H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics online course: IBT

Module: Sequence Alignment Theory and Applications

Session: Introduction to Searching and Sequence Alignment

Learning Objectives

- Understand the applications of sequence similarity searching and alignment
- Understand the concepts of homology, identity, orthologues, paralogues
- Sequence evolution: introducing concepts of point mutations, deletions, insertions etc.
- Introduction to pair-wise sequence alignment
- Overview of the different approaches to sequence alignment - exhaustive vs. heuristic

Learning Outcomes

- Understand the concept of sequence alignment
- Understanding the concepts of sequence evolution- mutations, deletions, insertions etc
- Sequence diversity: understand the difference between homologues, paralogues and orthologues
- Analyzing similarity and differences between sequences
- Understanding why and how to find sequences similar to the one of your interest

What is Sequence Alignment

- **Sequence alignment** is a way of arranging two or more sequences (DNA, RNA, or A.a.(proteins)) to identify regions of similar character patterns
- Sequence similarity could be a result of **functional**, **structural**, or **evolutionary** relationships between the sequences
- Procedure involves searching for series of identical or similar characters/patterns in the same order between the sequences
- Non identical characters aligned as mismatches or opposite a gap in the other sequence
- Alignment made between a known sequence and unknown sequence or between two unknown sequences

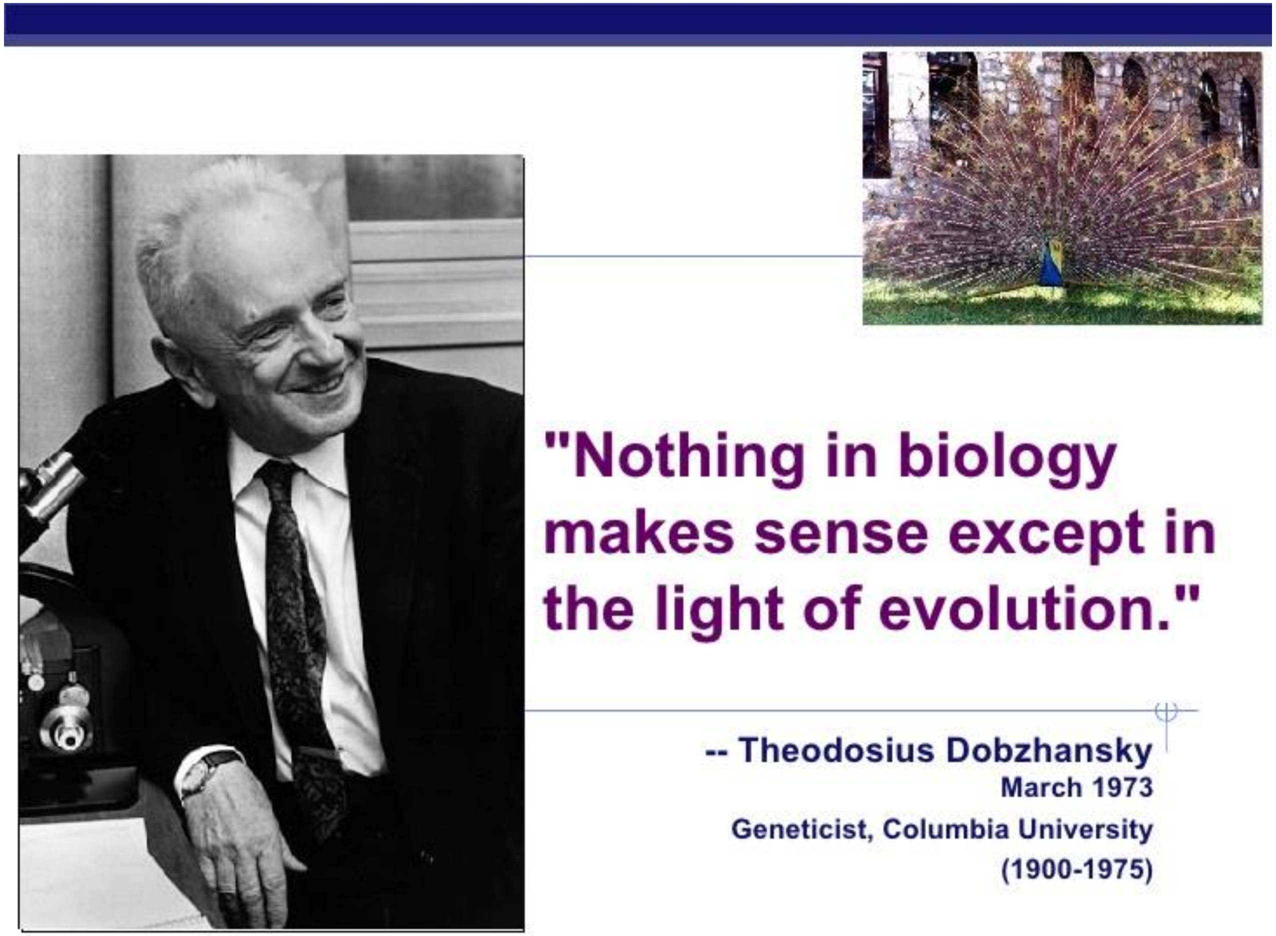
Why Sequence Alignment (uses)?:

- Useful in DNA and Protein sequences for:
  - Discovering **functional** information
  - Predicting molecular **structure**
  - Discovering **evolutionary** relationships

- Sequences that are very much alike probably have:
  - Same function
  - Similar secondary and 3-D structure (if proteins)
  - Shared ancestral sequence (**though not always**)

- Sequence alignment enables the following:
  - Annotation of new sequences
  - Modeling of protein structures
  - Phylogenetic analysis

Sequence Evolution



"Nothing in biology makes sense except in the light of evolution."

-- Theodosius Dobzhansky
March 1973
Geneticist, Columbia University
(1900-1975)

Evolutionary basis of Sequence Alignment

- One goal of sequence alignment is to enable inference of homology (origin from common ancestor) through observed shared sequence similarity.

- Changes that occur during sequence divergence from common ancestor include:
  - Substitutions
  - Deletions
  - Insertions

Sequence Relationships-1

- **Identity/ Similarity:**
  - **Sequence Identity:** Exactly the same Amino acid or nucleotide in the same position
  - **Sequence Similarity:** Content includes substitutions (A.a residues) with similar chemical properties
  - **Similarity:** A quantifiable property- Two sequences are similar if order of sequence characters is recognizably the same and they can be aligned

Sequence Relationships-2

- How similar is **very similar?:**

Sequences be at least 100 A.a or 100 nucleotides long, then:

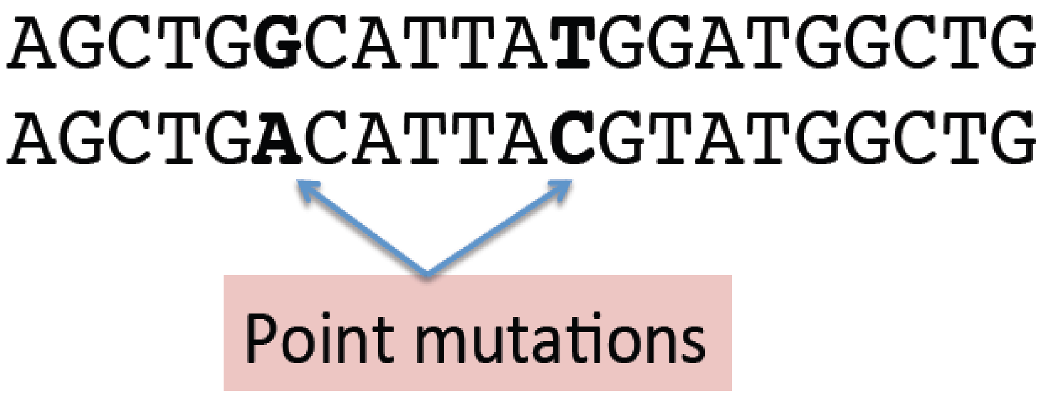
  - **25% Amino acid identity** required to call protein homology
  - **70% nucleotide identity** required to call gene homology

