Sequence Pairwise

Alignment Techniques

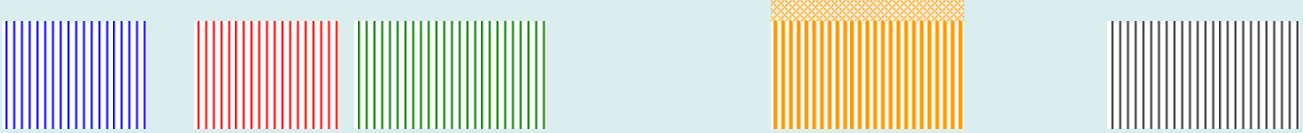
WHEN ALIGNING SEQUENCES…

- There are a lot of possible solutions.
- Two sequences could ALWAYS be aligned!
- Alignments should be valued and evaluated.
- There could be frequently more than one solution of the same value.

POSSIBLE SOLUTIONS

ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERNALEDLAGERDFWGALSTOUTWRARWATERA



POSSIBLE SOLUTIONS

POSSIBLE SOLUTIONS

POSSIBLE SOLUTIONS

POSSIBLE SOLUTIONS

WHAT IS THE CORRECT ONE?

```
actaccagttcatttgatacttctcaaa
 taccattaccgtgttaactgaaaggacttaaagact
```

```
actaccagttcatttgatacttctcaaa
taccattaccgtgttaactgaaaggacttaaagact
```

Sequence 1
Sequence 2

```
actaccagttcatttgatacttctcaaa
    taccattaccgtgttaactgaaaggacttaaagact
```

```
actaccagttcatttgatacttctcaaa
  taccattaccgtgttaactgaaaggacttaaagact
```

```
actaccagttcatttgatacttctcaaa
   taccattaccgtgttaactgaaaggacttaaagact
```

```
actaccagttcatttgatacttctcaaa
     taccattaccgtgttaactgaaaggacttaaagact
```

```
actaccagttcatttgatacttctcaaa
      taccattaccgtgttaactgaaaggacttaaagact
```

```
  actaccagttcatttgatacttctcaaa
taccattaccgtgttaactgaaaggacttaaagact
```

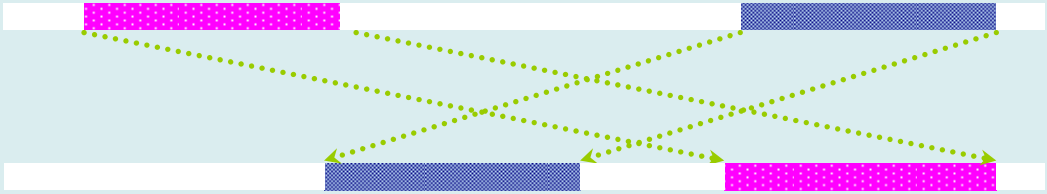
```
      actaccagttcatttgatacttctcaaa
taccattaccgtgttaactgaaaggacttaaagact
```

```
 actaccagttcatttgatacttctcaaa
taccattaccgtgttaactgaaaggacttaaagact
```

LOCAL and GLOBAL implementations

LOCAL Alignment

GCG : **bestfit**
Staden : **spin**
Emboss : **water/matcher**



GLOBAL Alignment

GCG : **gap**
Staden : **spin**
Emboss : **needle/stretcher**

These algorithms are very precise

PAIRWISE COMPARISON

- Local alignment (Smith-Waterman)
  - Identifies the most similar region shared between two sequences

```
14  TCAGAAGCAGCTAAAGCGT  32
    ||||||||| |||||||||
42  TCAGAAGCA.CTAAAGCGT  59
```

water *(Smith-Waterman)*

LOCAL ALIGNMENT

```
14   TCAGAAGCAGCTAAAGCGT   32
     ||||||||| |||||||||
42   TCAGAAGCA.CTAAAGCGT   59

 1   AGGATTGGAATGCT   14
     ||||||||||||||
 1   AGGATTGGAATGCT   14

39   AGGATTGGAAT   49
     ||||||||||||
 1   AGGATTGGAAT   11

62   AGACCG   67
     ||||||
66   AGACCG   71
```

Similarity *(X. Huang)* **shows all similar regions**

PAIRWISE COMPARISON

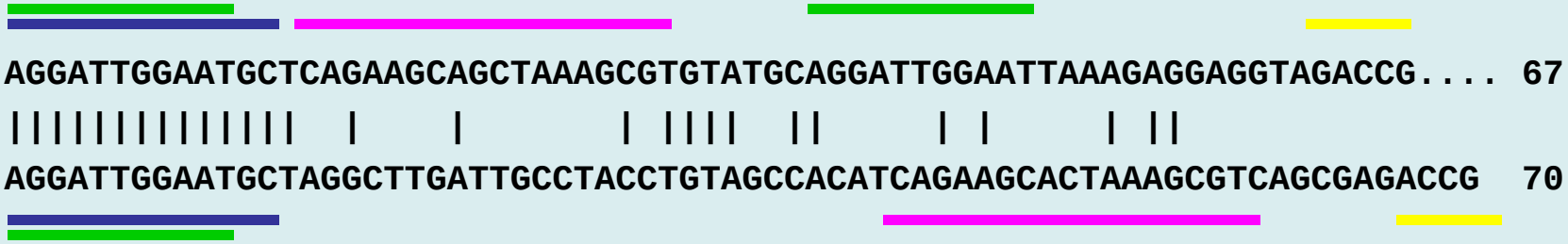
- Global alignment (Needleman-Wunsch) aligns on whole sequences

```
1 TGTCGATTAAGCGGTCGTAGCTGACCTGAGATTGCCCGATGGCGTAGTAGCTGACC 56
  |||||||||| |||||||||||   |||||||| || ||||||||||||||||| |||||
1 TGTCGATTATGCGGTCGTAG..GACCTGAGTTTCCCCGATGGCGTAGTAGGTGACC 54
```

needle *(Needleman & Wunsch)* **alignment from beggining to end**

GLOBAL ALIGNMENT

Very frequently, two sequences share several regions with local similarity:



How is alignment done?

DNA SCORING SYSTEM

Sequence 1 `actaccagttcatttgatacttctcaaa`

Sequence 2 `taccattaccgtgttaactgaaaggacttaaagact`

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

Match: 1
Mismatch: 0
Score = 5

DNA SCORING SYSTEM

Sequence 1 `actaccagttcatttgatacttctcaaa`

Sequence 2 `taccattaccgtgttaactgaaaggacttaaagact`

Negative values to penalise no coincidences (“mismatches”):

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

Matches: 5
Mismatches: 19

Score: 5 x 5 + 19 x (-4) = - 51

PROTEINS SCORING SYSTEM

Sequence 1 **PTHPLASKTQILPEDLASEDLTI**

|||||| | || ||

Sequence 2 **PTHPLAGERAIGLARLAEEDFGM**

Scoring matrix

	C	S	T	P	A	G	N	D	.	.
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	5			
D	-3	0	-1	-1	-2	-1	1	6		
.										
.										

