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

- ***Sequence alignment***: Understand Alignment strategies and Applications
- ***Sequence relationships***: Distinguish between similarity and identity
- ***Sequence divergence***: Introducing concepts of point mutations, deletions, insertions etc.
- ***Sequence evolution:*** Understand the concepts of homology, identity, orthologues, paralogues
- ***Pairwise Sequence alignment***: Key aspects

CC BY NC SA

Learning Outcomes

Understanding concepts of:

- Sequence alignment

- Sequence evolution & relatioships:- mutations, deletions, insertions, homologs, paralogs and orthologs etc

- Similarity and differences between sequences

- Sequence alignment methods and when the local or global approach is the most appropriate

CC BY NC SA

What is Sequence Alignment?

- **Sequence alignment** is the procedure of arranging two or more sequences (DNA, RNA, or A.a.(proteins)) to identify regions of similar/different 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

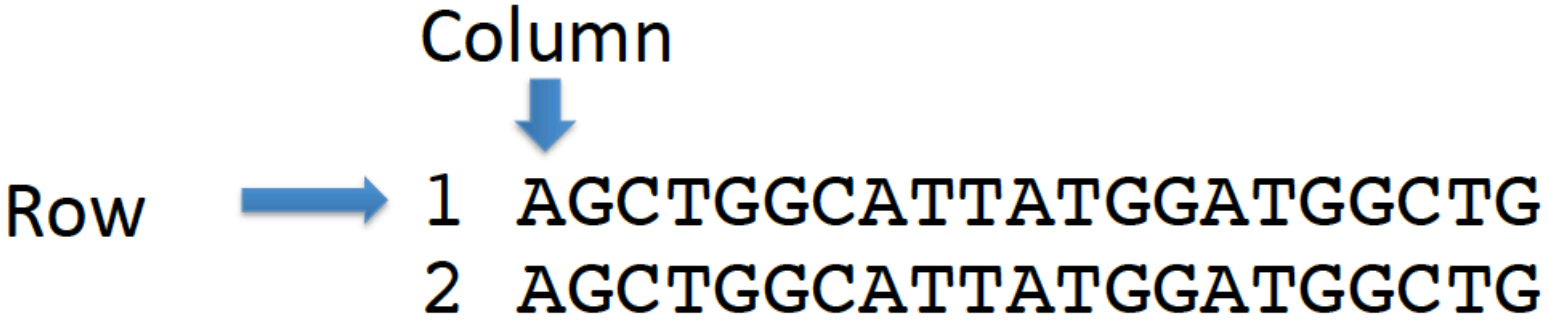


Fig 1: Sample Alignment of a sequence pair

Why Sequence Alignment? Uses:- 1

- Useful in DNA and Protein sequence analysis for:
  - Predicting **function** of a gene or protein
  - Predicting molecular **structure**
  - Discovering **evolutionary/phylogenetic** relationships

- Sequences that are **very alike** (highly similar) probably have:
  - Same function (**should be treated as hypothetical until experimentally tested**)
  - Similar secondary and 3-D structure (if proteins)
  - Shared ancestral sequence (**though not always**)