- **Caution**: Homology or non-homology is more than just sequence similarity

Sequence Relationships-3

- To ascertain homology , also consider other information reported by the sequence comparison/search:
  - Expectation Value (E-value, see local alignments later): tells how likely observed similarity is due to chance
  - Length of segments similar between the two sequences
  - The patterns of A.a. conservation
  - The number of insertions and deletions

Similarity/Identity: Nucleotides



AGCTG**G**CATTA**T**GGATGGCTG

AGCTG**A**CATTA**C**GTATGGCTG

Point mutations

90% identity

90% similarity

Sequence similarity and sequence identity are synonymous for nucleotide sequences

Credit Pandam Salifu , IBT 2016

% Similarity/Identity: Nucleotides-1

Equal Length:

- Two sequences of equal length, percentage of similarity S or identity I

$$= [2L/(L_y + L_z)] \text{ X } 100$$

Where

L is the number of aligned residues with similar or identical characteristics

L_y is the total length of sequence y

L_z is the total length of sequence z

Similarity/Identity: Nucleotides-2

Un equal Length:

- Two sequences of unequal length, percentage of similarity S or identity I

$$I(S) = (L_{i(s)}/L_y)100$$

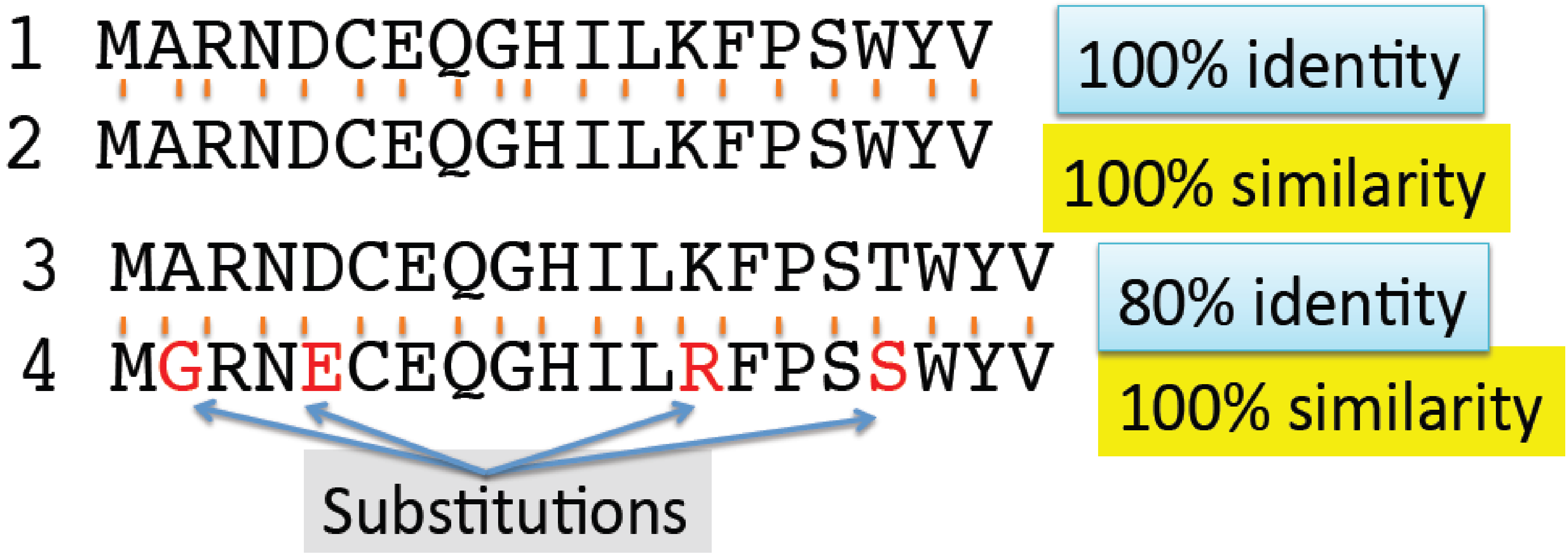
Where

$L_{i(s)}$ is the number of aligned residues with similar or identical characteristics