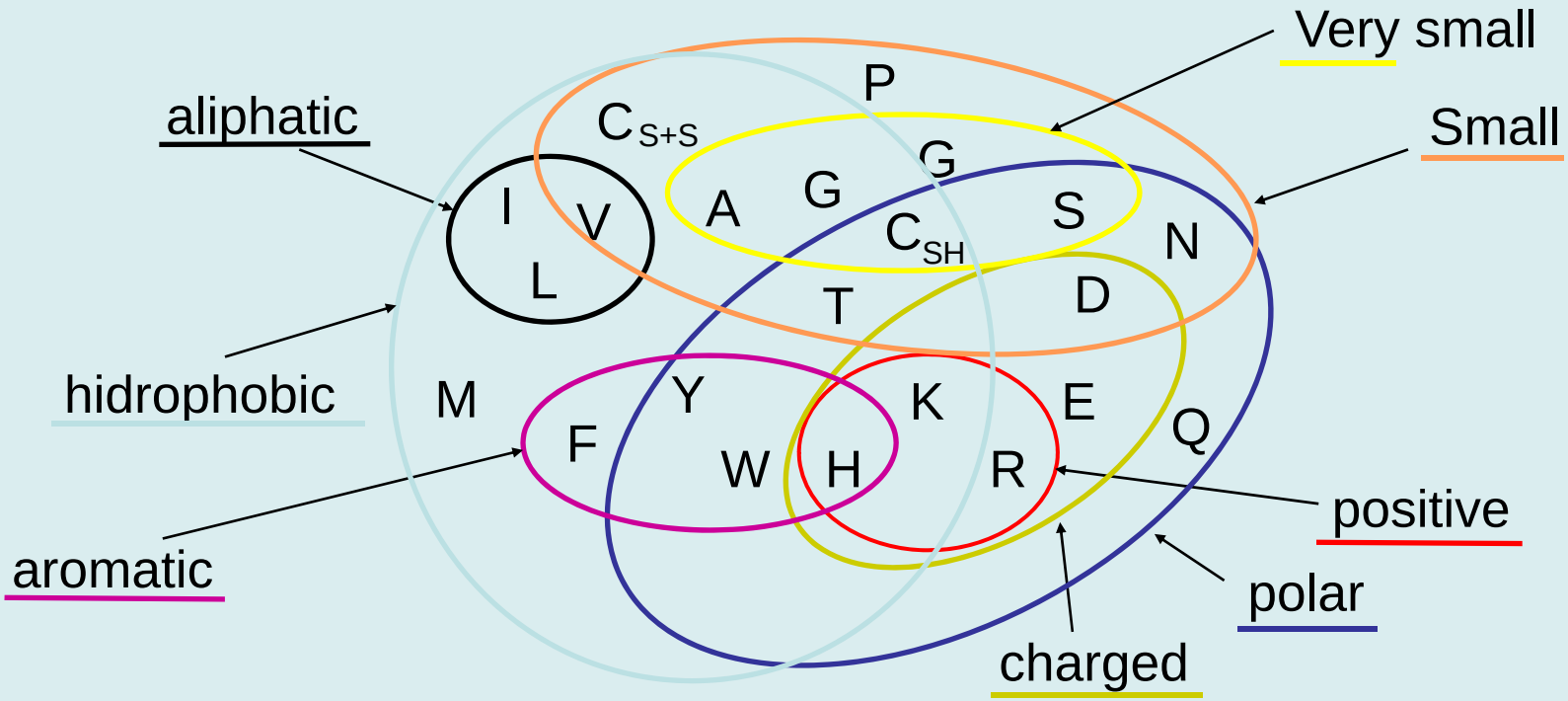
T:G = -2

T:T = 5

Score = 48

PROTEINS SCORING SYSTEM

• Amino acids have different biophysical and biochemical characteristics which influence their replacement ability (mutation) in the evolutive process.



SCORING SYSTEMS

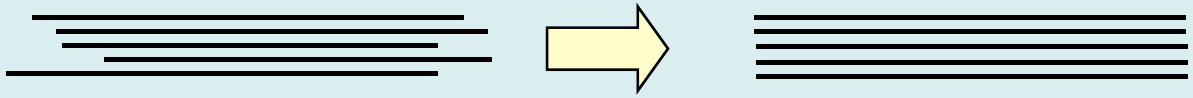
- There exists different scoring systems depending on the physical and chemical characteristics which determine their replacement ability in evolution.

- Scoring matrices show:
  - mutual substitution probability
  - probability of appearance of each amino acid

- Most commonly used matrices:
  - **PAM**
  - **BLOSUM**

MATRICES PAM (Percent Accepted Mutations)

• Come from global alignments from protein families, sharing about 85% identity.



•Construction of a phylogenetic tree and ancestor sequences for each protein family

• Calculation of the number of replacements for each pair of amino acids

MATRICES PAM (Percent Accepted Mutations)

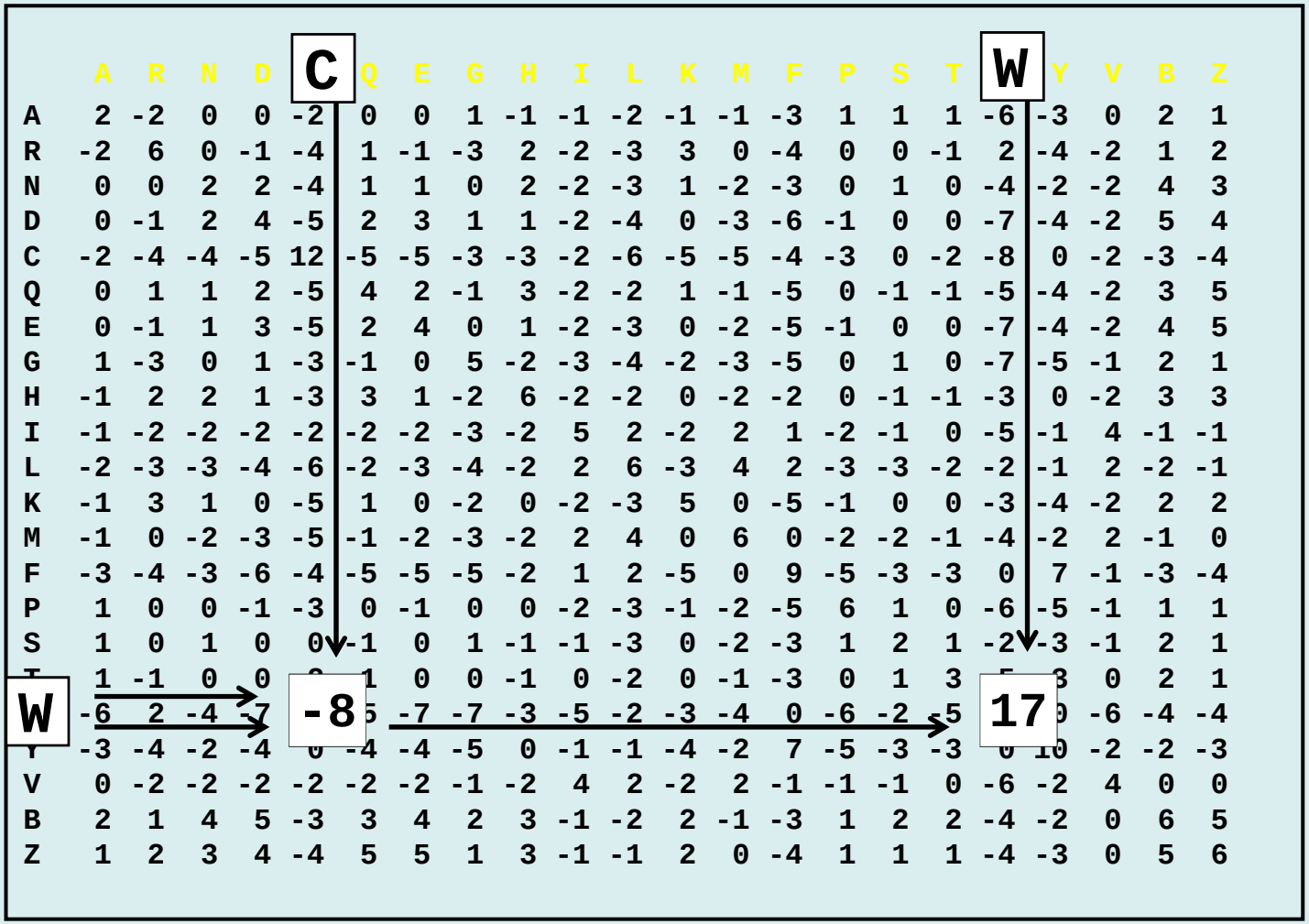
- The number of replacements is used to calculate the matrix PAM-1.

- PAM-1 reflects an change average of 1% in all positions. Matrices PAM for more distant evolution changes are extrapolated from matrix PAM-1.

- PAM250 = 250 mutations for each 100 residues.

