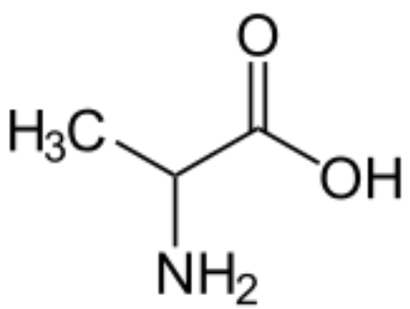
Thr



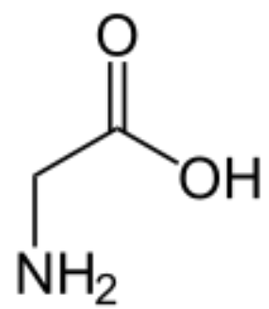
C

Cys

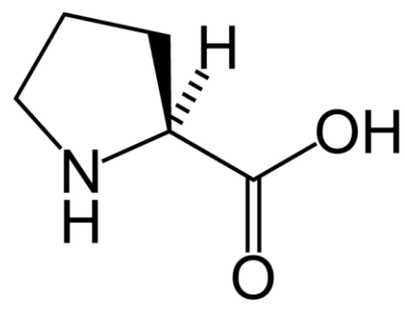
Amino acids with small non-polar side chains



A
Ala



G
Gly



P
Pro

Calculation of sequence similarity/identity

- Percentage sequence similarity, ***S***:

$$S = [(L_{\mathrm{s}} \times 2)/(L_{\mathrm{a}} + L_{\mathrm{b}})] \times 100$$

L_s : number of aligned residues <u>with similar characteristics</u>

L_a, L_b : total lengths of each individual sequence

- Percent sequence identity, ***I***:

$$I = [(L_{\mathrm{i}} \times 2)/(L_{\mathrm{a}} + L_{\mathrm{b}})] \times 100$$

L_i : number of aligned <u>identical residues</u>

Alternative formulas for sequence similarity/identity

- Alternatively, percentage of identical/similar residues over the full length <u>of the smaller sequence</u> can be calculated:

$$\mathrm{I(S)\%} = L_{\mathrm{i(s)}} / L_{\mathrm{a}} \,\%$$

L_a : length of the shorter of the two sequences

Methods for pairwise sequence alignment

Goal: look for the alignment of the two sequences with maximum correspondence among residues

- One sequence needs to be shifted relative to the other, so as to find the position with the largest number of matches
- Two different strategies:
  - global alignment
  - local alignment

Global sequence alignment

- The two sequences are assumed to be similar over their entire length
- Alignment is carried out across the entire length
- Applicable for closely related sequences of roughly the same length

Local sequence alignment

- Finds local regions with the highest level of similarity between two sequences
- Aligns these regions without regard for the alignment of the rest of the sequence
- Sequences can be of different length
- Useful for aligning more divergent (more distantly related) sequences with the goal of searching for conserved patterns in DNA or protein sequences
- Useful for identifying similar *motifs* or *domains*

Example of global and local sequence alignments

```
seq1   EARDF-NQYYSSIKRSGSIQ
          . :  .:::::::::. . .
seq2   LPKLFIDQYYSSIKRTMG-H
```

Figure 3.2: An example of pairwise sequence comparison showing the distinction between global and local alignment. The global alignment (*top*) includes all residues of both sequences. The region with the highest similarity is highlighted in a box. The local alignment only includes portions of the two sequences that have the highest regional similarity. In the line between the two sequences, ":" indicates identical residue matches and "." indicates similar residue matches.

global sequence alignment

```
seq1   NQYYSSIKRS
       .:::::::::.
seq2   DQYYSSIKRT
```

local sequence alignment

The dot matrix method

- Also known as "dot plot method"
- Graphical way of comparing two sequences in a two-dimensional matrix that contains dots
- The two sequences are written on the horizontal and vertical axes of the matrix
- Each residue of one sequence is scanned for identity with all residues of the other sequence
- A dot is placed within the graph for each match
- Dot matrix alignment at EBI:

https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_dotmatcher

Interpreting a dot matrix

- When the two sequences have substantial regions with sequence identity <u>many dots line up to form contiguous diagonal lines</u>
- Interruptions indicate insertions or deletions
- Parallel diagonal lines reflect repetitive regions within the sequences

