Sequence Alignment

Many types of biological objects can be represented as sequences such as DNA, RNA, or proteins. To study for example, either:

- Similarity between the genome of humans or chimpanzees.
- The causing gene of obesity in mice is the same as in humans.
- Determining the mutations causing sickle cell anemia in 100 patients and 100 normal people.
- Determining the conserved sequences of a protein for the purpose of assessing the degree of similarity and the possibility of homology.

 \rightarrow Preforming an **alignment** makes it easy to compute the similarity between sequences.

<u>Alignment</u>

The process of matching up the nucleotide or amino acid residues of two or more biological sequences to achieve maximal levels of identity.



A gap is the space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another.

 \rightarrow To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.

 \rightarrow So, a good alignment is one with few substitutions and indels.

Types of alignments:

- 1) Depending on the number of sequences:
 - a) Pairwise sequence alignment: the alignment of two biological sequences.
 - b) Multiple sequence alignment: the alignment of three or more biological sequences.
- 2) Depending on the part of sequences aligned:
 - a) Local alignment identify regions of similarity within long sequences that are often widely divergent overall. Most common tool used is BLAST (Basic Local Alignment Search Tool).

 \rightarrow It is useful for example in finding motifs common to two unrelated sequences.

b) Global alignment is a form of global optimization that "forces" the alignment to span the entire length of all query sequences. One of the most common tools used is CLUSTALW.
→ it is useful for example to compare sequences in cases where we have reason to believe that the sequences are related along their entire length to find conserved sequences, or finding mutations in closely related gene or protein sequences and identification of single nucleotide polymorphisms (SNPs).

	conservation conservation
	High Low
HT2A_CRIGR/91-380	GNILVIMAVSLE <mark>K</mark> KLQNATNYFLMSLAIADMLLGFLVMPVSMLTILYGY
DRD2_BOVIN/51-427	GNVLVCMAVSREKALQTTTNYLIVSLAVADLLVATLVMPWVVYLEVVG-
NR2_HUMAN/50-299	ENVAVLYLILSS <mark>H</mark> QLRRKPSYLFIGSLAGADFLASVVFACSFVNFHVF
NR1_HUMAN/133-397	ENLLVLCVILHSRSLRCRPSYHFIGSLAVADLLGSVIFVYSFIDFHVF
1PR1_HUMAN/62-311	ENIFVLLTIWKTKKFHRPMYYFIGNLALSDLLAGVAYTANLLLSGAT
IC3R_MOUSE/55-299	ENILVILAVVRNGNLHSPMYFFLCSLAAADMLVSLSNSLETIMIAVINSDSLTL
A1R_BOVIN/26-288	GNVLVIWAVKVNQALRDATFCFIVSLAVADVAVGALVIPLAILINIGP
CM1_HUMAN/42-418	GNLLVLISFKVNTELKTVNNYFLLSLACADLIIGTFSMNLYTTYLLMG-
22R4 HUMAN/34-329	GNLVAIVVLCKSRKEQKETTFYTLVCGLAVTDLLGTLLVSPVTIATYMKG-
A2R HUMAN/41-308	SNLLALSVLAGARQGGSHTRSSFLTFLCGLVLTDFLGLLVTGTIVVSQHAALF
LF15_MOUSE/41-290	GNLTIILLSRLDARLHTPMYFFLSNLSSLDLAFTTSSVPQMLKNLWGP
10J1 HUMAN/52-300	GNIIIVTIIRMDLHLHTPMYFFLSMLSTSETVYTLVILPRMLSSLVGM
L287_RAT/44-293	GNLAIISLVGAHRCLQTPMYFFLCNLSFLEIWFTTACVPKTLATFAPR
LF1_CHICK/41-290	TNLGLIALISVDLHLQTPMYIFLQNLSFTDAAYSTVITPKMLATFLEE
L867_RAT/22-271	GNLLIILAVSSNSHLHNLMYFFLSNLSFVDICFISTTIPKMLVNIHSQ
SHR_BOVIN/379-626	GNILVLVILITSQYKLTVPRFLMCNLAFADLCIGIYLLLIASVDVHTKTEYHNYAI
PS3_DROME/75-338	GNGLVIWVFSAA <mark>K</mark> SLRTPSNILVINLAFCDFMMMVKTPIFIYNSFHQG-
PS1_DROME/67-329	G <mark>N</mark> GVVIYIFATT <mark>K</mark> SLRTPANLLVINLAISDFGIMITNTPMMGINLYFE
PSD_LOLFO/51-315	GNGVVIYLFTKTKSLQTPANMFIINLAFSDFTFSLVNGFPLMTISCFMK
PSB_HUMAN/51-303	LNAMVLVATLRYKKLRQPLNYILVNVSFGGFLLCIFSVFPVFVASCNG-

Alignment matrices and scoring

Given a pair of aligned sequences, we want to assign a score to the alignment that gives a measure of the relative likelihood that the sequences are related as opposed to being unrelated. We do this by having models that assign a probability to the alignment in each of the two cases, and then we compare these two probabilities.

1) PAM-matrices:

- PAM (point accepted mutation) matrices are based on global alignments of closely related sequences.
- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.
- Other PAM matrices are extrapolated from PAM1. The number with the matrix (e.g. PAM40, PAM100) refers to the evolutionary distance; the greater numbers mean greater distances.

2) BLOSUM matrices:

- **BLOSUM** (BLOcks SUbstitution Matrix) are based on local alignments. It is used to score alignments between evolutionarily **divergent** sequences.
- The number after "BLOSUM" refers to the **minimum percentage identity** of the blocks used to construct the matrix; hence greater numbers mean smaller distances. For example, BLOSUM 62 is a matrix calculated from comparisons of sequences with no less than 62% similarity.
- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins.

→ BLOSUM matrices with higher numbers and PAM matrices with low numbers are both designed for comparisons of closely related sequences.

→ BLOSUM matrices with low numbers and PAM matrices with high numbers are designed for comparisons of distantly related proteins.

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent	≺	→ More divergent

BLAST (Basic Local Alignment Search Tool):

It is an algorithm for comparing biological sequences, such as the amino-acid sequences of different proteins or the DNA sequences. It is one of the most widely used bioinformatics programs, probably because it addresses a fundamental problem emphasizing speed over sensitivity.

There are many different ftypes of BLAST searches:

- 1) **BLASTN** preforms nucleotide-nucleotide sequence comparison.
- 2) **BLASTP** performs protein-protein sequence comparison.
- 3) **BLASTX** searches a nucleotide "query" (refers to the term used in the search) against a protein database, translating the query on the fly.
- 4) **TBLASTN** searches a protein query against a nucleotide database, translating the database on the fly.

The Statistics of BLAST

From a search with BLAST we obtain a series of sequence alignments, each one with a similarity score. However, an important issue is how to know the statistical significance of such a score.

E-value: The Expectation value or Expect value represents the number of different alignments with scores equivalent to.

 \rightarrow The lower the E value, the more significant the score and the alignment.

Note:

ο

In case of studying sequence homology using alignment tools, most biochemists consider 25% identity the cutoff for sequence homology, meaning that if two proteins are less than 25% identical in sequence, more evidence is needed to determine whether they are homologs.

Tools that could be used are:

• BLAST, by aligning two or more sequences rather than searching in the NCBI databases.

Enter Subject Sequence					
Enter accession number(s), gi(s), or FASTA sequence(s) 😯	Clear	Subject subrange 😯			
		From			
		То			
MBOSS Needle, <u>https://www.ebi.ac.uk/Tools/psa/emboss_needle/</u>					