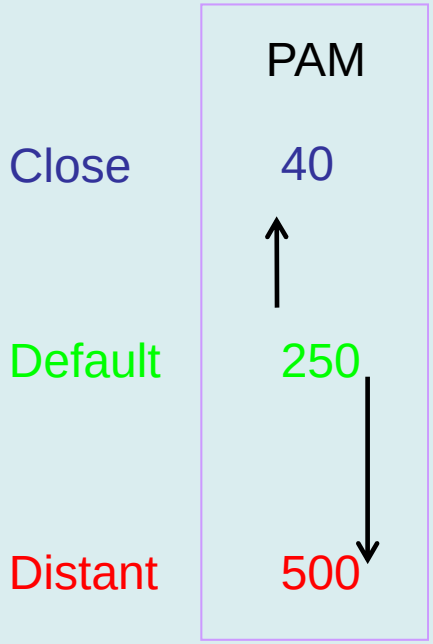
- Higher values reflect more evolutive distance

PAM 250



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

PAM DISTANCES

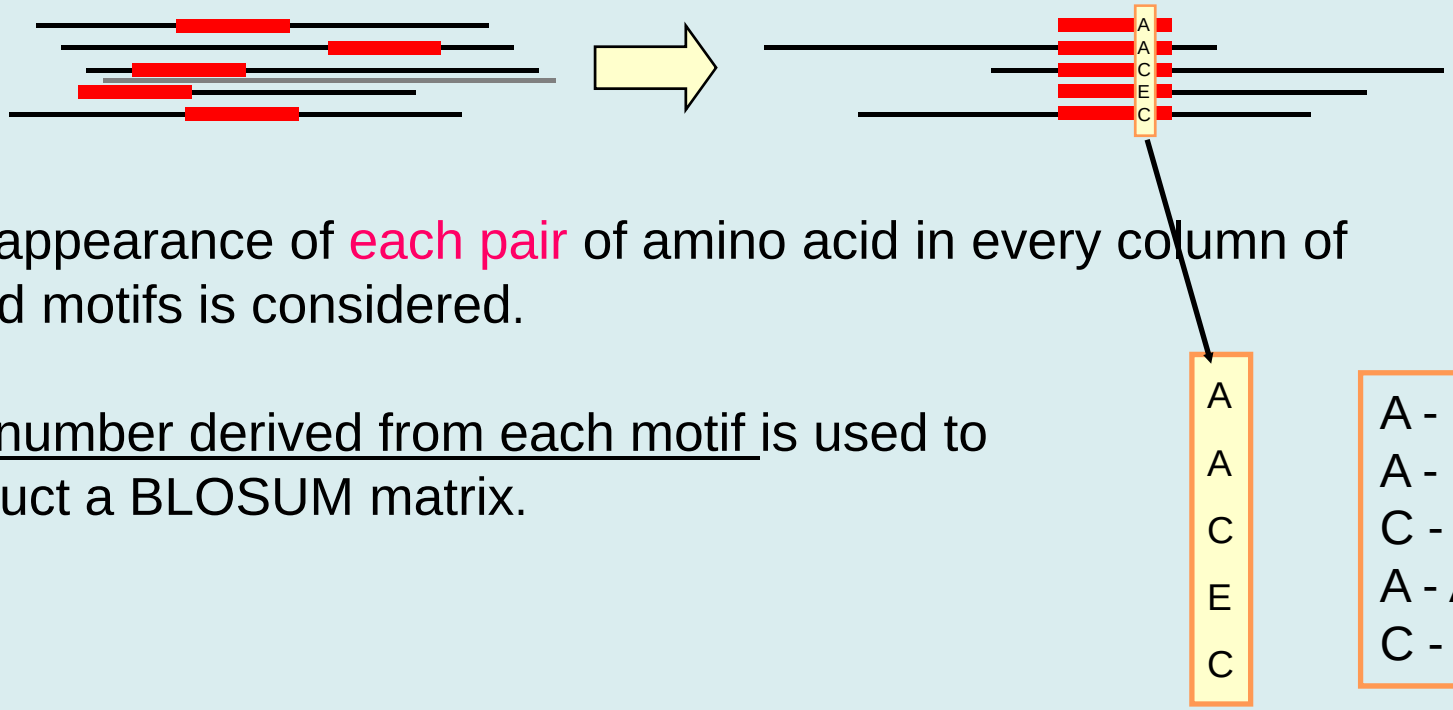


PAM – is a measure of evolutionary distance

An evolutionary distance of 1 PAM indicates the probability of a residue mutating during a distance in which 1 point mutation was accepted per 100 residues.

BLOSUM (Blocks Substitution Matrix)

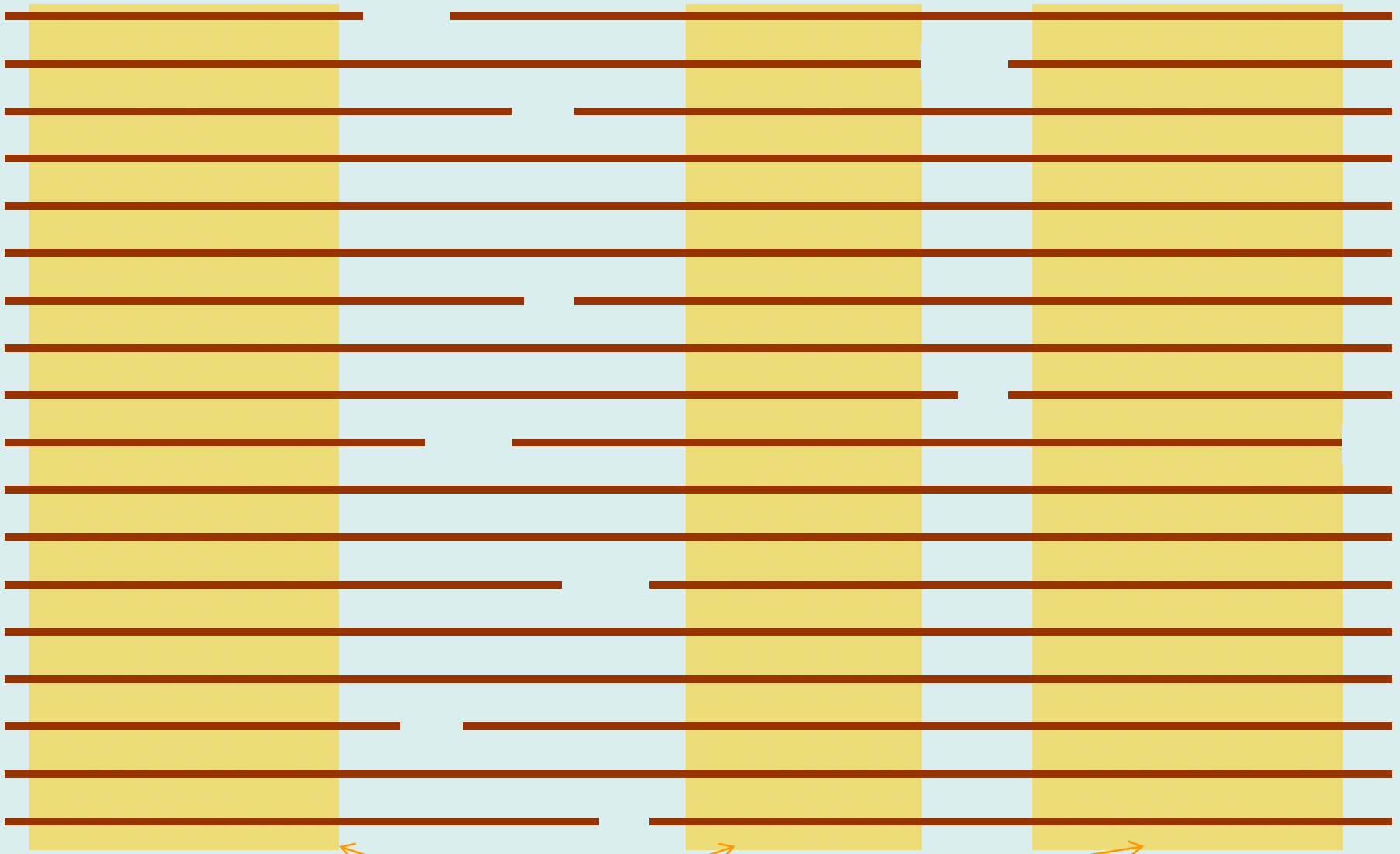
• Derives from alignment of domains of proteins evolutively distanced.



• The appearance of each pair of amino acid in every column of aligned motifs is considered.

• The number derived from each motif is used to construct a BLOSUM matrix.

