Sequence Alignment

10/03/2017

Bioinformatics Data

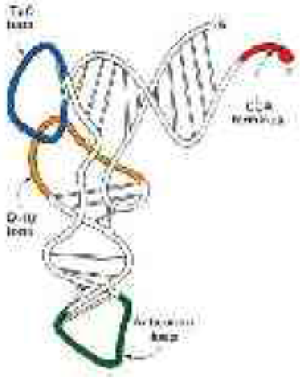
Sequences of nucleotides

>gi|8886401|gb|AF162269.1|
CCCACTCCTCCATCTCACAAACACTTCTCTATACCCAACAATCCCTTTTACAATCCCTGCTCATTTAGTC
AAAATGGTCAAGATTGCTGCTATCATCCTCCTCATGGGCATTCTCGCCAATGCTGCCGCCATCCCTGTCA
TTTCAACACCCAAATTACAGAGCCAACCGGCGAGGGCGACCGTGGGGACGTGGCCGAC

Sequences of amino acids

>P25032
MASSSSATSGDDRPPAAGGGTPAQAHAEWAASMHAYYAAAASAAGHPYAAWPLPPQAQQHGLVAAGAGAAYG
AGAVPHVPPPPAGTRHAHASMAAGVPYMA

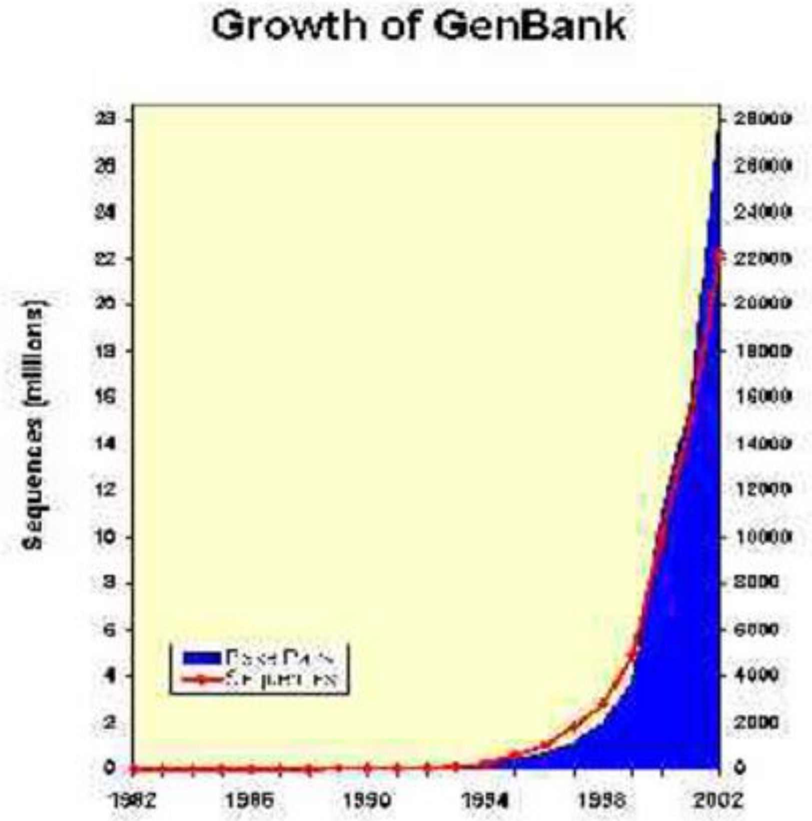
Structure of macro-molecules



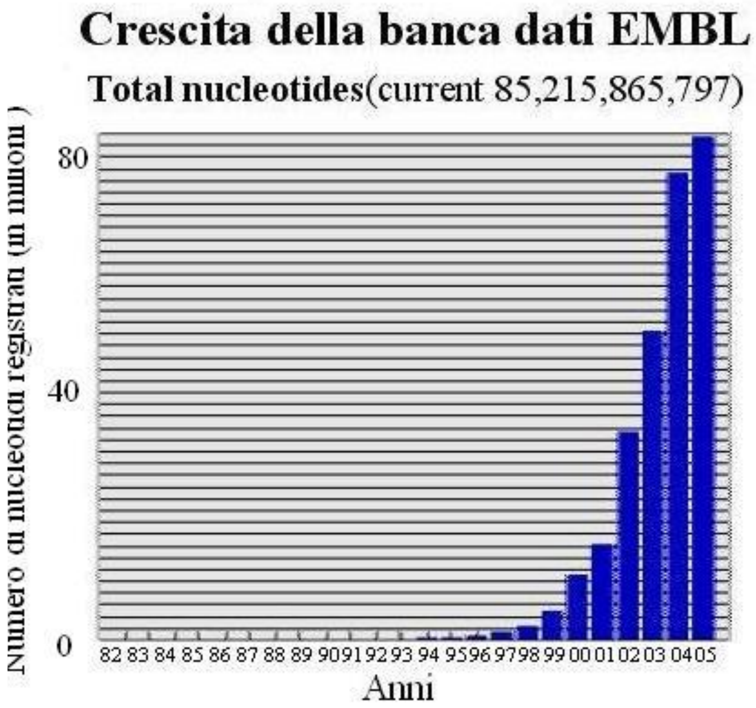
Bioinformatics aims

- Management of biological data
  - Storage, organization, distribution, etc.
- Analysis of biological data
  - Inferences and prediction on their biological meaning

Why Bioinformatics is important



Growth of GenBank



Crescita della banca dati EMBL

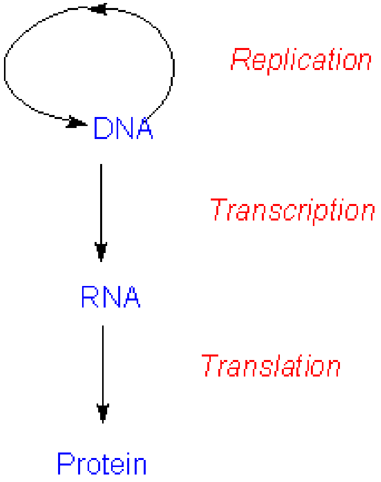
Total nucleotides(current 85,215,865,797)

We have sequenced the complete genome of many species!

Central Dogma

- **Transcription and translation are the two main processes linking gene to protein**
- Genes provide the instructions for making specific proteins.
- The bridge between DNA and protein synthesis is RNA.
- RNA is chemically similar to DNA, except that it contains ribose as its sugar and substitutes the nitrogenous base uracil for thymine.
  - An RNA molecules almost always consists of a single strand.