L_y is the length of the shorter of the two sequences

Similarity/Identity: Amino Acids-1

- % Identity and similarity not synonymous for Amino acid sequences:



```
1  MARNDCEQGHILKFPSWYV
2  MARNDCEQGHILKFPSWYV

3  MARNDCEQGHILKFPSTWYV
4  MGRNECEQGHILRFPSSWYV
```

100% identity

100% similarity

80% identity

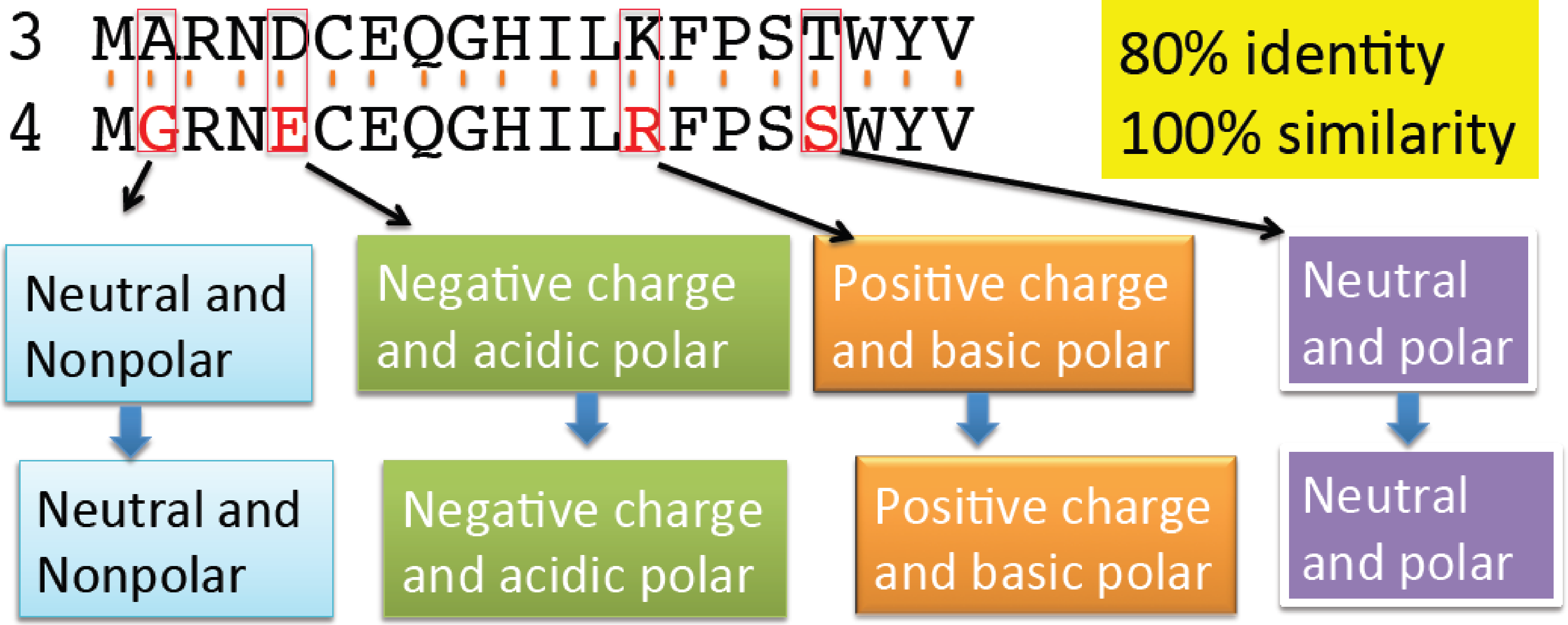
100% similarity

Substitutions

Credit Pandam Salifu , IBT 2016

Similarity/Identity: Amino Acids-2

❖% Identity and similarity not synonymous for Amino acid sequences:

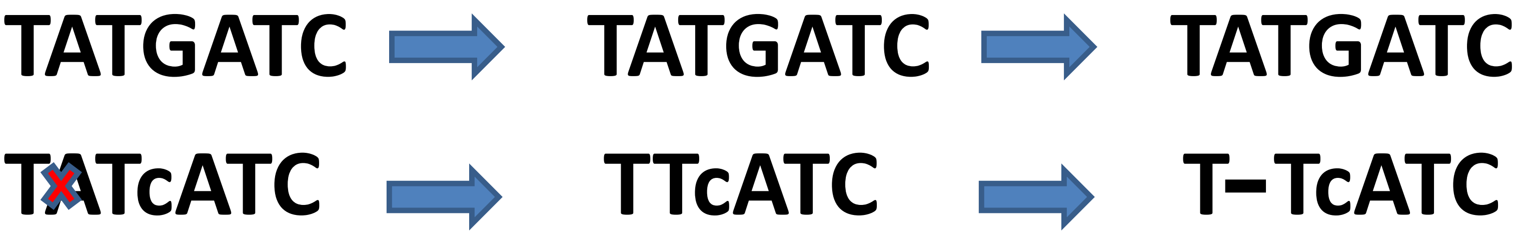

3 MARNDCEQGHILKFPSTWYV
4 MGRNECEQGHILRFPSSWYV
Neutral and Nonpolar
Negative charge and acidic polar
Positive charge and basic polar
Neutral and polar
Neutral and Nonpolar
Negative charge and acidic polar
Positive charge and basic polar
Neutral and polar

80% identity
100% similarity

Credit Pandam Salifu , IBT 2016

Sequence Relationships-2a

- **Homology:**
  - **Homologous sequences** (related by descent): Two or more sequences, readily aligned ,i.e. very similar such that they have a shared ancestry
  - **Homologous positions**



TATGATC → TATGATC → TATGATC

TATcATC → TTcATC → T–TcATC

Sequence Relationships-2b

Similarity Vs Homology

- Similarity means likeness or % identity between two sequences

- Similarity means having statistically significant number of Amino acids or nucleotide base matches