Example of a dot plot

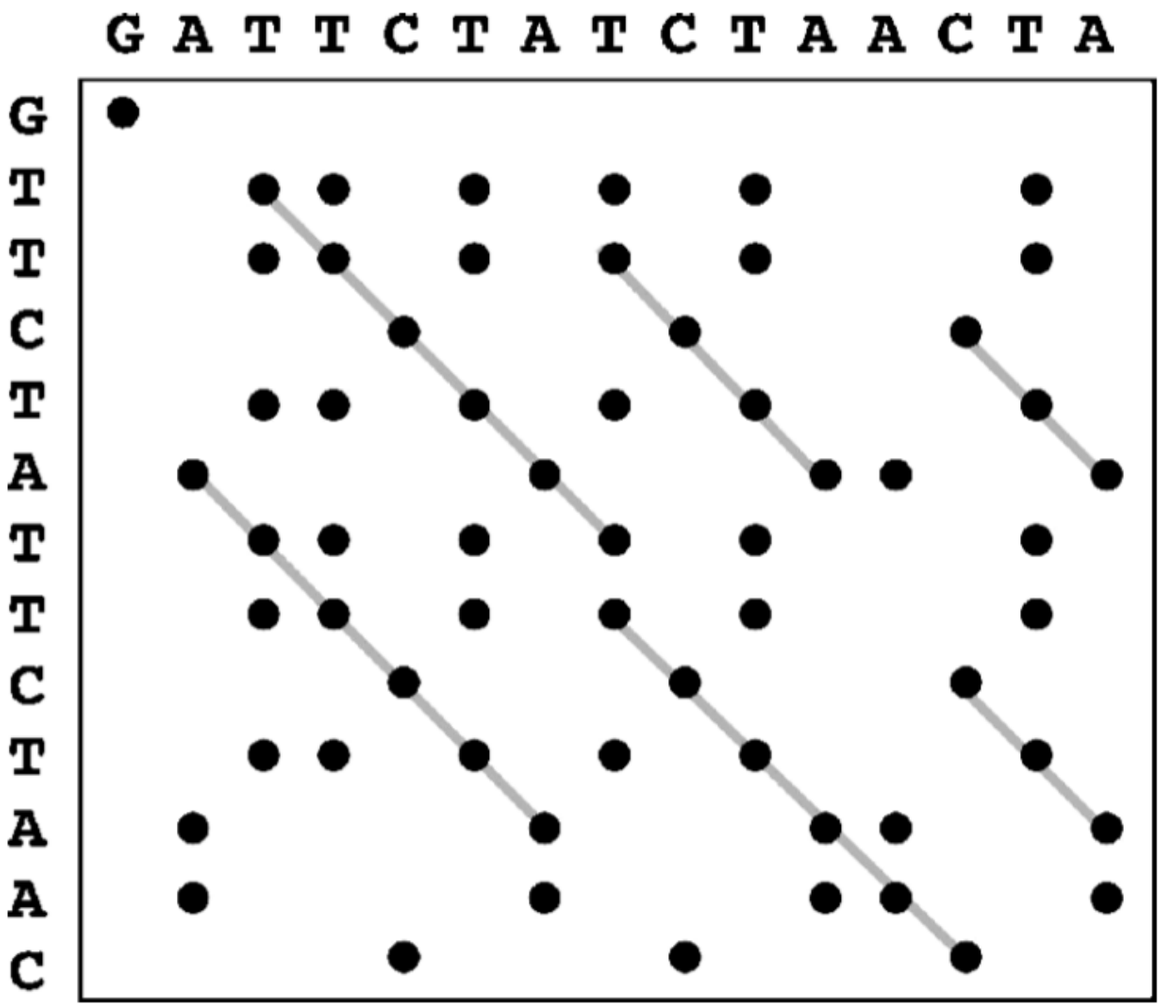


Figure 3.3: Example of comparing two sequences using dot plots. Lines linking the dots in diagonals indicate sequence alignment. Diagonal lines above or below the main diagonal represent internal repeats of either sequence.

Strengths of the dot matrix method

- *Self-alignment* (alignment of a sequence with itself) for the identification of repeated elements
- Alignment of a sequence with its *reverse complement* for the identification of complementary DNA or RNA sequences that are capable of forming hairpin structures

Drawback of the dot matrix method: noise

- Usually a high level of noise, *i.e.* dots are present all over the graph
- Especially problematic for DNA sequences, as 25% of the fields contain a dot merely by chance
- Noise reduction: usage of a "window" (= *"tuple"*) of a fixed length covering a stretch of residue pairs:
  - dots are only placed when sequences match over the entire window size

Other drawbacks of the dot matrix method

- Final alignment remains up to the user, who has to link nearby diagonals "by hand"
- Lack of statistical rigor in assessing the quality of the alignment
- Restricted to pairwise alignment