DNA→RNA→Protein



Replication

DNA

Transcription

RNA

Translation

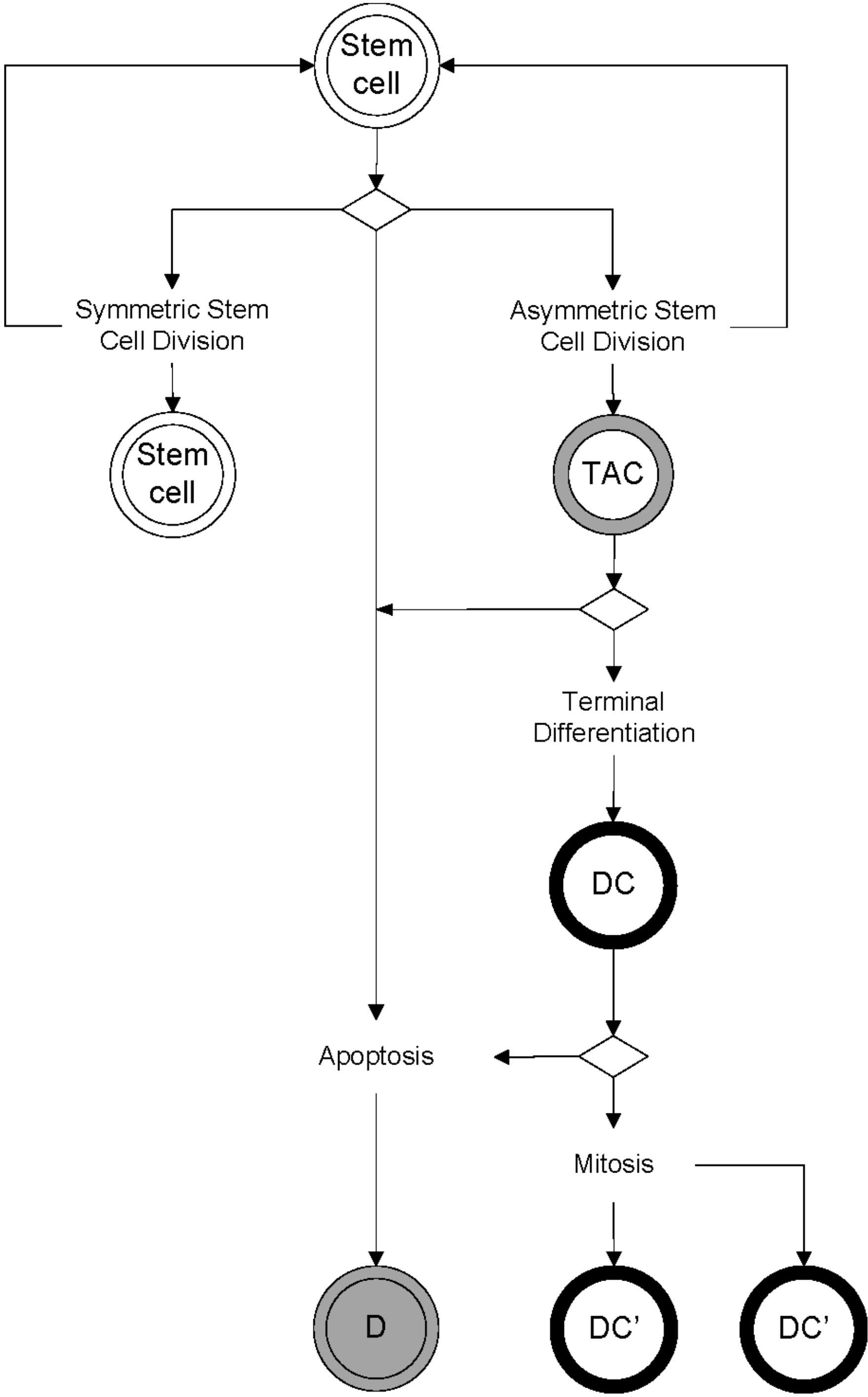
Protein

- DNA is **TRANSCRIBED** to messenger RNA (mRNA)
- mRNA carries the message to tranfer RNA (tRNA)
- tRNA is **TRANSLATED** to an amino acid chain, which makes up proteins

- In DNA or RNA, the four nucleotide monomers act like the letters of the alphabet to communicate information.
- The specific sequence of hundreds or thousands of nucleotides in each gene carries the information for the primary structure of a protein (the linear order of the 20 possible amino acids)
- To get from DNA, written in one chemical language, protein, written in another language, requires two major stages, transcription and translation.

Mutations

- Several “copy” actions of DNA: mitosis, stem asymmetric division (cell specialization), stem symmetric division.



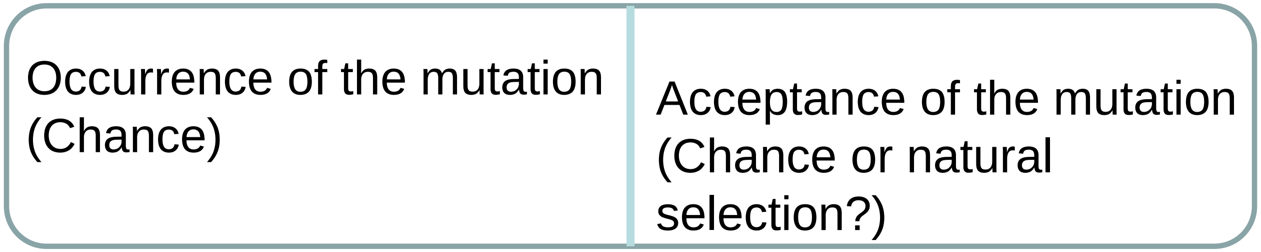
Mutations

- **Mutations:** alterations of the coded information in DNA
- A mutation can occur due to
  - **Substitutions:** change of a single base
  - **Insertions:** additions of nucleotides
  - **Deletions:** removal of nucleotides

Mutations

- ***substitutions***
  - ***synonymous:*** *do not change the amino acid*
  - ***meaning:*** *change an amino acid in a different one*
  - ***non-sense:*** *change an amino acid in a stop codon*
- ***Insertions / deletions***
  - *Framed kept reading* ***(multiples of three)***
  - ***Frameshift:*** A **frameshift mutation** (also called a **framing error** or a **reading frame shift**) is a genetic mutation caused by indels (insertions or deletions) of a number of nucleotides in a DNA sequence that is not divisible by three.