- Similarity does not imply homology

- Homology refers to shared ancestry

- Two sequences are homologous if derived from a common ancestral sequence

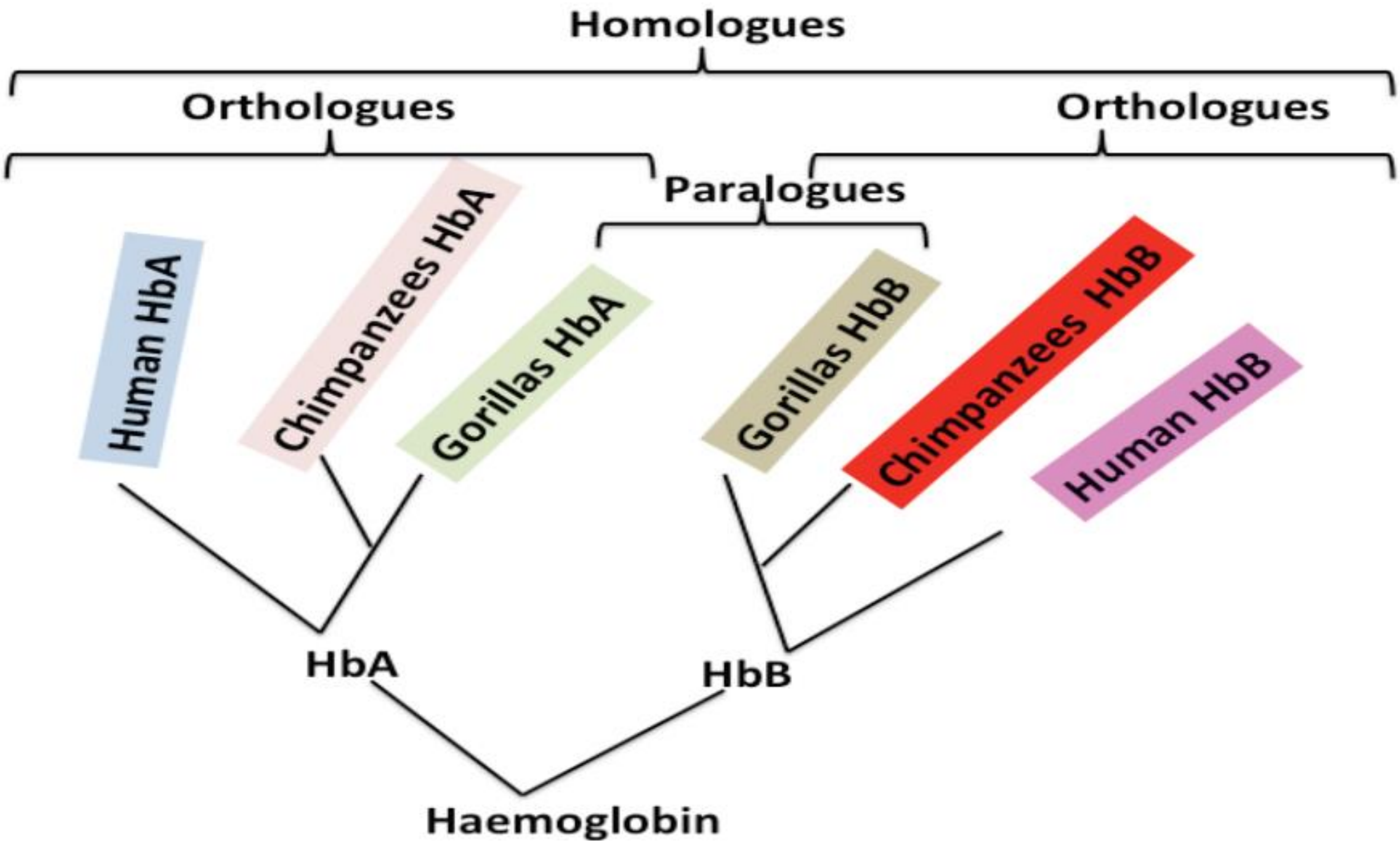
- Homology usually implies similarity

- sequences are either homologous or not, so no % homology

Sequence Relationships-2c

- **Orthologous sequences:** quite similar sequences found in different species (i.e. due to a speciation event), and carrying out a similar biological function
- **Paralogous sequences:** Sequences related through gene duplication events. Can have variable biological function within a species
- Sequences may be both orthologous and paralogous
- Orthology is a form of homology