The dynamic programming method

- Much like the dot matrix method, this method works by creating a two-dimensional alignment grid with the two sequences along the axes
- Instead of dots, a score is attributed and placed into the matrix (*e.g.* +1 for a match, 0 for a mismatch)
- Scores are calculated row after row, from left to right
- Scoring takes into account the scores of previous rows: to each score, the highest of the three neighbouring scores to the left (in the same row), left-above and straight-above (in the previous row) is added

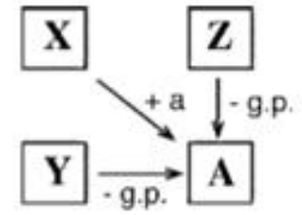
Example of a dynamic programming matrix

Goal: aligning the sequences

ATTGC

and

AGGC



A is the maximum score from one of the three directions plus matching score at the current position

	A	T	T	G	C
A	1	0	0	0	0
G					
G					
C					

	A	T	T	G	C
A	1	0	0	0	0
G	0	1			
G					
C					

	A	T	T	G	C
A	1	0	0	0	0
G	0	1	1	2	
G					
C					

	A	T	T	G	C
A	1	0	0	0	0
G	0	1	1	1	
G					
C					

	A	T	T	G	C
A	1	0	0	0	0
G	0	1	1	2	2
G	0	1	1	3	3
C	0	1	1	3	4

Dynamic programming: interpreting the matrix

- The optimal alignment is found by tracing back through the matrix in reverse order from the highest score in lower right-hand corner of the matrix toward the upper left-hand corner
- The best matching path is the one that has the maximum total score (sum of all scores in the path)
- Deciding between paths that reach the same highest score is arbitrary

Dynamic programming: interpreting the matrix

Goal: aligning the sequences

ATTGC

and

AGGC

		A	T	T	G	C
		1	0	0	0	0
A	1	0	0	0	0	
G	0	1	1	2	2	
G	0	1	1	3	3	
C	0	1	1	3	4	

		A	T	T	G	C
		(1)	0	0	0	0
A	(1)	0	0	0	0	
G	0	(1)	(1)	2	2	
G	0	1	1	(3)	3	
C	0	1	1	3	(4)	

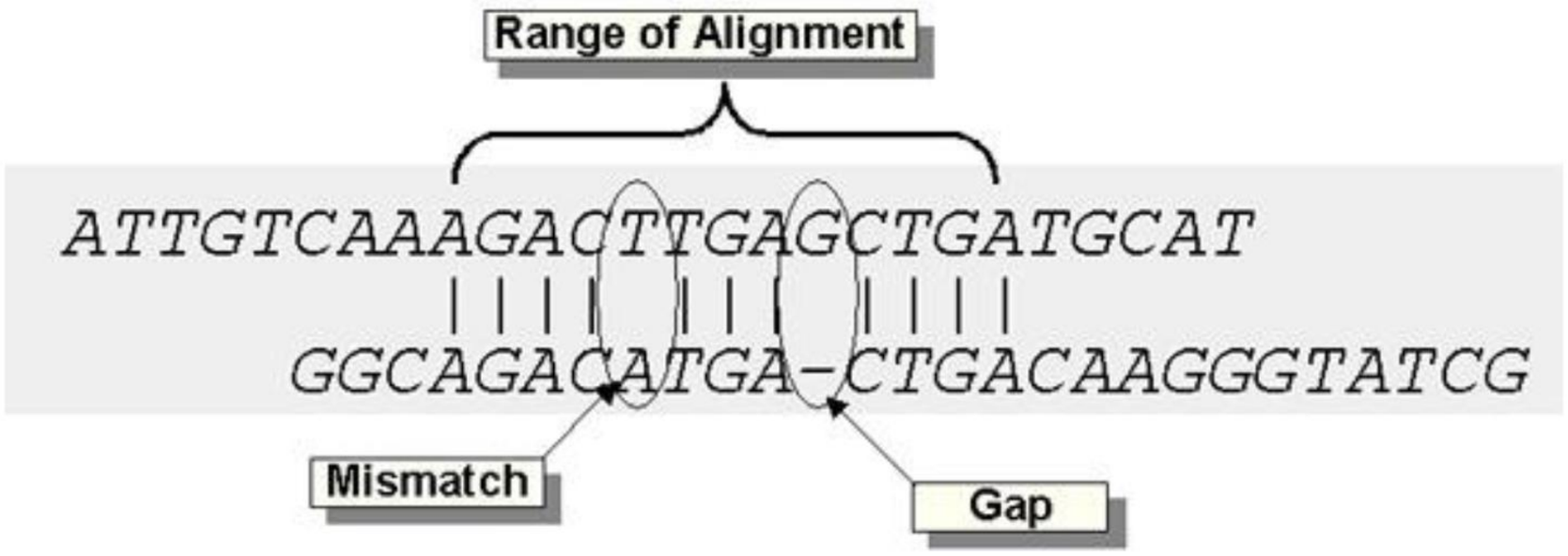
Final Alignment:

```
A T T G C
|     | |
A - G G C
```