Occurrence and outcome of mutations

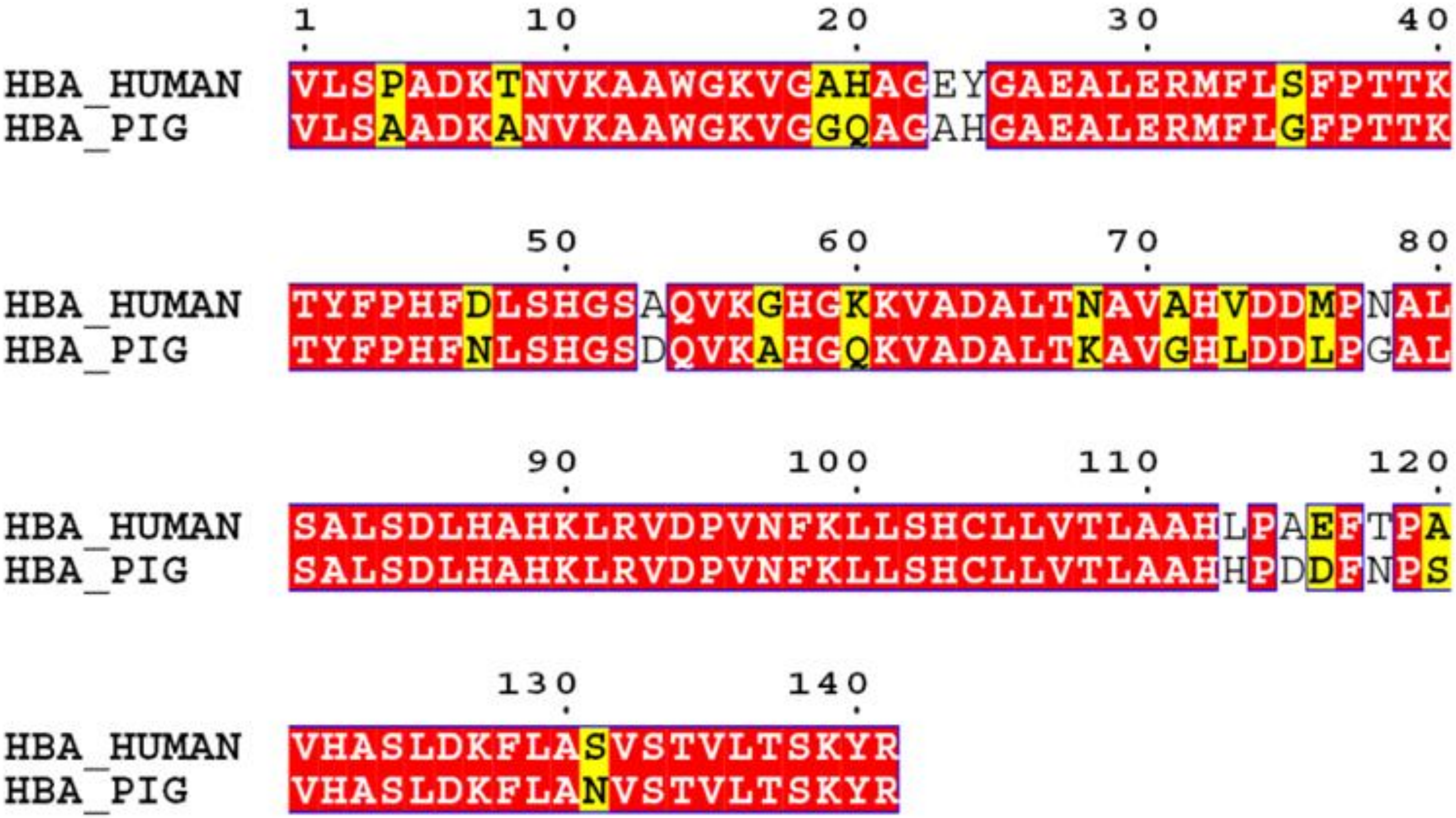


Occurrence of the mutation (Chance)

Acceptance of the mutation (Chance or natural selection?)

A mutations can be:

- Neutral
- Advantageous
- Disadvantageous



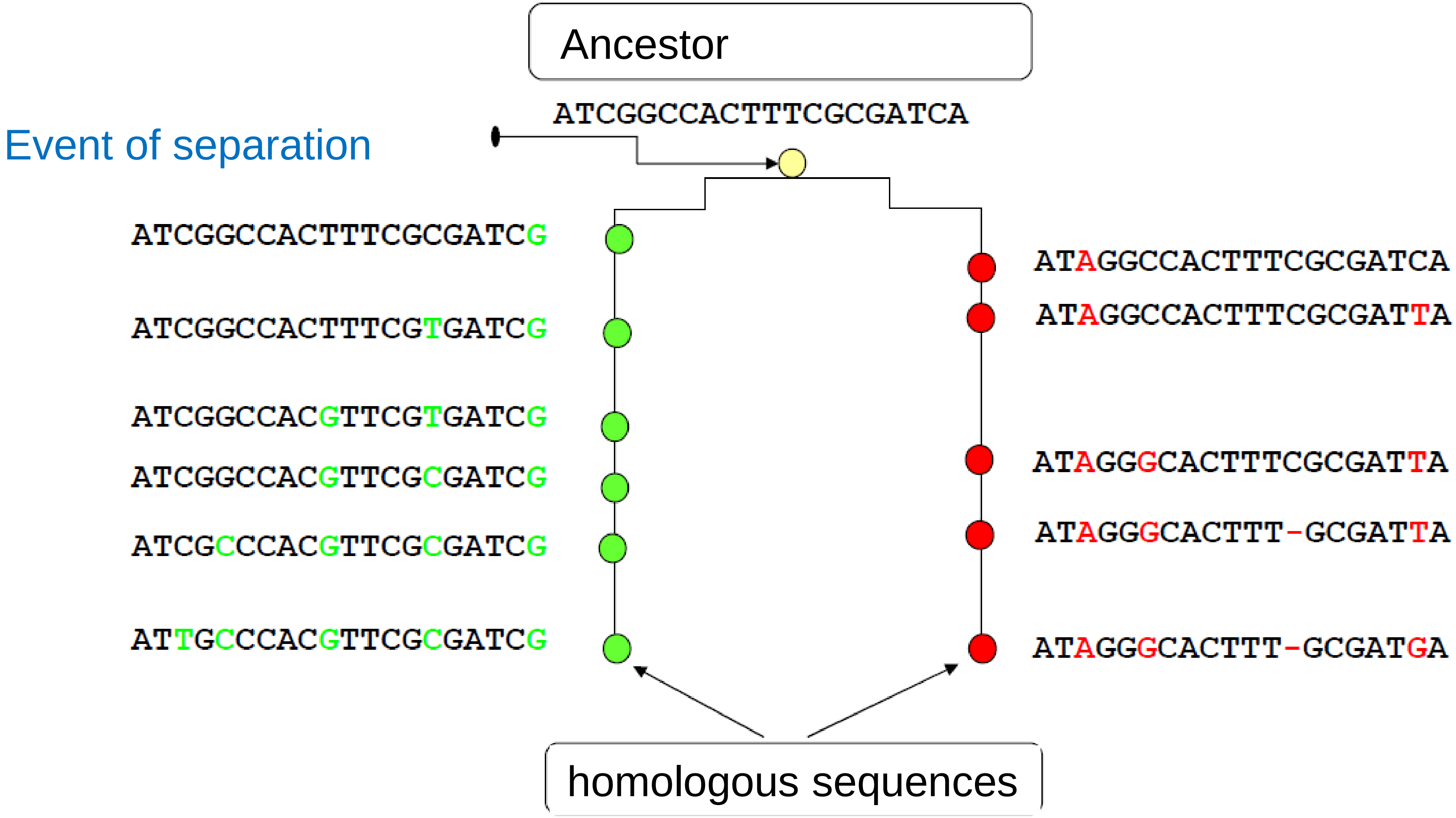
```
                  1        10        20        30        40
                  .         .         .         .         .
HBA_HUMAN   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTK
HBA_PIG     VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTK

                          50        60        70        80
                           .         .         .         .
HBA_HUMAN   TYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNAL
HBA_PIG     TYFPHFNLSHGSDQVKAHGQKVADALTKAVGHLDDLPGAL

                          90       100       110       120
                           .         .         .         .
HBA_HUMAN   SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
HBA_PIG     SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPDDFNPS

                         130       140
                           .         .
HBA_HUMAN   VHASLDKFLASVSTVLTSKYR
HBA_PIG     VHASLDKFLANVSTVLTSKYR
```

Evolution


Ancestor

ATCGGCCACTTTCGCGATCA

Event of separation

ATCGGCCACTTTCGCGATCG

ATCGGCCACTTTCGTGATCG

ATCGGCCACGTTCGTGATCG

ATCGGCCACGTTCGCGATCG

ATCGCCCACGTTCGCGATCG

ATTGCCCACGTTCGCGATCG

ATAGGCCACTTTCGCGATCA

ATAGGCCACTTTCGCGATTA

ATAGGGCACTTTCGCGATTA

ATAGGGCACTTT-GCGATTA

ATAGGGCACTTT-GCGATGA

homologous sequences

homology: sharing a common ancestor

Analisys of biological data

- The main approach of data analysis in Bioinformatics is based on **comparison**.
- Example: in case of an epidemic, the sequenced data coming from infected organisms are compared with the data in data banks:
  - To find similarities with what is known
  - To find the ancestor
  - To guess the source of the epidemic and how it evolved