Sequence Relationships-2d



Credit Pandam Salifu , IBT 2016

Sequence Alignment Example- Homology

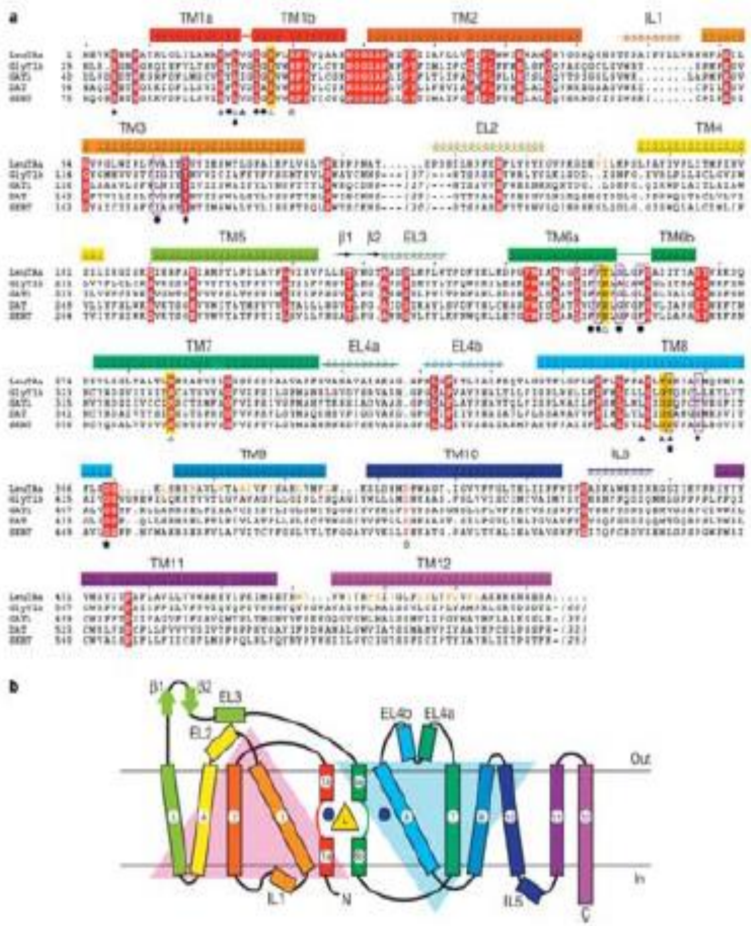
```
               10        20        30        40        50        60
HUMAN  MNPLLILTFVAAALAAPFDDDDKIVGGYNCEENSVPYQVSLNSGYHFCGGSLINEQWVVS
       :. ::::..:.::.: :..:::::::::.: :.:::::::::::::::::::::.:::::
RAT    MSALLILALVGAAVAFPLEDDDKIVGGYTCPEHSVPYQVSLNSGYHFCGGSLINDQWVVS
               10        20        30        40        50        60
               70        80        90       100       110       120
HUMAN  AGHCYKSRIQVRLGEHNIEVLEGNEQFINAAKIIRHPQYDRKTLNNDIMLIKLSSRAVIN
       :.::::::::::::::::.::::.::::::::::.::.:.  ::::::::::::: . .:
RAT    AAHCYKSRIQVRLGEHNINVLEGDEQFINAAKIIKHPNYSSWTLNNDIMLIKLSSPVKLN
               70        80        90       100       110       120
              130       140       150       160       170       180
HUMAN  ARVSTISLPTAPPATGTKCLISGWGNTASSGADYPDELQCLDAPVLSQAKCEASYPGKIT
       :::. ..::.:   .::.::::::::: :.:.. :: :::.:::::::: :::.:::.::
RAT    ARVAPVALPSACAPAGTQCLISGWGNTLSNGVNNPDLLQCVDAPVLSQADCEAAYPGEIT
              130       140       150       160       170       180
              190       200       210       220       230       240
HUMAN  SNMFCVGFLEGGKDSCQGDSGGPVVCNGQLQGVVSWGDGCAQKNKPGVYTKVYNYVKWIK
       :.:.::::::::::::::::::::::::::::.:::: :::  ..::::::: :.: ::.
RAT    SSMICVGFLEGGKDSCQGDSGGPVVCNGQLQGIVSWGYGCALPDNPGVYTKVCNFVGWIQ
              190       200       210       220       230       240

HUMAN  NTIAAN
       .:::::
RAT    DTIAAN
```

Human (247 aa) vs Rat (246 aa) Trypsin : show 76.4% identity (91.9% similarity) in 246 aa overlap (1-246:1-246) , **E(1) < 2e-86**

The similarity is statistically significant (> expected by chance) , so sequences can be considered **homologous**

Sequence Alignment:- Structure

Global + local sequence alignment example – a protein structure analysis



Nature 437, 215-223 (8 September 2005) | doi:10.1038/nature03978; Received 23 May 2005; Accepted 4 July 2005; Published online 24 July 2005

Crystal structure of a bacterial homologue of Na$^+$/Cl$^-$-dependent neurotransmitter transporters

Atsuko Yamashita[1], Satinder K. Singh[1], Toshimitsu Kawate[1], Yan Jin[2] & Eric Gouaux[1,2]

Sequence Alignment Problems: global & local-1

- Sequences can be aligned:
  - Matching as many characters as possible across their entire length (**Global alignment**)
    - The tool for global alignment is based on the **Needleman-Wunsch algorithm**
  - Focusing on just the best –matching (highest scoring) regions (**Local alignment**)
    - The tool for local alignment is based on **Smith-Waterman algorithm**
  - Both algorithms are derivatives from the basic dynamic programming algorithm (see later, **session 2**).