Gap penalties

- The introduction of a gap in a protein sequence in the course of evolution is a relatively rare event (in comparison to single amino acid substitutions)
- To reflect this, many alignment methods apply penalty values for introducing gaps
- No evolutionary theory exists that allows establishing the exact "cost" for introducing insertions and deletions
- The penalty values are therefore more or less arbitrary, but:
  - values that are too low lead to too many gaps and the alignment of unrelated residues within the sequences
  - values that are too high force the matching of unrelated regions (in cases where a gap is truly present)

Applying gap penalties



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

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

http://www.ncbi.nlm.nih.gov/books/NBK62051/

The exact gap penalty values are usually derived empirically, this seems to be suitable for most alignment purposes

Affine gap penalties

- If insertions and deletions occur, several adjacent residues are likely to have been inserted or deleted together
- *Affine gap penalties*: much higher gap penalties for gap opening than for gap extension, *e.g.* -12 / -1 for gap opening / extension
- "Gaps" at terminal regions allowed with no penalty at all, to account for differences in sequence length

Example of a dynamic programming method for global alignment: the Needleman-Wunsch algorithm

- Designed for full-length sequence alignment, risks missing best local similarity
- Only suitable for aligning two closely related sequences of the same length
- Web-based Needleman-Wunsch alignment at EBI: https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle

Example: Needleman-Wunsch algorithm

X	Δ	A	V	C	N	E	R	C	K	L	C	K	P	M
Δ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	1	1	1	1	1	1	1	1	1	1	1	1	1
E	0	1	1	1	1	2	2	2	2	2	2	2	2	2
C	0	1	1	2	2	2	2	3	3	3	3	3	3	3
E	0	1	1	2	2	3	3	3	3	3	3	3	3	3
N	0	1	1	2	3	3	3	3	3	3	3	3	3	3
R	0	1	1	2	3	3	4	4	4	4	4	4	4	4
C	0	1	1	2	3	3	4	5	5	5	5	5	5	5
K	0	1	1	2	3	3	4	5	6	6	6	6	6	6
C	0	1	1	2	3	3	4	5	6	6	7	7	7	7
R	0	1	1	2	3	3	4	5	6	6	7	7	7	7
D	0	1	1	2	3	3	4	5	6	6	7	7	7	7
P	0	1	1	2	3	3	4	5	6	6	7	7	8	8

```
AEC ENRCK CRDP              AECEN RCK CRDP
AVCNE RCKLC KPM             AVC NERCKLC KPM
Score=8 (8 residues matched)
```

Local Alignment: Smith-Waterman algorithm

- More complex scoring system with gap penalties included in the horizontal and vertical steps
- Negative scores set to zero
- Alignment path may begin and end internally along the main diagonal
- Suitable for more divergent sequences
- Smith-Waterman alignment at EBI: https://www.ebi.ac.uk/jdispatcher/psa/emboss_water