CC BY NC SA

Conserved sequence patterns may represent shared Functional Domains

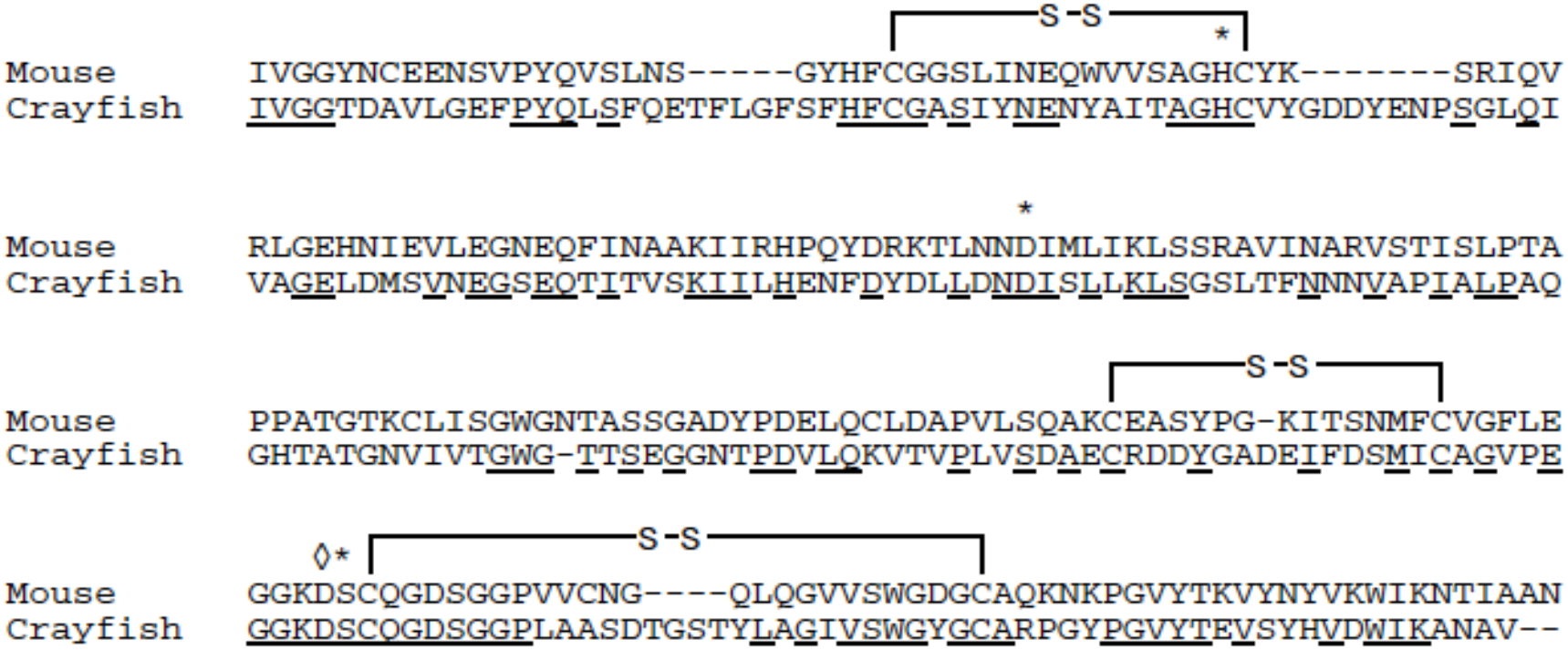


Fig 2: Mouse trypsin (SWISS-PROT P07146) / Crayfish trypsin (SWISS-PROT P00765) sequence alignment. Identical residuals are underlined. Indicated above the alignments are three disulfide bonds (-S-S-), with participating cysteine residues conserved, amino acids side chains involved in the charge relay system (asterisk), and active side residue governing substrate specificity (diamond). Credit: "Bioinformatics A practical Guide to the Analysis of Genes and Proteins" *Wiley InterScience 2nd Edition*

Warning: Always some exceptions to consider

```
Human-ZCr       MATGQKLMRAVRVFEFGGPEVLKLRSDIAVPIPKDHQVLIKVHACGVNPVETYIRSGTYS
Ecoli-QOR       ------MATRIEFHKHGGPEVLQA-VEFTPADPAENEIQVENKAIGINFIDTYIRSGLYP
                      .   .     ******.   . .   * .... .  .* *.* ..****** *

Human-ZCr       RKPLLPYTPGSDVAGVIEAVGDNASAFKKGDRVFTSSTISGGYAEYALAADHTVYKLPEK
Ecoli-QOR       -PPSLPSGLGTEAAGIVSKVGSGVKHIKAGDRVVYAQSALGAYSSVHNIIADKAAILPAA
                  * **   *.. **..  **  .    * ****  . .  * *.           **

Human-ZCr       LDFKQGAAIGIPYFTAYRALIHSACVKAGESVLVHGASGGVGLAACQIARAYGLKILGTA
Ecoli-QOR       ISFEQAAASFLKGLTVYYLLRKTYEIKPDEQFLFHAAAGGVGLIACQWAKALGAKLIGTV
                . * * **  .   * *  * .. .*  * * * *.***** *** * .* * * ..**

Human-ZCr       GTEEGQKIVLQNGAHEVFNHREVNYIDKIKKYVGEKGIDIIIEMLANVNLSKDLSLLSHG
Ecoli-QOR       GTAQKAQSALKAGAWQVINYREEDLVERLKEITGGKKVRVVYDSVGRDTWERSLDCLQRR
                **  .  .  *. ** .* * **    ....*   * * . .. . .   .  . *  * .

Human-ZCr       GRVIVVG-SRGTIEINPRDTMAKES----SIIGVTLFSSTKEEFQQYAAALQAGMEIGWL
Ecoli-QOR       GLMVSFGNSSGAVTGVNLGILNQKGSLYVTRPSLQGYITTREELTEASNELFSLIASGVI
                * ..  * * *..       . .      .   .  . .*.** . .  * . .  * .

Human-ZCr       KPVIGSQ--YPLEKVAEAHENIIHGSGATGKMILLL
Ecoli-QOR       KVDVAEQQKYPLKDAQRAHE-ILESRATQGSSLLIP
                *  .  *  ***     *** *.    . *  .*.
```

Fig 3. *Human zeta-crystallin (SWISS-PROT Q08257) / E-coli quinone oxidoreductase (SWISS-PROT P28304) global sequence alignment via Clustaw W program. (Higgins et al., 1996). Identical residues are marked by asterisks below the alignment, and dots indicate conserved residues.*

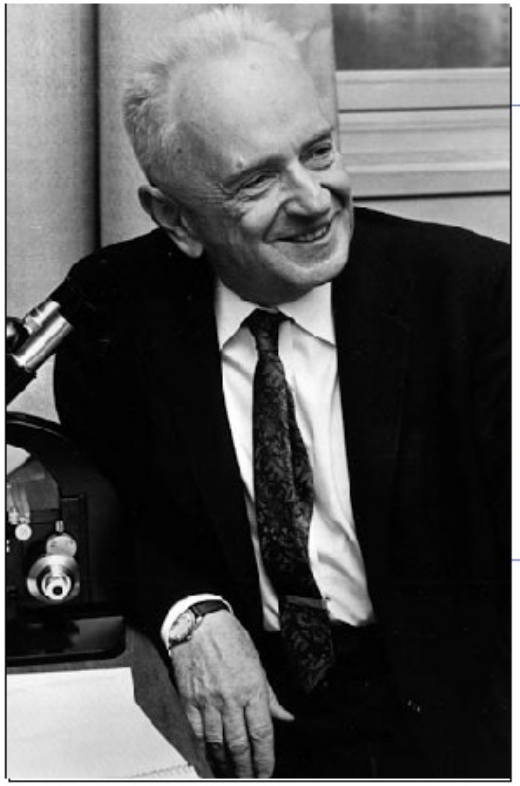
Fig 4. ***Structure Vs Function: Phantom ship-as Dinner*** *@ The Cargo hold restaurant Ushaka Marine World , Durban SA*

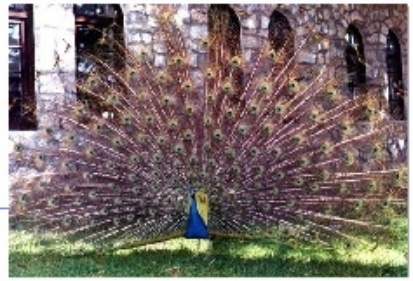
Why Sequence Alignment? Uses:- 2

- Sequence alignment also enables the following:
  - Annotation of new sequences
  - Fragment/ genome assembly- Merging strings of DNA or RNA sequences
  - Detect gene mutations