We talk about **Alignment**

Alignment

- The alignment of the sequences is a procedure in which two or more primary amino acid (proteins), or nucleotides (DNA or RNA) sequences are compared and aligned.
- The alignment allows to identify identical or similar regions that may have functional, structural, or phylogenics relationships.
- In general, the alignment is used to verify whether a sequence of interest is already present within a database of known sequences, or if a similar one already exists.

Alignment

- In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.
- Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix.
- Sequence alignments are also used for non-biological sequences, such as calculating the edit distance cost between strings in a natural language or in financial data.

Alignment

- Some alignment uses gap to maximize the similarities.
- Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.
- If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations)

Alignment

- It can be **single** (also called pairwise) or **multiple**, **local** or **global**.
- It **finds the correspondence** between the characters of the input sequences.
- It can be applied to sequences of
  - nucleotides (alphabet of 4 symbols: A-T-G-C in the case of DNA, and A-U-G-C in the case of RNA)
  - amino acids (alphabet of 20 symbols)

Alignment

- The aim is to assess the conservation and the variation of the residues from the moment in which the sequences have diverged from the common ancestor sequence -> **identify identical or similar regions that have phylogenetic relationships.**
- With the “**residue**” term denotes a single character of a biological sequence

Alignment

- In case of proteins, the degree of **similarity** between amino acids **occupying a particular position** in the sequence can be interpreted as a rough **measure of how conserved** a particular region or sequence **motif is among lineages**.
- The **absence of substitutions, or the presence of only very conservative substitutions** (that is, the substitution of amino acids whose side chains have similar biochemical properties) in a particular region of the sequence, suggest that this **region has structural or functional importance**.

Alignment

- Computational approaches to sequence alignment generally fall into two categories:
  - **global** alignments: Calculating a global alignment is a form of global optimization that "forces" the alignment to span the entire length of all query sequences.
  - **local** alignments: identify regions of similarity within long sequences that are often widely divergent overall.
- Local alignments are often preferable, but can be more difficult to calculate because of the additional challenge of identifying the regions of similarity.

Alignment

- A variety of computational algorithms have been applied to the sequence alignment problem.
  - These include slow but formally correct methods like dynamic programming.
  - These also include efficient, heuristic algorithms or probabilistic methods designed for large-scale database search, that do not guarantee to find best matches.

SINGLE OR PAIRWISE ALIGNMENT

Single (or pairwise) alignment

- From a computer science/engineering point of view, a protein is a more or less long string of an alphabet of four/twenty letters
- ALIGNING TWO OR MORE PROTEINS MEANS TO FIND THE BEST WAY TO MATCH THEM, LOOKING FOR PATTERN AND COMMON MOTIFs, or in other words, to locate the evolutionary reason why they are alike and try to understand the common structural features
- How it works: it tries to find a way to convert the input sequences, one to another, by inserting the smallest possible number of changes.

Problem Definition

Let us consider two sequences $S = s_1s_2\ldots s_n$ and $T = t_1t_2\ldots t_m$ built on the alphabet Σ, a global alignment of S and T consists in a pair of sequences $S' = s'_1s'_2\ldots s'_l$ and $T' = t'_1t'_2\ldots t'_l$ on the alphabet $\Sigma \cup \{-\}$ (where – represents a gap), satisfying the following properties:

$|S'| = |T'| = l \quad (\max(n,m) \leq l \leq (m+n))$

Removing gaps from S' we obtain S

Removing gaps from T' we obtain T

If $s'_i = -$, then $t'_i \neq -$ and viceversa

Example

```
S    ttcgagccttagcgta

T    ttatagcgtagtcgta


S'   ttc-gagccttag-cgta

T'   ttat-agcg-tagtcgta
```

- **MATCH:** positive similarity
- **MISMATCH:** negative similarity
- **GAP:** negative similarity introduced to allow a better global alignment

Which alignment is the best one?

To determine the best alignment we use a score function that assigns a value to each alignment.

The score function depends on the some assumptions.