Example: Smith-Waterman algorithm

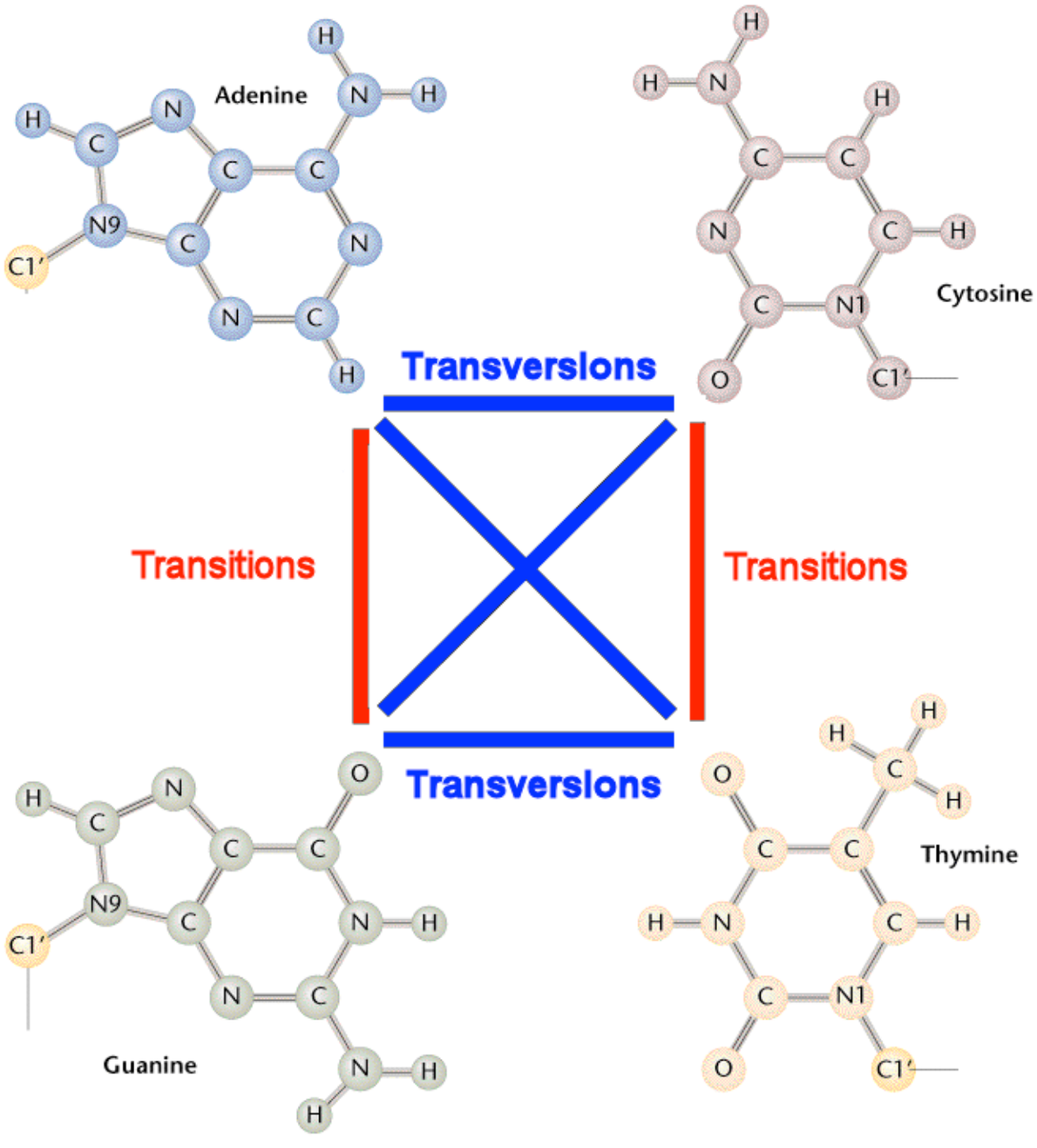
	Δ	A	V	C	N	E	R	C	K	L	C	K	P	M
Δ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	1	0	0	0	0	0	0	0	0
C	0	0	0	1	0	0	0	1	0	0	1	0	0	0
E	0	0	0	0	0	1	0	0	0	0	0	0	0	0
N	0	0	0	0	1	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0
C	0	0	0	1	0	0	0	2	0	0	1	0	0	0
K	0	0	0	0	0	0	0	0	3	1	0	2	0	0
C	0	0	0	1	0	0	0	1	1	2	2	0	1	0
R	0	0	0	0	0	0	1	0	0	0	1	1	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Scoring matrices

- Scoring matrices can be used instead of simple 1 / 0 scores for match / mismatch, which improves the quality of alignments
- The likelihood of one residue being substituted by another in an alignment = *substitution matrix*

Scoring matrices for nucleotide sequences

- Positive value for a match, negative value for mismatch
- *Transitions* (A↔G, C↔T) occur more frequently than *transversions* (A↔C, A↔T, G↔C, G↔T)



Scoring matrices for amino acid sequences

- Reflect similar physicochemical properties of groups of amino acid residues
- Substitutions between residues of different physicochemical properties are more likely to cause disruptions to the structure and function of a protein
- Disruptive substitutions render proteins non-functional and are unlikely to survive natural selection