THE BLOSUM SCORING MATRICES



Regions of high conservation, stored in the BLOCKS database

BLOSUM (Blocks Substitution Matrix)

- Sequences with motifs are grouped according to their level of identity.
- Groups are counted as unique sequences.
- Different BLOSUM matrices differ in the sequence identity percentage used in grouping.
- The number in the name of the matrix (ex. 62 in BLOSUM62) refers to the sequence identity percentage used to arrange the matrix.
- Higher numbers mean lower evolutive distance.

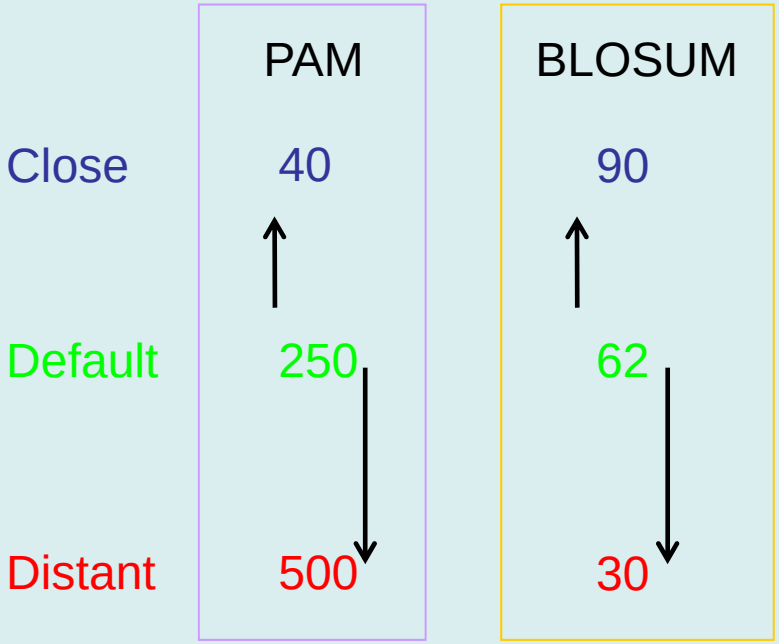
SCORING MATRIX ELECTION

• In general, BLOSUM matrices function better than PAM matrices for local identity searches.

• When comparing **very related** proteins **low PAM** or **high BLOSUM** matrices should be used.

•For proteins **evolutively distanced high PAM** or **low BLOSUM** matrices should be used.

•The most frequently used matrix is BLOSUM62.



INSERTIONS and DELETIONS (INDEL)

ACBEERGYALEDILAGERAFGSTOUTFAWATERM

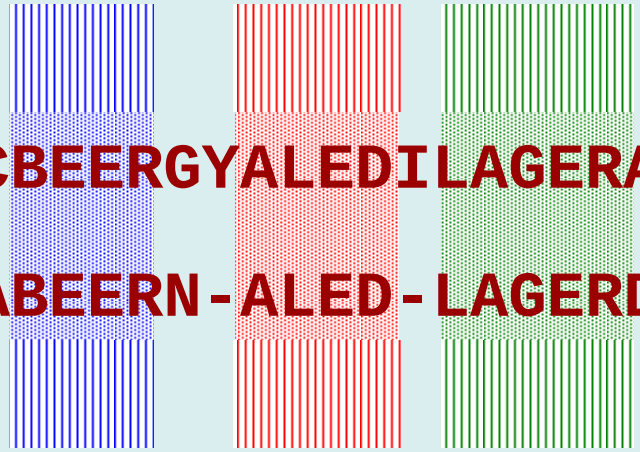
ABEERNALEDLAGERDFWGALSTOUTWRARWATERA

INSERTIONS and DELETIONS (INDEL)

ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERN-ALEDLAGERDFWGALSTOUTWRARWATERA

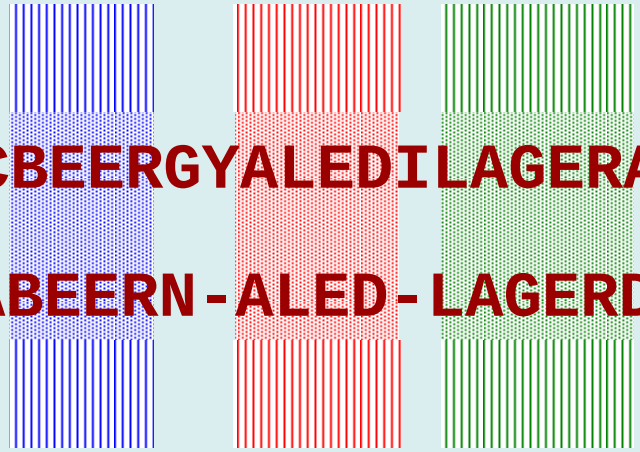
INSERTIONS and DELETIONS (INDEL)



ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERN-ALED-LAGERDFWGALSTOUTWRARWATERA

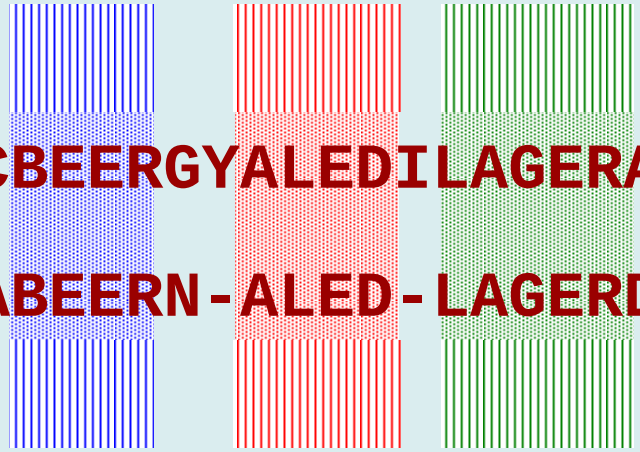
INSERTIONS and DELETIONS (INDEL)



ACBEERGYALEDILAGERAFG-STOUTFAWATERM

ABEERN-ALED-LAGERDFWGALSTOUTWRARWATERA

INSERTIONS and DELETIONS (INDEL)



ACBEERGYALEDILAGERAFG--STOUTFAWATERM

ABEERN-ALED-LAGERDFWGALSTOUTWRARWATERA

INSERTIONS and DELETIONS (INDEL)

INSERTIONS and DELETIONS (INDEL)

INSERTIONS and DELETIONS (INDEL)

OPTIMIZING THE ALIGNMENT

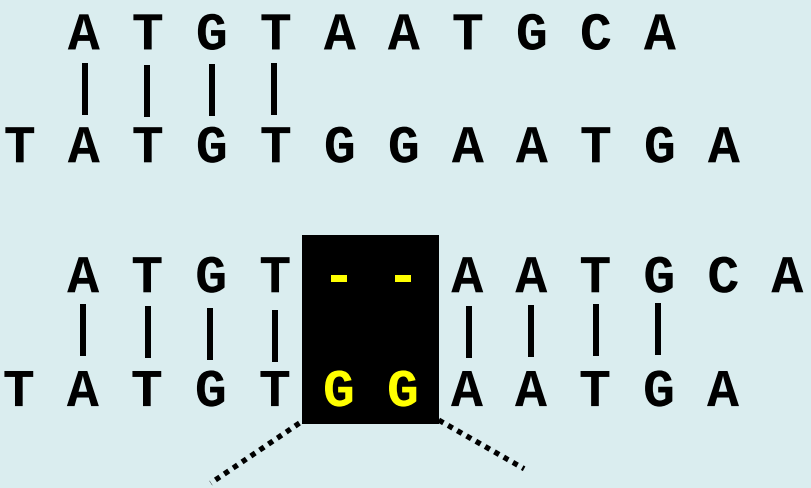
- The optimal alignment for two similar sequences is the one that
  - **maximizes** the number of coincidences
  - **minimizes** the number of gaps

- Allowing the arbitrary insertion of a not defined number of gaps leads to good scoring with non-homologue sequences.

- Penalizing gaps forces the alignment to have few ones.

- The raw score of a gapped alignment is the sum of all amino acid substitution from which we substract the gap opening and extension penalties.