CC BY NC SA

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

CC BY NC SA

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

cc BY NC SA

Sequence Relationships-2a

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

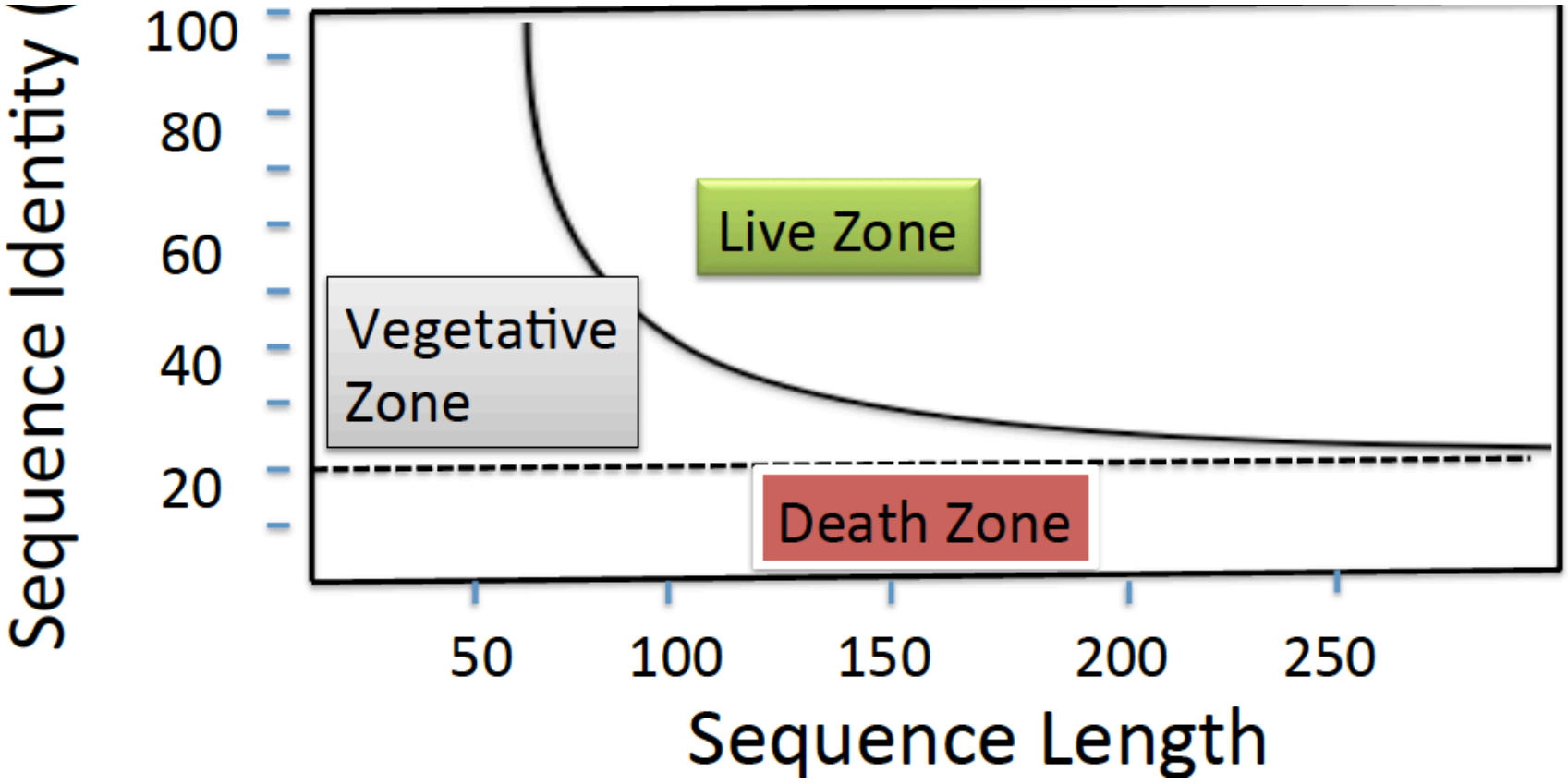
Sequences be at least 100 A.a or 100 nucleotides long, then:

  - **20 - 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

CC BY NC SA

Sequence Relationships-2b: Infering homology



Sequence Relationships-2c

- 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

CC BY NC SA

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

CC BY NC SA

% Similarity/Identity: Nucleotides-1

Equal Length:

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

$$= [2L/(L_y + L_z)] \times 100$$

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

CC BY NC SA

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

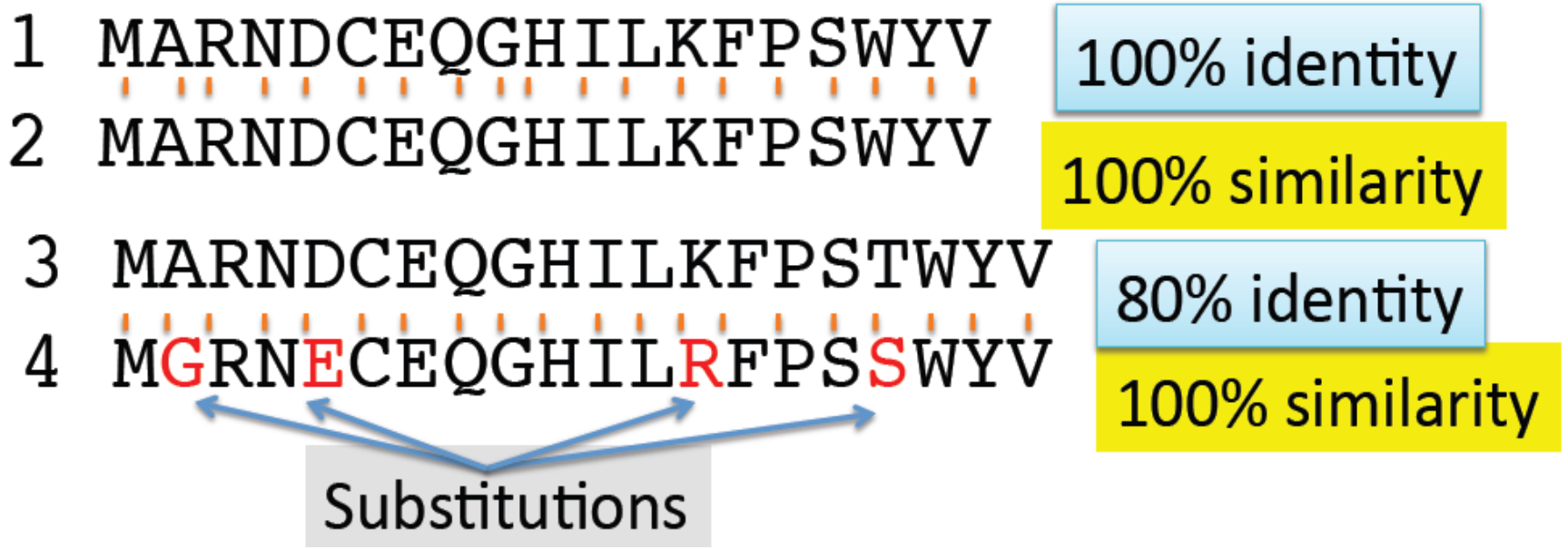
$L_{i(s)}$ is the number of aligned residues with similar or identical characteristics

L_y is the length of the shorter of the two sequences

BY NC SA

Similarity/Identity: Amino Acids-1

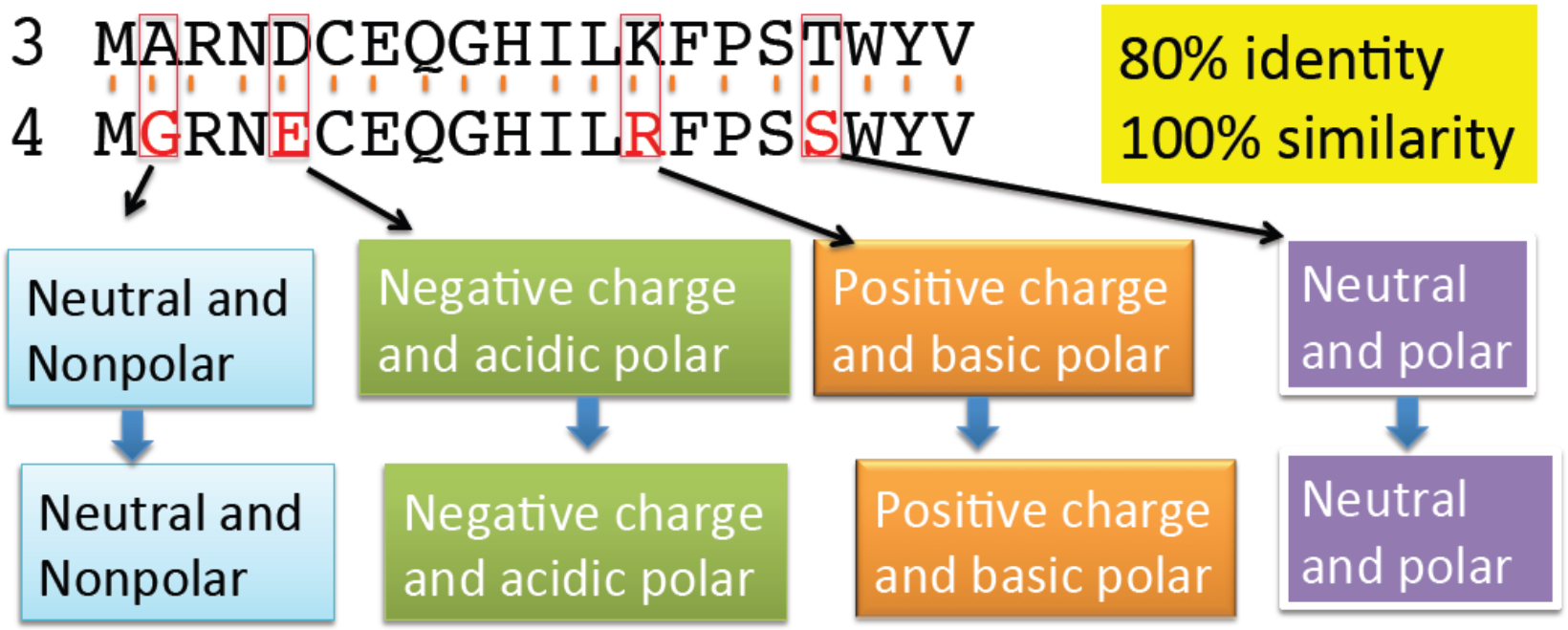
❖% Identity and similarity not synonymous for Amino acid sequences:



Credit Pandam Salifu , IBT 2016

Similarity/Identity: Amino Acids-2

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



Credit Pandam Salifu , IBT 2016

Sequence Relationships-3a

- **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 → ∓ TcATC

CC BY NC SA

Sequence Relationships-3b

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

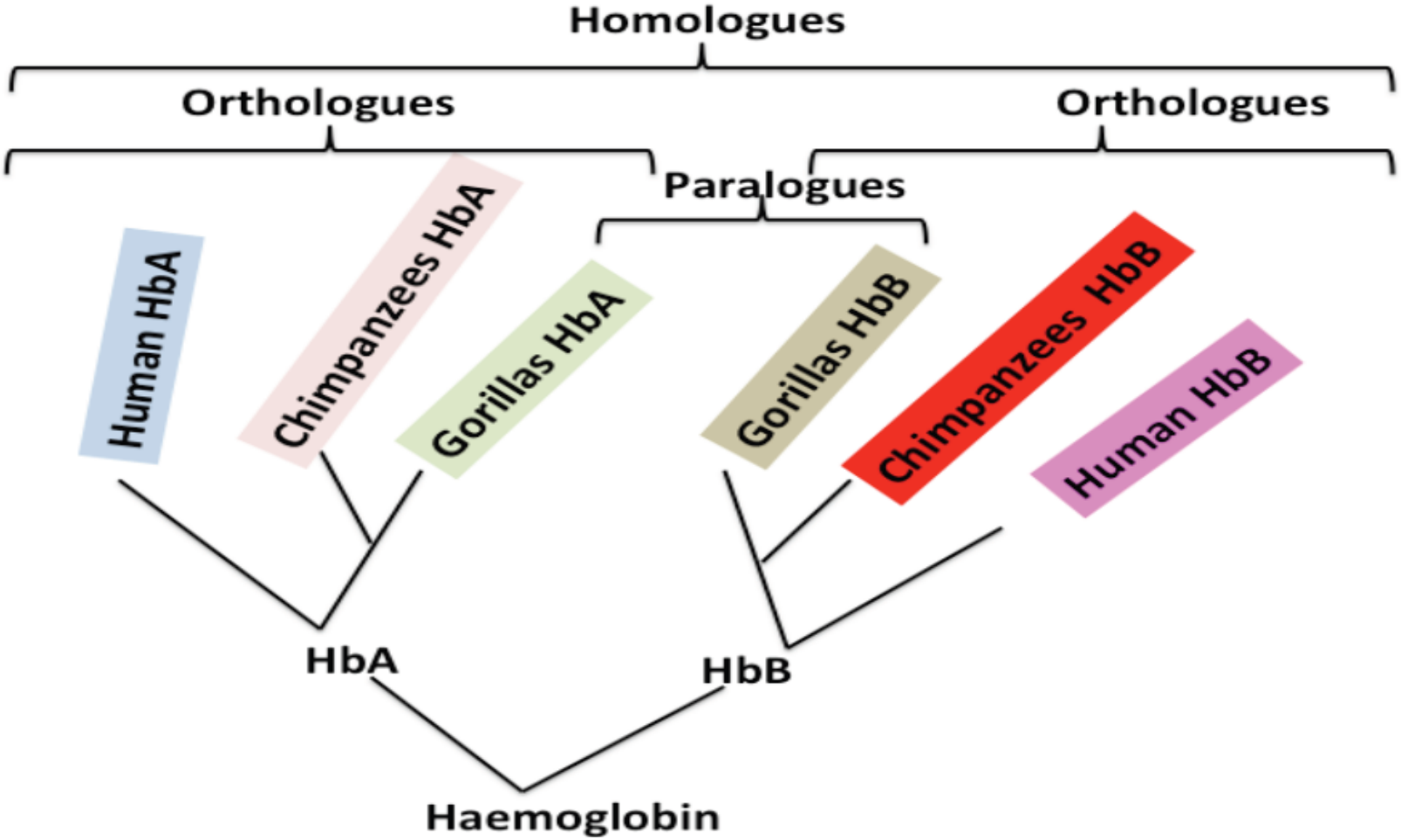
CC BY NC SA

Sequence Relationships-3c

- **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
- *Orthologs & Paralogs are forms of homologs*
- **Analogous sequences:** related through convergency
- **Xenologous sequences:** related through direct transfer of genetic material between species

CC BY NC SA

Sequence Relationships-3d



Credit Pandam Salifu , IBT 2016

CC BY NC SA

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
```

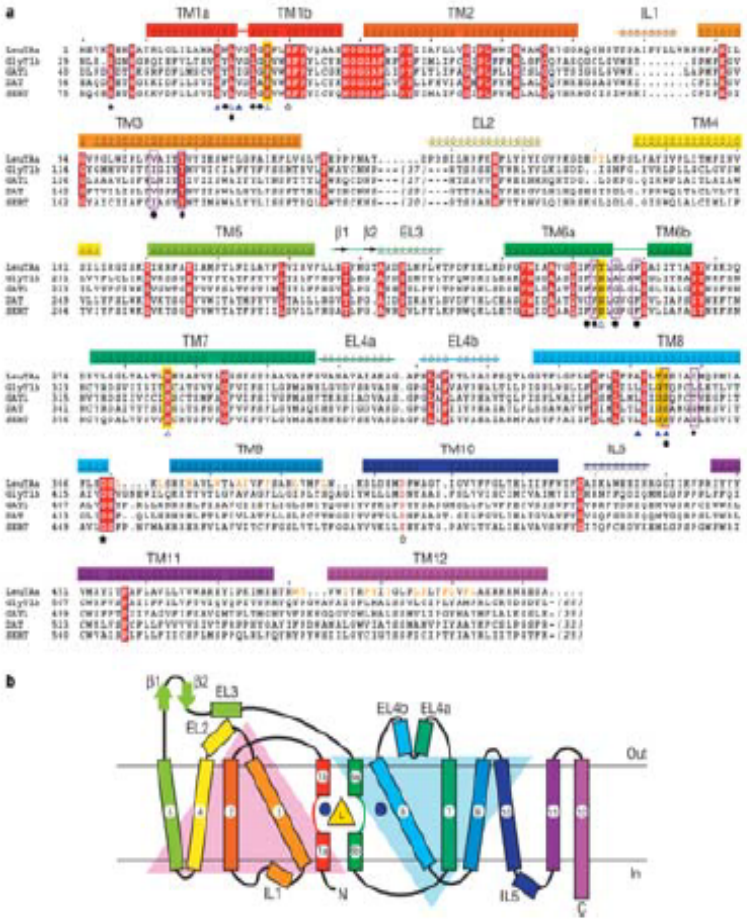
Human (247 aa) vs Rat (246 aa) Trypsin : show 76.4% identity (91.9% similarity) in 246 aa overlap (1-246:1-246) , **E(1) < 2e-86**

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

CC BY NC SA

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]

CC BY NC SA

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**).

CC BY NC SA

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 - - - -  - - - -
```


CC BY NC SA

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 less tolerated within local alignment

CC BY NC SA

Pair-wise Sequence Alignment- 1

- **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)

CC BY NC SA

Pair-wise Sequence Alignment- Optimal

Which of the following alignment is more likely optimal alignment for sequences GATTCGGACCTCGGATCCCGT and ATCGACTCGATCGT?