Amino Acid Group	Amino Acid Name	Three- and One-Letter Code	Main Functional Features
Small and nonpolar	Glycine Alanine Proline	Gly, G Ala, A Pro, P	Nonreactive in chemical reactions; Pro and Gly disrupt regular secondary structures
Small and polar	Cysteine Serine Threonine	Cys, C Ser, S Thr, T	Serving as posttranslational modification sites and participating in active sites of enzymes or binding metal
Large and polar	Glutamine Asparagine	Gln, Q Asn, N	Participating in hydrogen bonding or in enzyme active sites
Large and polar (basic)	Arginine Lysine Histidine	Arg, R Lys, K His, H	Found in the surface of globular proteins providing salt bridges; His participates in enzyme catalysis or metal binding
Large and polar (acidic)	Glutamate Aspartate	Glu, E Asp, D	Found in the surface of globular proteins providing salt bridges
Large and nonpolar (aliphatic)	Isoleucine Leucine Methionine Valine	Ile, I Leu, L Met, M Val, V	Nonreactive in chemical reactions; participating in hydrophobic interactions
Large and nonpolar (aromatic)	Phenylalanine Tyrosine Tryptophan	Phe, F Tyr, Y Trp, W	Providing sites for aromatic packing interactions; Tyr and Trp are weakly polar and can serve as sites for phosphorylation and hydrogen bonding

Empirical amino acid scoring matrices

- Scoring matrices are 20 × 20 matrices corresponding to "mutation likelihood" for amino acid pairs
- Usually not derived from amino acid properties directly, but from empirical studies of amino acid substitutions, *e.g.* PAM, BLOSUM:
  - Derived from actual alignments of highly similar sequences
  - Higher scores for more likely substitutions and lower scores for rare substitutions

Constructing amino acid scoring matrices

- First calculate the ratio of the observed frequency of a particular substitution in alignments (e.g. Ala->Gly) to the expected frequency if all substitutions were equally likely: $F_{Observed, Ala->Gly} / F_{Expected, Ala->Gly}$
- Then, the matrix is filled with the logarithms of these ratios (these values are called *log-odds scores* or *log-odds ratios*)
- *E.g.* if a mutation happens 100x more often in an alignment than expected if all mutations were equally likely, the ratio is 100 / 1 = 100, and the score in the matrix is log(100) = +2
- Logarithm either to the base of 10 or the base of 2

The values in amino acid scoring matrices

- Positive score means that frequency of an amino acid substitution is greater than expected by chance (i.e. if every mutation were equally likely to happen)
- Zero score: frequency of a substitution is equal to that expected by chance
- Negative score: frequency of a substitution is less than would by expected by chance

Dayhoff PAM matrices (1966)

Observed mutations were not expected to significantly change the function of the proteins, *i.e.* considered to be accepted by natural selection, therefore: **PAM** = "point accepted mutations"

- Margaret Dayhoff analysed alignments of 71 groups of very closely related protein sequences (with 1% of amino acids changed during evolution)
- This avoids problems due to inaccurate alignment or multiple mutations occurring at the same position
- Frequencies of specific amino acid substitutions within these groups were determined in the context of phylogenetic trees

Extrapolation of PAM matrices

- PAM<N> matrices for use with more divergent sequences are extrapolated from PAM1 *via* matrix multiplication
- *E.g.* PAM250 corresponds to 20% amino acid identity, 250 mutations per 100 residues and approximately corresponds to an expected evolutionary span of 2,500 million years
- PAM matrices with lower <N> are more suitable for aligning more closely relates sequences

Examples of PAM matrices

TABLE 3.1. Correspondence of PAM Numbers with Observed Amino Acid Mutational Rates

PAM Number	Observed Mutation Rate (%)	Sequence Identity (%)
0	0	100
1	1	99
30	25	75
80	50	50
110	40	60
200	75	25
250	80	20

PAM250

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

Figure 3.5: PAM250 amino acid substitution matrix. Residues are grouped according to physicochemical similarities.

BLOSUM matrices

BLOSUM matrices (Henikoff & Henikoff, 1992)

- BLOSUM = block substitution matrix
- Directly derived from multiple sequence alignments of more than 2000 conserved amino acid patterns (= blocks) representing 500 groups of protein sequences
- No extrapolation
- BLOSUM62: matrix constructed from sequences with average identity value of 62% (reverse ordering as PAM numbering system!)

BLOSUM62

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

Figure 3.6: BLOSUM62 amino acid substitution matrix.