Insertion and deletion (INDEL) evaluation



A gap creation is penalized with a negative value

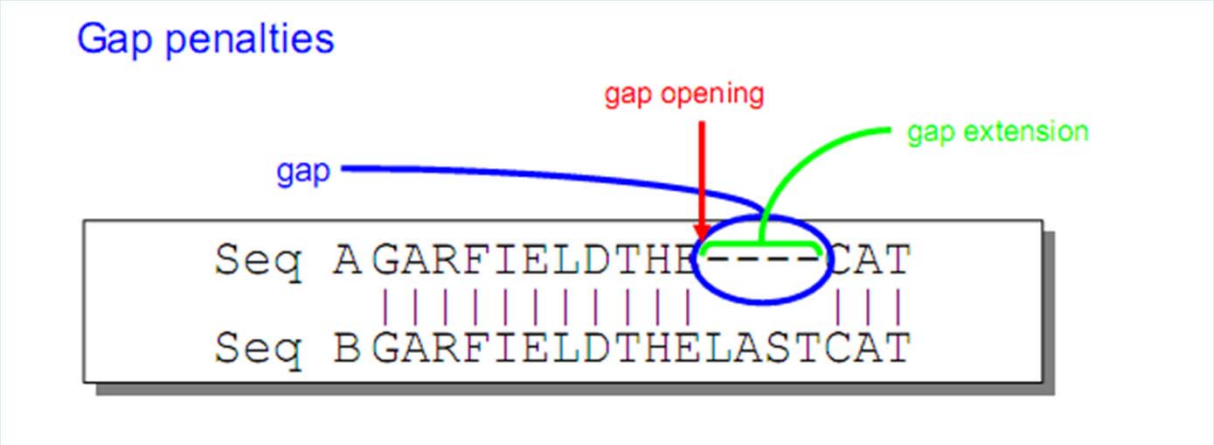
SEQUENCE ALIGNMENT PARAMETERS

Scoring systems:

• Each possible pairwise is assigned a numeric value and arranged in a table.

“Gap Penalties”:

• Gap: The cost of introducing a gap
• Extension: The cost of extending a gap



WHY GAPS SHOULD BE PENALISED?

Gaps not allowed Score: 0

```
1 GTGATAGACACAGACCGGTGGCATTGTGG 29
  |||   |  | |||    |   || || |
1 GTGTCGGGAAGAGATAACTCCGATGGTTG 29
```

Match = 5
Mismatch = -4

Gaps allowed but not penalised Score: 88

```
1 GTG.ATAG.ACACAGA..CCGGT..GGCATTGTGG 29
  ||| || | | | |||  ||  |  |  || || |
1 GTGTAT.GGA.AGAGATACC..TCCG..ATGGTTG 29
```

TYPES OF PENALTIES

Lineal:

$$\gamma(g) = - gd$$

Related:

$$\gamma(g) = -d - (g -1)e$$

$\gamma(g)$ = gap penalty of lenght g

d = gap open penalty

e = Extended penalty

g = gap length

DELETION AND INSERTION SCORING

match = 1
mismatch = 0

Scoring: 4

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

Scoring: 8 - 3.2 = 4.8

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

insertion/ deletion

Parametres:

$d = 3$ (gap opening)

$e = 0.1$ (gap extension)

$g = 3$ (gap length)

$\gamma(g) = -3 - (3 - 1)\ 0.1 = -3.2$

GAP PENALTY MODIFICATIONS

Scoring matrix: BLOSUM62

Gap opening penalty = 3
Gap extension penalty = 0.1
Scoring = **6.3**

```
1  ...VLSPADKFLTNV 12
        ||||
1 VFTELSPAKTV.... 11
```

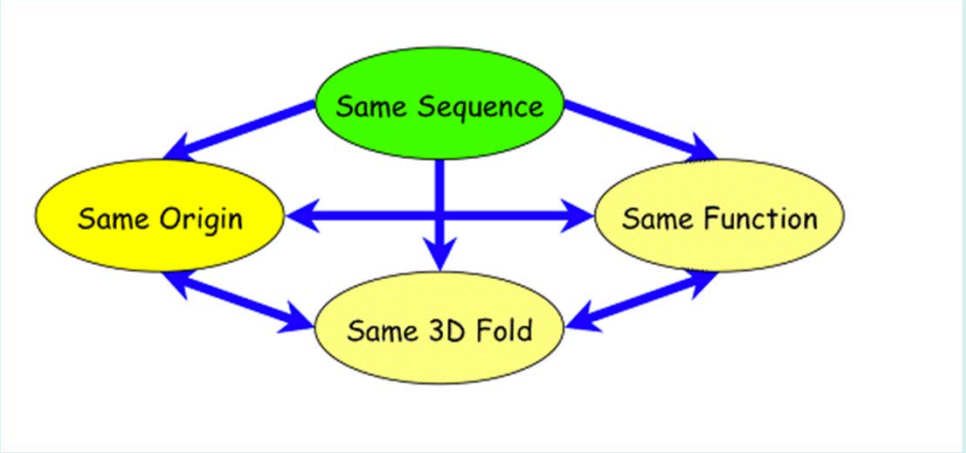
Gap opening penalty = 0
Gap extension penalty = 0.1
Scoring = **11.3**

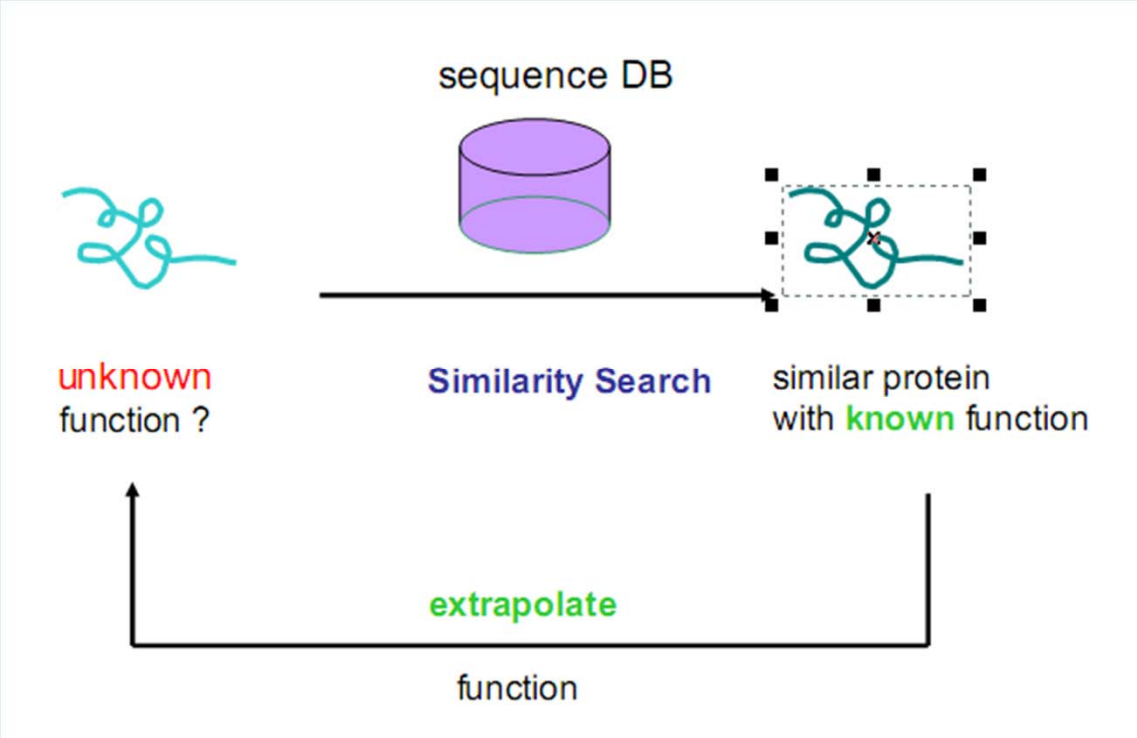
```
1 V...LSPADKFLTNV 12
  |   ||||  |  | |
1 VFTELSPA.K..T.V 11
```

Similarity Searches

IMPORTANCE OF SIMILARITY

Similar sequences: probably have the same ancestor, share the same structure, and have similar biological function





ALGORITHM TYPES

- Dynamic Algorithms (precise)
  - Global: Complete alignment between the sequences A and B (Needleman-Wunsch)
  - Local: Alignment between a sub-sequence of A and a sub-sequence of B N(Smith-Waterman)

- Heuristic Algorithm (imprecise)
  - BLAST
  - FASTA

COMPUTATIONAL PROBLEM

Dynamic Programming: computational method that provide in mathematical sense the best alignment between two sequences, given a scoring system.