1
```
GATTCGGACCTCGGATCCCGT
-AT-C-GA-CTC-GAT-C-GT
```

2
```
GATTCGGACCTCGGATCCCGT
-A-TCG-AC-TC-GAT--CGT
```

3
```
GATTCGGACCTCGGATCCCGT
-AT-C-GAC-TCG-ATC--GT
```

- It is difficult to choose one
- = ~ 2^{39} possible alignments for two sequences

Credit Pandam Salifu , IBT 2016

- 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

CC BY NC SA

Optimal Alignment

- Employ scoring scheme- to reward matches/ similarity, punish mismatches & gaps
- Similarity of two sequences- High score
- Dissimilarity of two sequences- Low score
- Account for substitutions, insertions, deletions

CC BY NC SA

Optimal Alignment: Gaps

- It is desirable to allow some gaps to be introduced into an alignment to compensate for insertions and deletions

- Gaps come with a cost & process not arbitrary

- Several strategies been proposed to penalize for gaps

CC BY NC SA

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

cc BY NC SA

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

BY NC SA

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

BY NC SA

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

CC BY NC SA

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 TA 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

cc BY NC SA

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

CC BY NC SA

Dot Matrix Sequence Comparisons

- Good practice to start with a dot plot:
  - Provides an overview of the possible relationships between your two sequences
    - Regions of similar sequence
    - Repeat identifications (Direct or Inverted)
    - Predicting self-complementary regions (in RNA can form 2^0 structures)
  - Helpful tool in deciding next steps

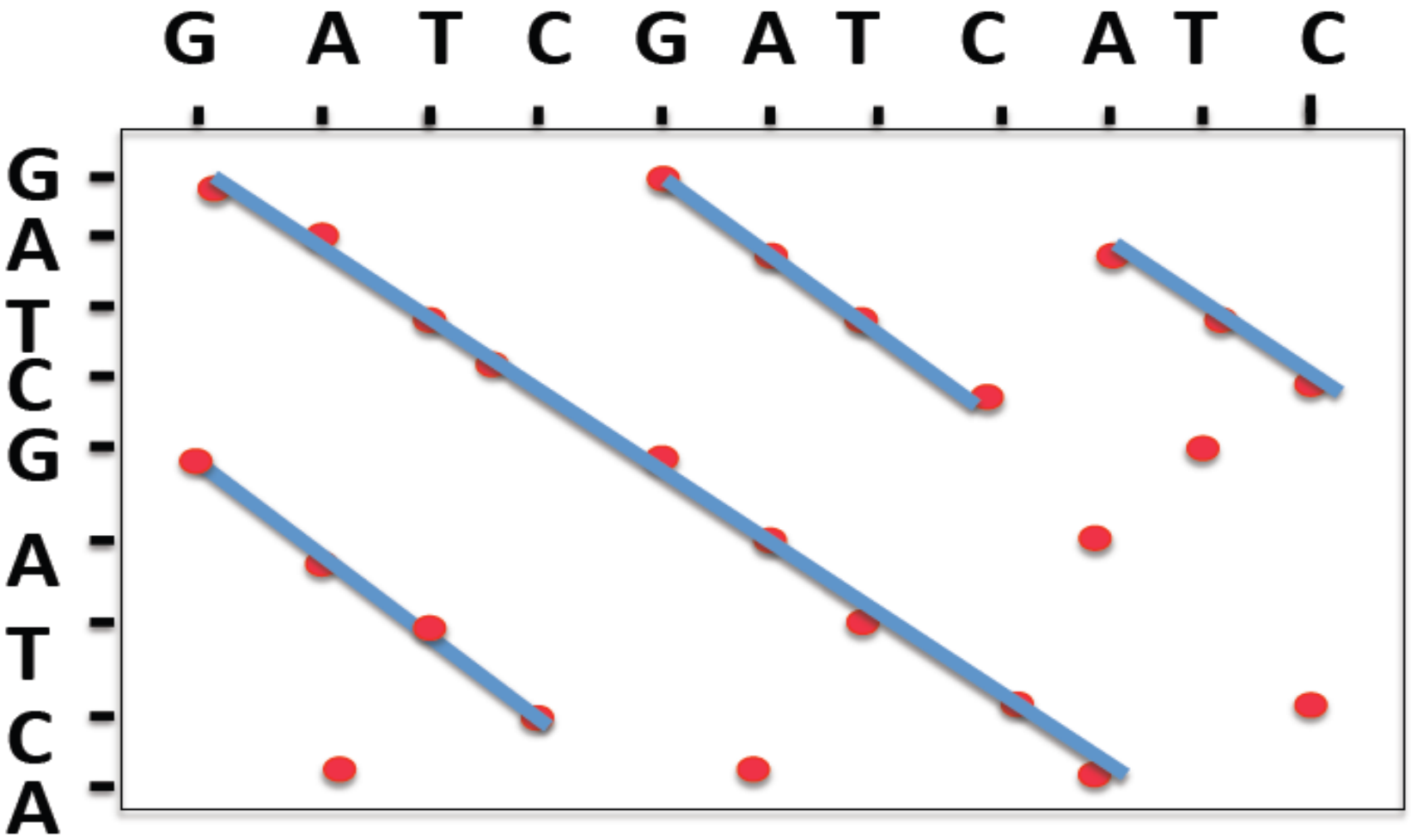
CC BY NC SA

Making a Dot Plot

- Dot plot compares two sequences for possible similarities
- **Dot plot algorithm**:
  1. Draw a grid to write out the sequences
  2. One sequence (A) is listed across the top and the second sequence (B) listed down the left side
  3. Starting from the first character in B, one moves across the page keeping in the first row and placing a dot in any column where the character in A is the same
  4. The process is continued until all possible comparisons between A and B are made
  5. Any region of similarity is revealed by a diagonal row of dots
  6. Isolated dots not on diagonal represent random matches