It depends on the substitution matrix that assigns a score to each mismatch or gap

Example of score

No gap: alignment score 10 (with a score function that assign 1 to each match)

```
IPLMTRWDQEQESDFGHKLPIYTREWCTRG
CHKIPLMTRWDQQESDFGHKLPVIYTREW
```

With gap: alignment score 25 (with a score function that assign 1 to each match)

```
   IPLMTRWDQEQESDFGHKLP-IYTREWCTRG
CHKIPLMTRWDQ-QESDFGHKLPVIYTREW
```

Single Alignment: **Substitution matrices**

- In the case of proteins, a substitution matrix is a grid of 20 x 20 in which for each position is given a score relative to the specific pair of considered amino acids.

- At the position (a,b) there is the score of the substitution of amino acid a to the amino acid b.

- The defined score comes from observations on the similarities in terms of the following criteria:

  - phylogenetic
  - Structural
  - Statistical

Single Alignment: **Substitution matrices**

A substitution matrix is a matrix that assigns to each pair of characters $(a,b)\in(\Sigma\cup\{-\})^2$ a score d that expresses the cost (or benefit) of the replacement of the symbol a by the symbol b.

Score A of the alignment S’, T’

$$A = \sum_{i=1}^{l} d(S'(i), T'(i))$$

A Matrix represents the score of the substitution of two characters. When d(a,b) = d(b,a) the matrix is triangular.

Single Alignment:
Substitution matrices

Identity Matrix

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

BLAST Matrics

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

Transition
Transversion matrix

	A	T	C	G
A	1	-5	-5	-1
T	-5	1	-1	-5
C	-5	-1	1	-5
G	-1	-5	-5	1

For nucleotides.

Single Alignment: **Substitution matrices**

Example of matrix:

- $d(x,x) = 1$, $d(-,x) = d(x,-) = -a$, $d(x,y) = -u$
  - Se $a = 0$, $u = \infty \Rightarrow$ LCS (*Longest Common Subsequence*)
- PAM or BLOSUM for proteins

Single Alignment: **Substitution matrices**

Matrice PAM [1] **(P**oint **A**ccepted **M**utation **o P**ercent **A**ccepted **M**utation **)**

They were introduced by Margaret Dayhoff in 1978 based on the observation of 1572 mutations in 71 families of closely related proteins among them. It was observed that:

- some substitutions were not random and
- some occurred more easily than others.

The **purpose of a PAM matrix** is:
Given two related sequences, calculate the probability of each amino acid can undergo a mutation and relate it to the evolutionary distance.

[1] Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. (1978). "**A model of Evolutionary Change in Proteins**". *Atlas of protein sequence and structure* (volume 5, supplement 3 ed.). Nat. Biomed. Res. Found.. pp. 345–358.

PAM (Point Accepted Mutations) Matrix

- The calculation of these matrixes is based on an evolutionary model based on memoryless property.

- Underlying assumption: the probability that in a certain site a certain mutation occurs is independent of previous mutational events at the same site (**Markov process**):

  ? → A → B

- **This probability does not depend on the position in the sequence of the site taken into account.**

PAM (Point Accepted Mutations) Matrix

- PAM1:
  - It corresponds to evolutionary distance 1 (1% of mutations)
- PAMn:
  - It corresponds to evolutionary distance n, PAM1^n

PAM 1 Calculation

- p_a, probability of occurrence of the amino acid a in a fairly wide range of protein sequences
- $f_{a,b} = f_{b,a}$ number of accepted mutation a ⇔ b
- $f_a = \Sigma f_{a,b}$
- $f = \Sigma f_a$
- $M_{a,b}$ = probability of a is substituted by b
- $M_{a,a}$ is calculated from the relative mutability of a:
  - $m_a = f_a/100\ f \cdot p_a$
  - $M_{a,a} = 1 - m_a$
- $M_{a,b} = f_{a,b} * m_a/f_a$

PAM (Point Accepted Mutations) Matrix

- We use PAM units to measure the distance between amino acid sequences;
- S1 and S2 sequences are 1 PAM unit if S1 can be transformed into S2 with an average of 1 point mutation per 100 amino acids;
- when k = 250 we get one of the most used matrices: PAM250

PAM250 to align sequences with approximately 14-27% similarity
PAM120 to align sequences with approximately 40% similarity
PAM80 to align sequences with approximately 50% similarity
PAM60 to align sequences with approximately 60% similarity

[1] Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. (1978). "**A model of Evolutionary Change in Proteins**". *Atlas of protein sequence and structure* (volume 5, supplement 3 ed.). Nat. Biomed. Res. Found.. pp. 345–358.