BUT

Exact algorithms are time consuming. Example:

2 Globins	1 sec
3 Globins	1 min
4 Globins	5 hour (heuristic wished)
5 Globins	3 weeks (heuristic really wished)
6 Globins	9 years (heuristic required)
7 Globins	1000 years (definitely required!)

Heuristic Methods (e.g. BLAST, FASTA) they prune the search space by using fast approximate methods to select the sequences of the database that are likely to be similar to the query and to locate the similarity region inside them.

=> Restricting the alignment process:

Only to the selected sequences

Only to some portions of the sequences (search a fraction as small as possible of the cells in the dynamic programming matrix)

HEURISTIC SEQUENCE ALIGNMENT: PRINCIPLE

•These methods are heuristic; i.e., an empirical method of computer programming in which rules of thumb are used to find solutions.

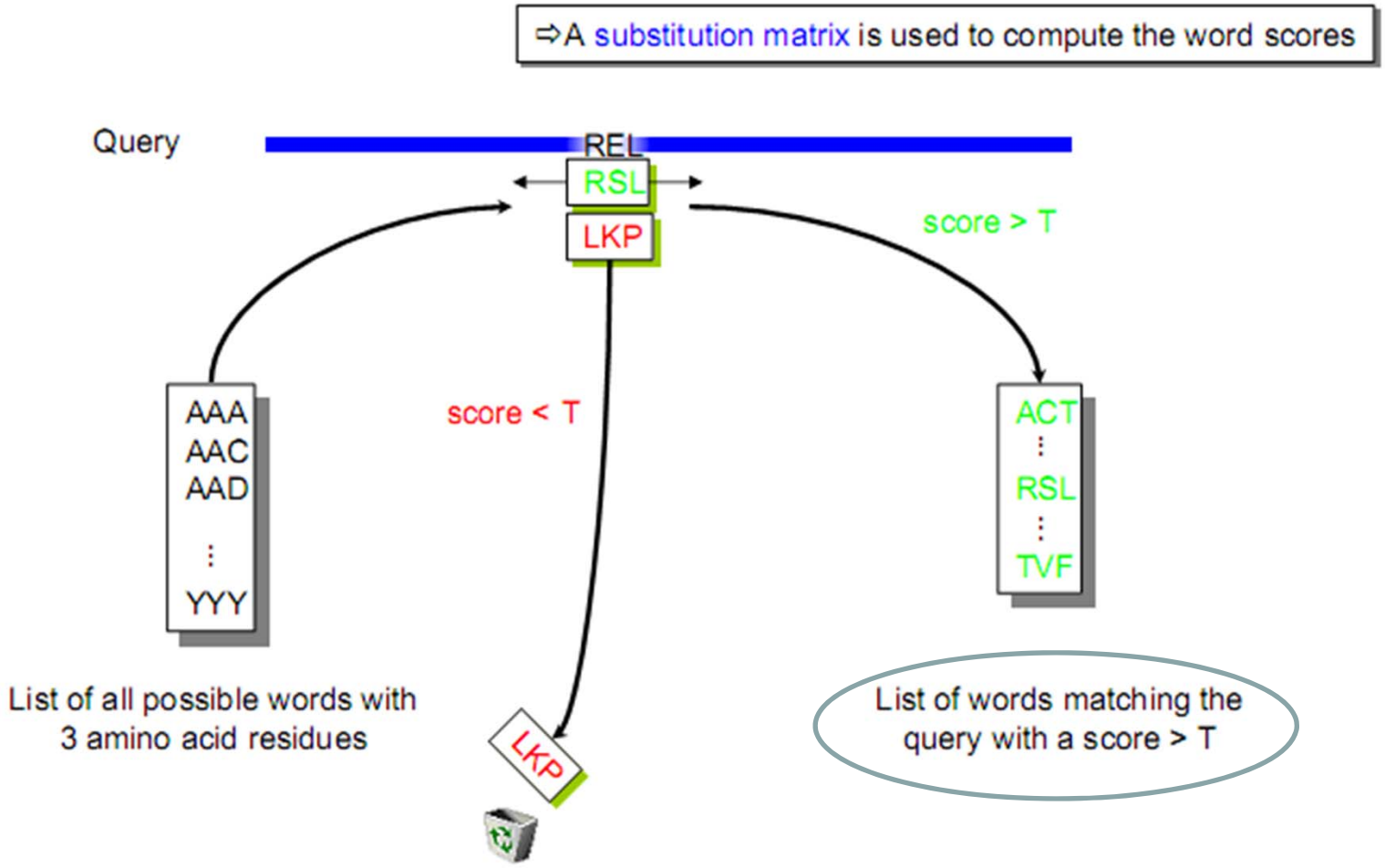
Rule-of-thumb: If your sequences are more than 100 amino acids long (or 100 nucleotides long) you can considered them as homologues if 25% of the amino acids are identical (70% of nucleotide for DNA). Below this value you enter the twilight zone.

•They almost always works to find related sequences in a database search but does not have the underlying guarantee of an optimal solution like the dynamic programming algorithm (But good ones often do).

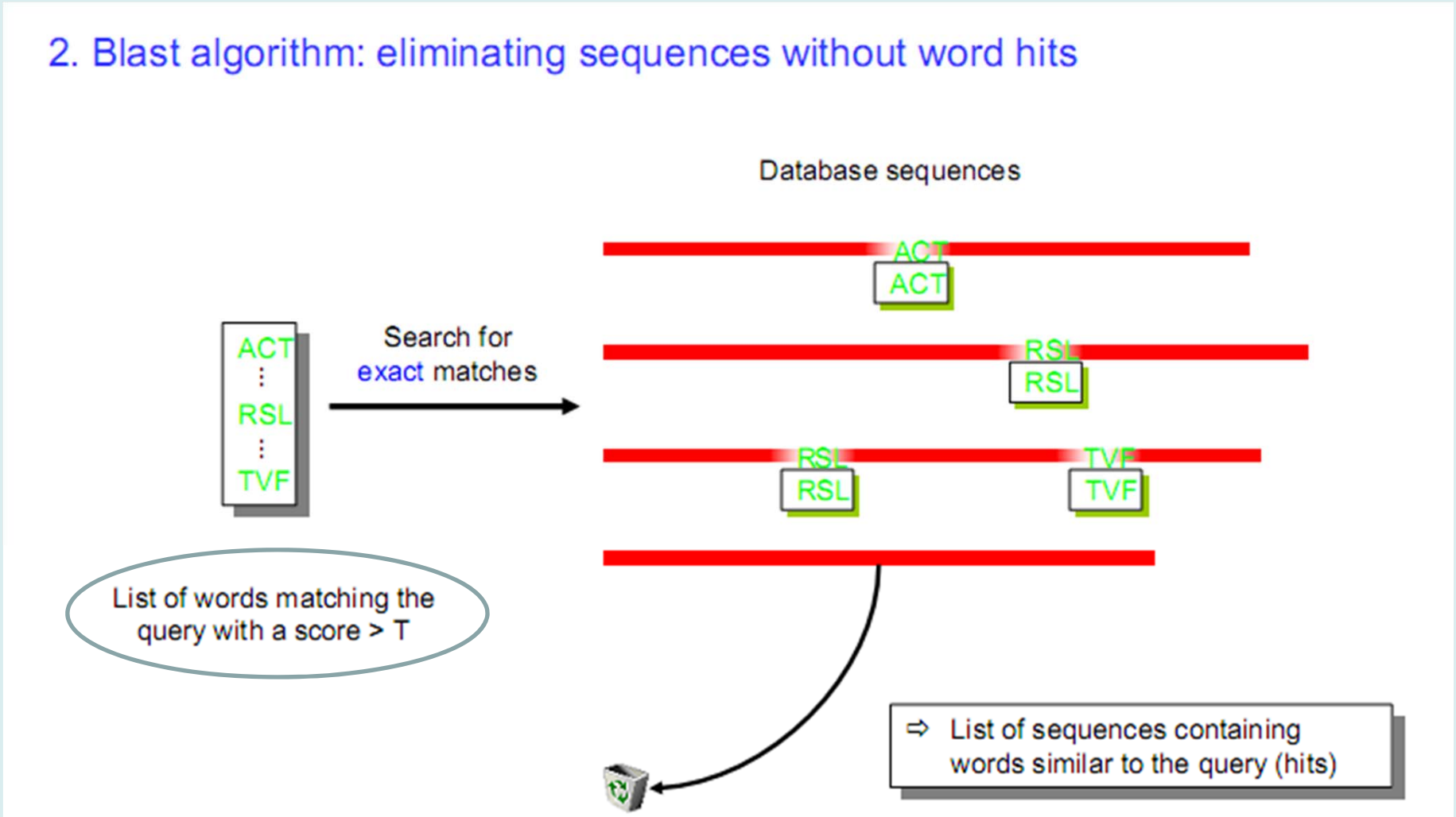
•Advantage: This methods that are least 50-100 times faster than dynamic programming therefore better suited to search databases.