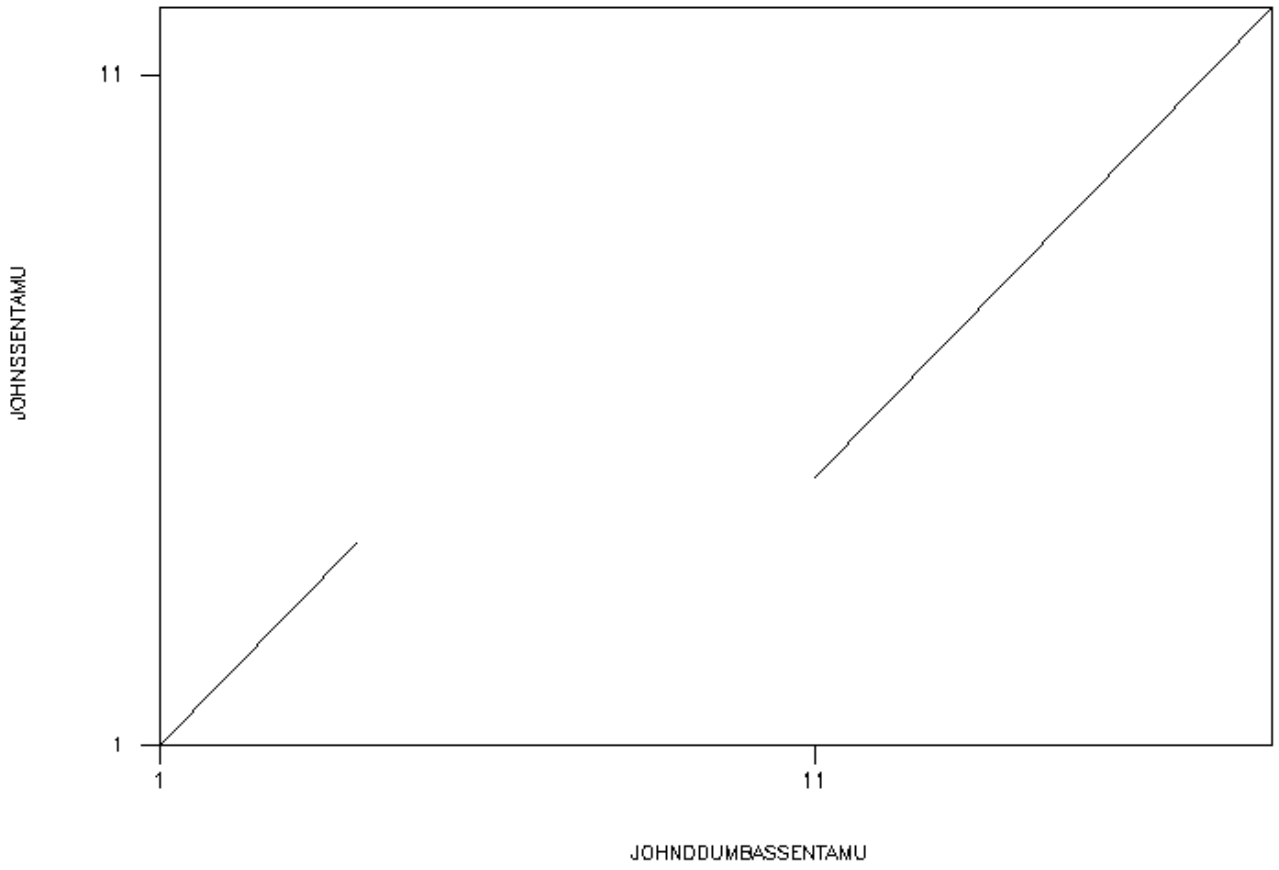
cc BY NC SA

Dot Plot: Example



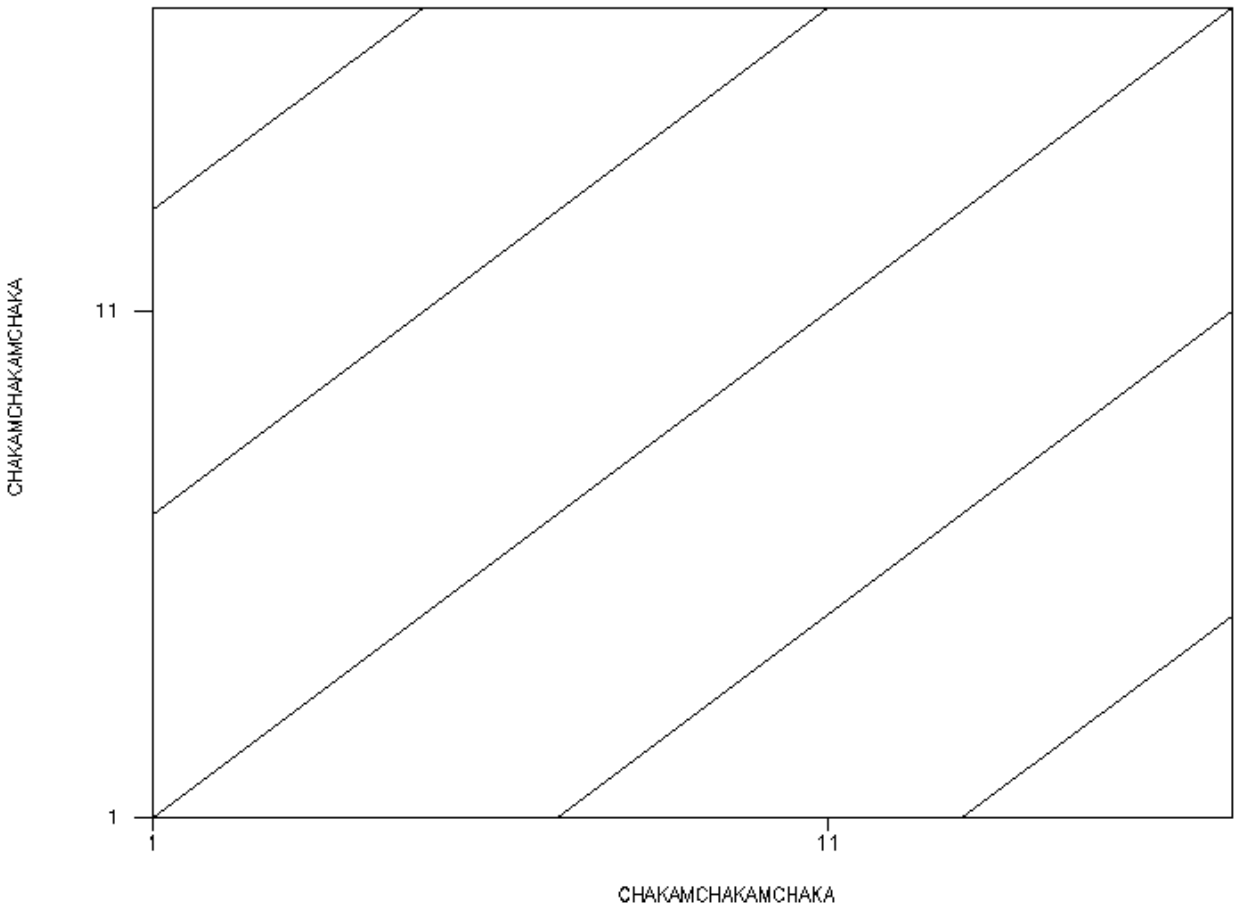
CC BY NC SA

Dot Plot: Insertions: JOHNDDUMBASSENTAMU