Sequence Alignment Problems: global & local-2

Global alignment:

```
L G P S S K Q T G K G S – S R I W D N
|         |     | | |       |     |
L N – I T K S A G K G A I M R L G D A
```

Local alignment:

```
- - - - - - - - G K G - - - - - - - -
                | | |
- - - - - - - - G K G - - - - - - - -
```

Sequence Alignment Problems: global & local-3

Global alignment:

- Suitable for-
  - Sequences that are quite similar (more closely related)
  - Sequences of approximately same length
- Global alignment made possible by including gaps either within the alignment or at the ends of the sequences

Local alignment:

- Suitable for-
  - Sequences similar along some of their lengths but dissimilar in others (i.e. sharing several conserved regions of local similarity/domains)
  - Sequences that differ in length
- Gaps not tolerated within local alignment

Pair-wise Sequence Alignment

- Pair-wise sequence alignment maps and compares residues between two sequences
- Aligning two sequences has many distinct alignment options possible
- The overall goal is to find the alignment that provides the best (optimal) pairing between the two sequences (i.e. maximum residue/character matches, gaps inclusive)
- Sequence alignments have to be scored to identify the best one/s among them.
- Scoring system can be simple **match/mismatch** scheme (DNA) or for protein comparisons , use of a more sensitive scheme by **substitution matrix**
- Often there is more than one solution with the same score

Treatment of gaps: Penalties-1

Constant gap penalty, a fixed – ve score "-a" is given as penalty of every gap, irrespective of length.

Aligning GCTGATTCAT Vs GCTTCAT

```
GCTGATTCAT
| |     |||||
GC - - -TTCAT
```

Score rules: Each match +1; The gap -1

Total score = 7-1 = 6

Treatment of gaps: Penalties-2

Linear gap penalty, a penalty of “-a” per unit length of a gap. Takes into account the length(L) of each insertion / deletion in the gap

Aligning GCTGATTCAT Vs GCTTCAT

```
GCTGATTCAT
| |   | | | | |
GC - - -TTCAT
```

Score rules: Each match +1; Each gap -1

Total score = 7-3 = 4

Treatment of gaps: Penalties-3

Constant and linear gap penalties do not consider whether gap is opening or extending

Gaps at terminal regions treated with no penalty since many true homologous sequences can be of different lengths

Treatment of gaps: Penalties-4

Affine gap penalty considers Introducing (opening) and extension of gaps: Total gap penalty G= O+E(L-1)

Where O= opening penalty; E= extension penalty; L= length of gap

Pros:

- Opening gap costs more than extending
- More evolutionary sound

Draw back:

- penalty points are arbitrary chosen

Pair-wise alignment score: Example

Data: A G T A C Vs G T A A C

Score rules: +1 for match, -2 for mismatch, -3 for gap

2 matches, 0 gaps (-4)

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

4 matches, 1 insertion (+1)

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

3 matches (2 end gaps) (+1)

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

4 matches, 1 insertion (+1)

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

Scoring scheme rewards matches and punishes mismatches and gaps

Methods of Pair-wise Sequence Alignment

- Short and very related sequences…**By hand-** slide sequences on two lines of a word processor
- General initial exploration of your sequence: to discover repeats, insertions, deletions **etc…Dot plot/matrix methods**- simplest comparison method
- Intensive comparisons to arrive at the optimal alignment ..**Rigorous mathematical approach**
  - Dynamic programming (slow, optimal)
- Extensive comparisons involving long sequences (e.g. entire genomes) or a large set of sequences(e.g. database entries) ….**Heuristic methods** (fast, approximate)
  - Word search methods e.g. BLAST, FASTA etc

Continued in session 2