BLAST ALGORITHM (step 1)

1. Blast algorithm: creating a list of similar words

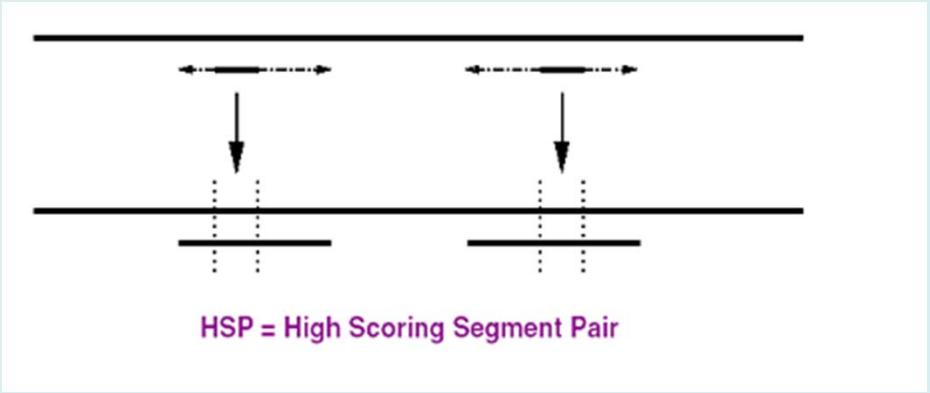


BLAST ALGORITHM (step 2)

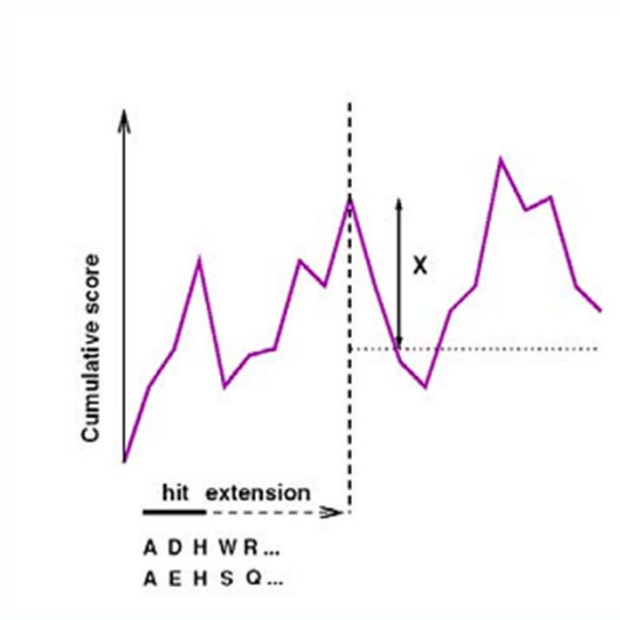


BLAST ALGORITHM (step 3)

3. Blast algorithm: extension of hits
For each word match (“hit”), extended ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.



Each match is then extended. The extension is stopped as soon as the score decreases more then X when compared with the highest value obtained during the extension process.



Reports all HSPs having score S above a threshold, or equivalently, having **E-value** below a threshold

ASSESSING THE SIGNIFICANCE OF SEQUENCE ALIGNMENT

Statistics derived from the scores:

P-value: (0-100%)
Probability that an alignment with this score occurs by chance in a database of this size
The closer the P-value is towards 0, the better the alignment

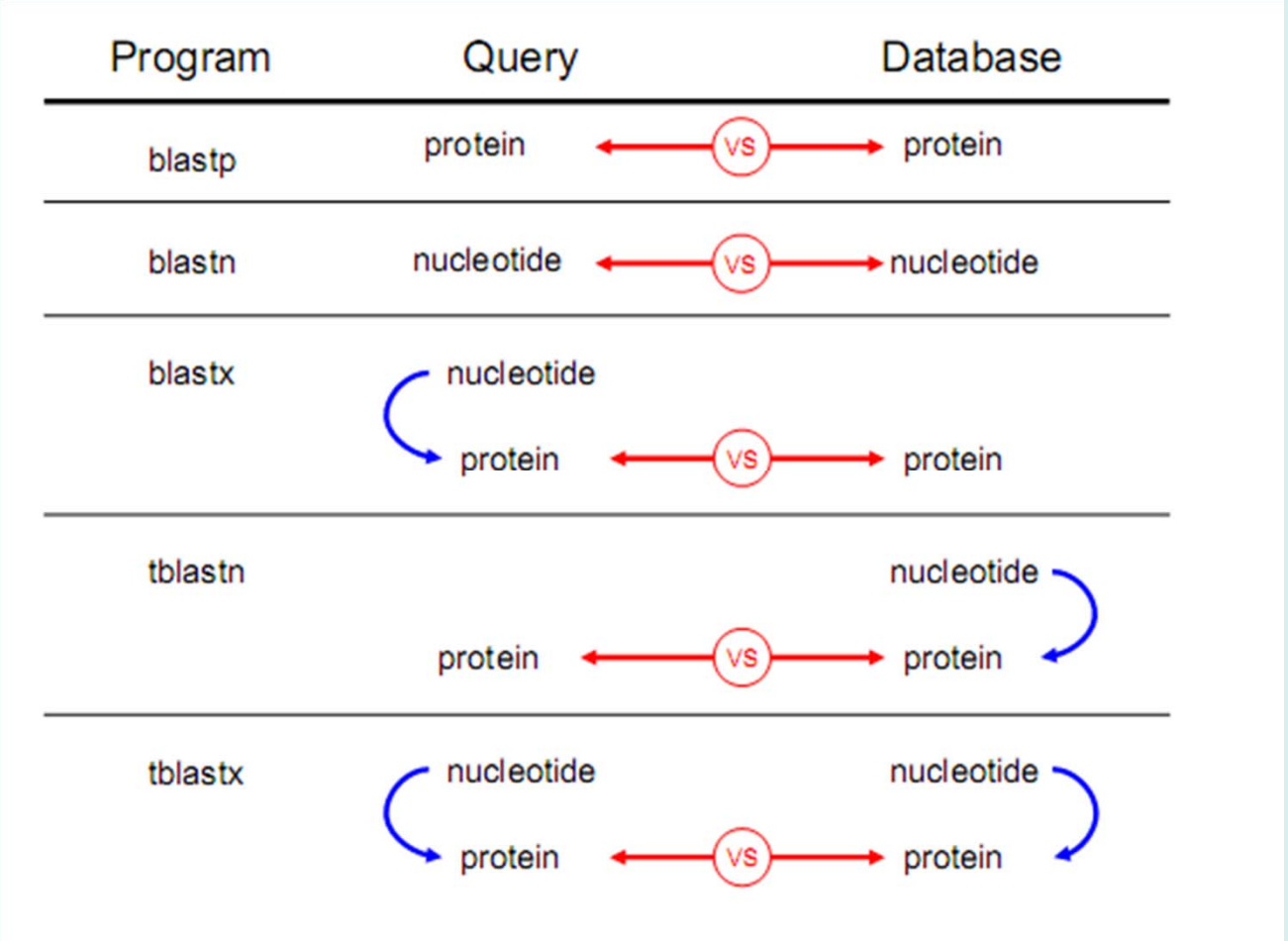
E-value: (0-N)
Number of matches with this score one can expect to find by chance in a database of size N
The closer the E-value is towards 0, the better the alignment