PAM (Point Accepted Mutations) Matrix

Calculation of PAM2

j

i

	A	B	C
A	A -> A	A -> B	A -> C
B	B -> A	B -> B	B -> C
C	C -> A	C -> B	C -> C

x

j

	A	B	C
A	A -> A	A -> B	A -> C
B	B -> A	B -> B	B -> C
C	C -> A	C -> B	C -> C

i

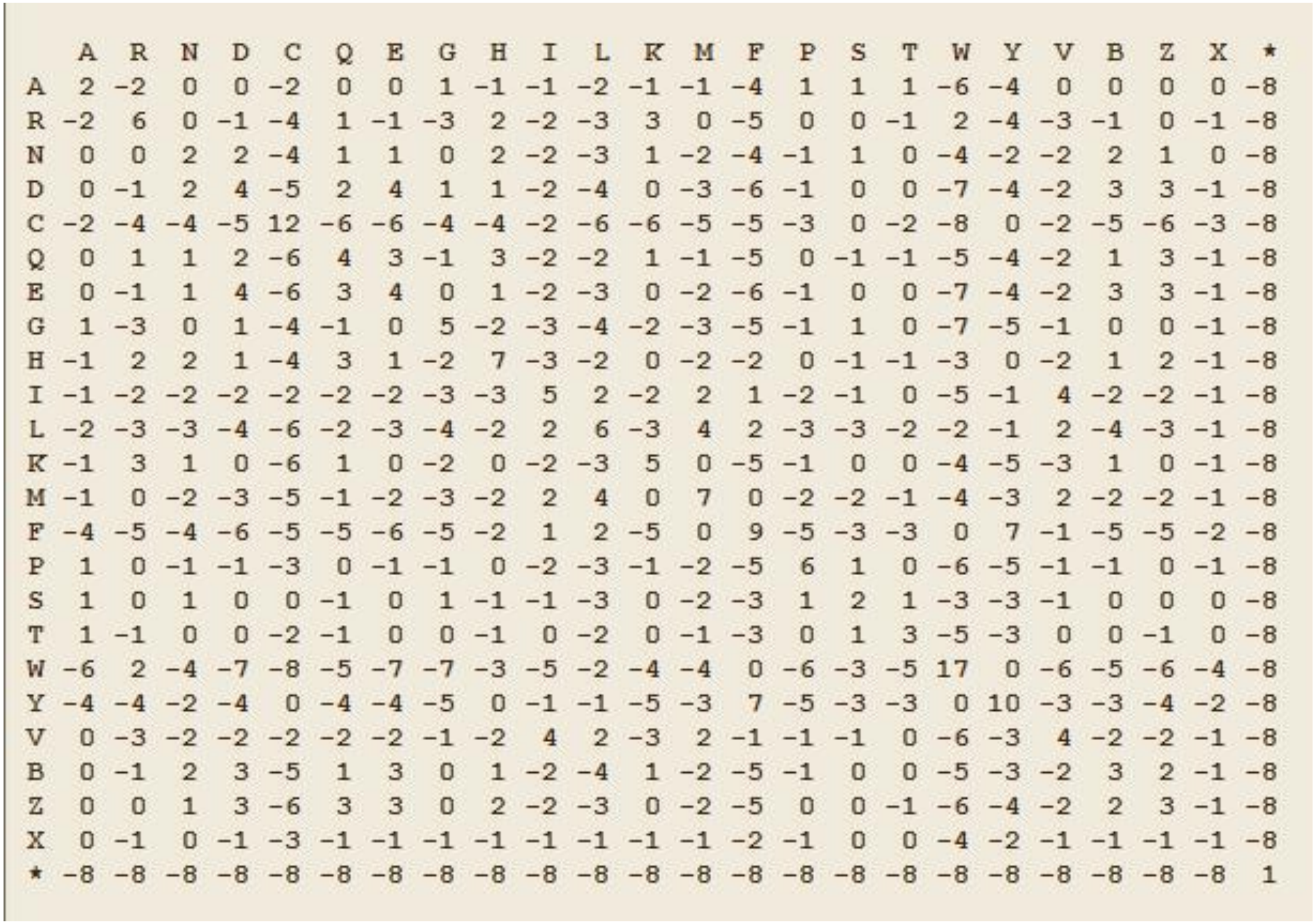
j

i

	A	B	C
A		(A -> A * A -> B)+ (A -> B * B -> B)+ (A -> C * C -> B).	
B			
C			

A → B =
(A → A → B)+(A → B → B)+ (A → C → B)

PAM 240



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

SEE YOU ON MARCH 16TH