Comparison between PAM and BLOSUM matrices

PAM:

- Derived from an evolutionary model and therefore used for reconstructing phylogenetic trees
- Based on global alignment of both conserved and variable regions

BLOSUM:

- Based entirely on direct observations (no extrapolation) and are perhaps more realistic for divergent sequences
- Based only on conserved sequence blocks, *i.e.* advantageous for finding conserved domains

PAM and BLOSUM matrices in practice

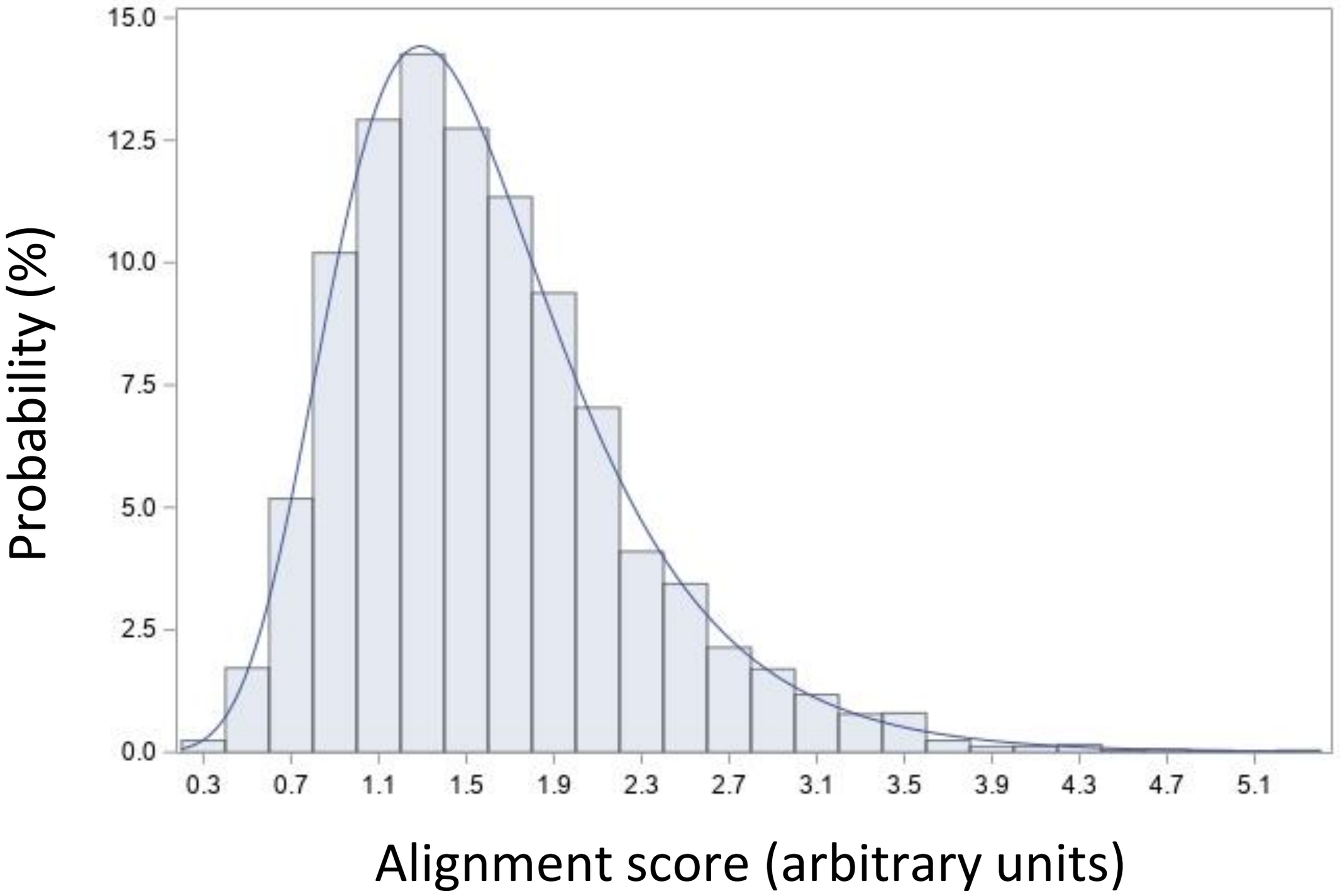
- BLOSUM outperforms PAM in terms of accuracy of local alignment
- Newer matrices based on PAM approach, *e.g.* Gonnet matrices, Jones-Taylor-Thornton (JTT) matrices:
  - Equivalent performance to BLOSUM in alignment
  - Particularly robust in phylogenetic tree construction