Relationship between **E-value** and **P-value**
In a database containing N sequences

$$E = P \times N$$

A BLAST FOR EACH QUERY

BLAST stands for Basic Local Alignment Search Tool



BLASTing PROTEIN SEQUENCES

blastp = Compares a protein sequence with a protein database

If you want to find something about the function of your protein, use **blastp** to compare your protein with other proteins contained in the databases; identify common regions between proteins, or collect related proteins (phylogenetic analysis);

tblastn = Compares a protein sequence with a nucleotide database

If you want to discover new genes encoding proteins (from multiple organisms), use **tblastn** to compare your protein with DNA sequences translated into their six possible reading frames; map a protein to genomic DNA;

BLASTing DNA SEQUENCES

blastn = Compares a DNA sequence with a DNA database;

Mapping oligonucleotides, cDNAs, and PCR products to a genome; annotating genomic DNA; screening repetitive elements; cross-species sequence exploration;

tblastx = Compares a DNA translated into protein with a DNA database translated into protein;

Cross-species gene prediction at the genome or transcript level (ESTs); searching for genes not yet in protein databases;

blastx = Compares a DNA translated into protein with a protein sequence database;

Finding protein-coding genes in genomic cDNA; determining if a cDNA corresponds to a known protein;

PROGRAMS FOR BLASTing PROTEINS

Program Selection for Protein Queries

Length [1]	Database	Purpose	Program
15 residues or longer	Peptide	Identify the query sequence or find protein sequences similar to the query	Standard Protein BLAST (blastp)
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST
		Find proteins similar to the query around a given pattern	PHI-BLAST
		Find conserved domains in the query	CD-search (RPS-BLAST)
		Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)
5-15 residues	Peptide	Search for peptide motifs	Search for short, nearly exact matches

[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in Section 4 below.

PROGRAMS FOR BLASTing NUCLEOTIDES

Program Selection for Nucleotide Queries

Length [1]	Database	Purpose	Program
20 bp or longer 28 bp or above for megablast	Nucleotide	Identify the query sequence	discontiguous megablast, megablast, or blastn
		Find sequences similar to query sequence	discontiguous megablast or blastn
		Find similar sequence from the Trace archive	Trace megablast, or Trace discontiguous megablast
		Find similar proteins to translated query in a translated database	Translated BLAST (tblastx)
	Peptide	Find similar proteins to translated query in a protein database	Translated BLAST (blastx)
7 - 20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches

[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in the Section 4 below. With default setting, the shortest unambiguous query one can use is 11 for blastn and 28 for MEGABLAST.

ON LINE SERVICES: BLASTing PROTEIN SEQUENCES

Four of the most popular blastp online services:

WU-BLAST2 at EBI (European Bioinformatics Institute)
http://www.ebi.ac.uk/Tools/blast2/

NCBI (National Center for Biotechnology Information):
http://www.ncbi.nlm.nih.gov/BLAST

ExPASy server:
http://www.expasy.org/tools/blast/

Swiss EMBnet server (European Molecular Biology network):
http://www.ch.embnet.org/software/bBLAST.html (basic)
http://www.ch.embnet.org/software/aBLAST.html (advanced)

ON LINE SERVICES: BLASTing DNA SEQUENCES

•BLASTing DNA requires operations similar to BLASTing proteins BUT does not always work so well.

•It is faster and more accurate to BLAST proteins (blastp) rather than nucleotides.

If you know the reading frame in your sequence, you're better off translating the sequence and BLASTing with a protein sequence.

Different BLAST Programs Available for DNA Sequences

Program	*Query*	*Database*	*Usage*
blastn	DNA	DNA	Very similar DNA sequences
tblastx	TDNA	TDNA	Protein discovery and ESTs
blastx	TDNA	Protein	Analysis of the query DNA sequence

T= translated

BLAST SUBSTITUTION MATRICES: SHORT SEQUENCES

In particular, short query sequences can only produce short alignments, and therefore database searches with short queries should use an appropriately tailored matrix.

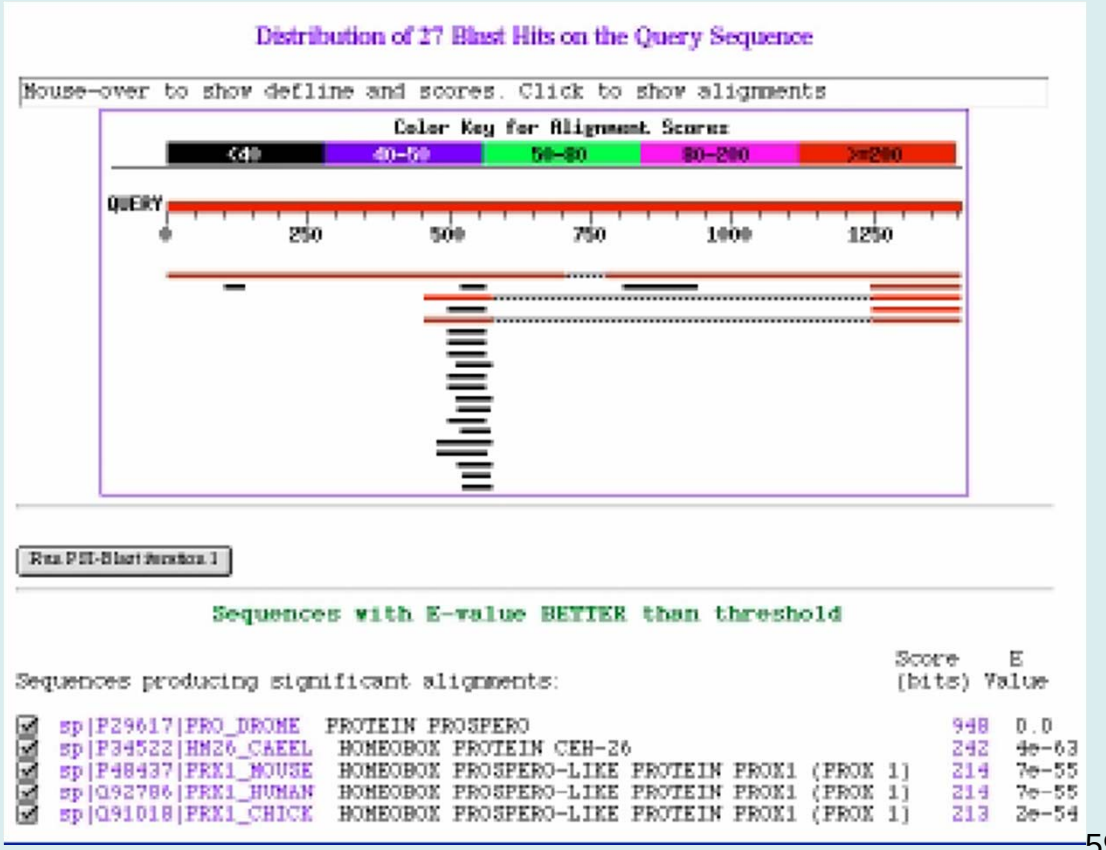
The BLOSUM series does not include any matrices with relative entropies suitable for the shortest queries, so the older PAM matrices may be used instead.

For **proteins**, a provisional table of recommended substitution matrices and gap costs for various query lengths is:

Query Length	Substitution Matrix	Gap Costs
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
85	BLOSUM-62	(10,1)

PSI-BLAST: SEARCH REMOTE HOMOLOGUES

- Position-Specific Iterated BLAST search
- Easy-to-use version of a profile-based search
  - –Perform BLAST search against protein database
  - –Use results to calculate a position-specific scoring matrix
  - –PSSM replaces query for next round of searches
  - –May be iterated until no new significant alignments are found:
    - Convergence
      - –all related sequences deemed found
    - Divergence
      - –query is too broad, make cutoffs more stringent



CONCLUSIONS

Blast: the most used database search tool

- Fast and very reliable even for a heuristic algorithm
- Does not necessarily find the best alignment, but most of the time it finds the best matching sequences in the database
- Easy to use with default parameters
- Solid statistical framework for the evaluation of scores

but...

- The biologist's expertise is still essential to the analysis of the results !

Tips and tricks

- For coding sequences always search at the protein level
- Mask low complexity regions
- Use a substitution matrix adapted to the expected divergence of the searched sequences (nevertheless most of the time BLOSUM62 works well)
- If there are only matches to a limited region of your query, CUT OUT that region and rerun the search with the remaining part of your query