Statistical significance of a given sequence alignment

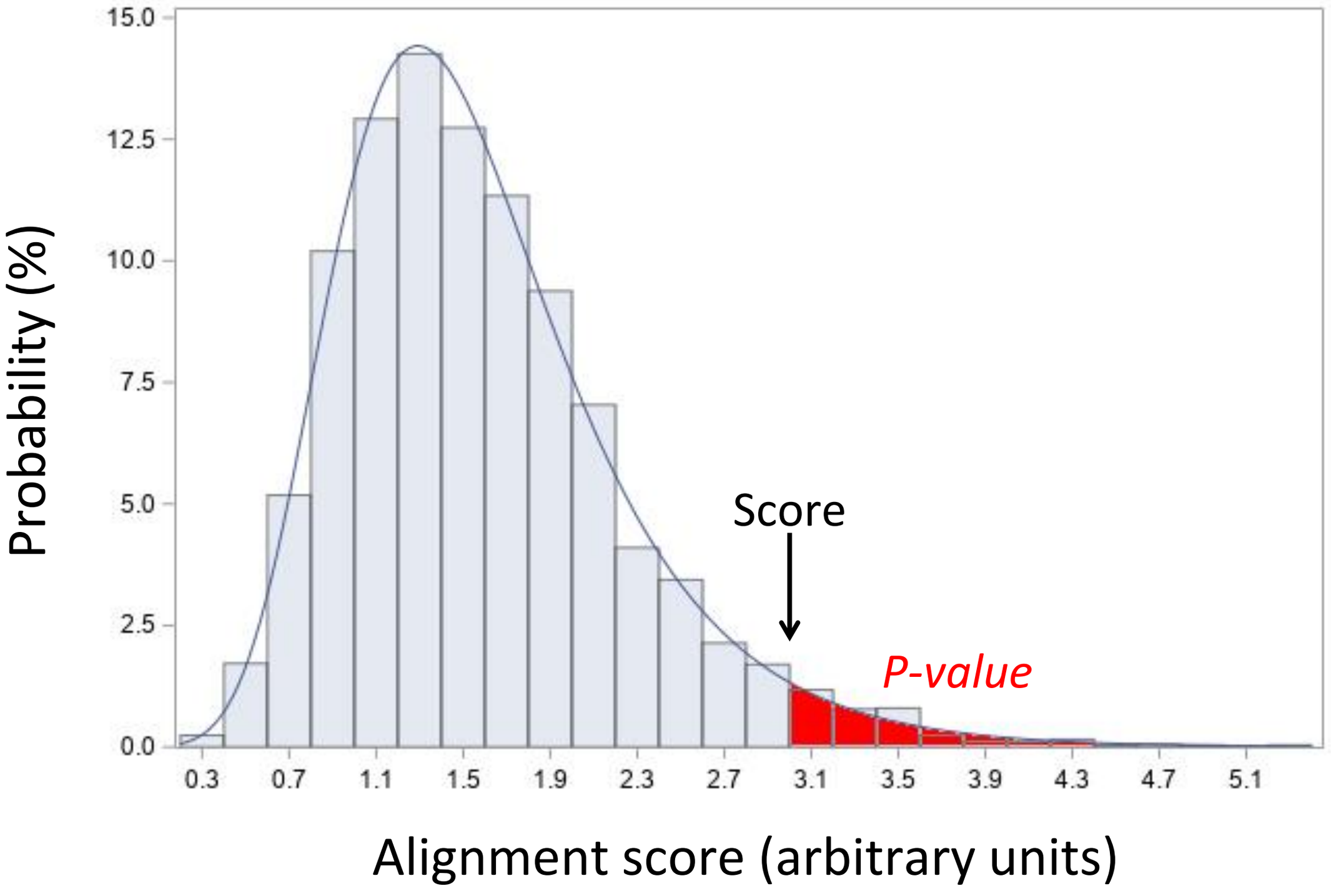
Only statistically significant sequence alignment provides evidence of homology

- Needed: comparison of the alignment score with those obtained when aligning random sequences (which are totally unrelated) of similar length
- Calculation of large numbers of alignment scores for randomised sequences gives the so-called *"Gumbel extreme value distribution"* (Emil Julius Gumbel 1891-1966)
- A Gumbel distribution can be used to find the probability that a given score occurs purely due to chance

Gumbel extreme value distribution



Gumbel extreme value distribution



P-value (probability value)

Probability of achieving an equal or higher score for an alignment of random sequences:

- $P = 10^{-1}$ means that there is a 10% chance that the two sequences may be randomly related, usually interpreted as no evidence for homology
- $P = 10^{-1}$ to 10^{-5} indicates possible distant homologs
- $P = 10^{-5}$ to 10^{-50} is interpreted as sequences having clear homology
- $P = 10^{-50}$ to 10^{-100} is considered a nearly identical match
- $P < 10^{-100}$ indicates an exact match