Vs JOHNSSENTAMU



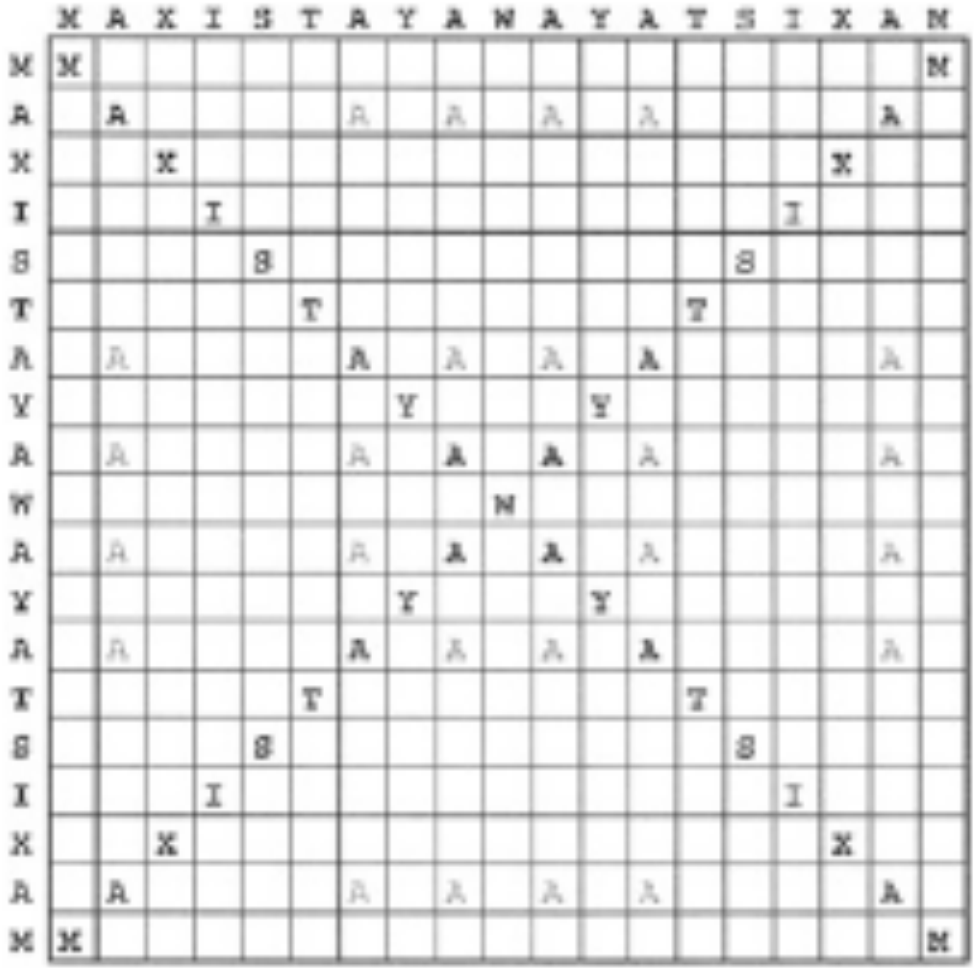
Dottup alignment http://www.bioinformatics.nl/cgi-bin/emboss/dottup: Gap in the sequential diagonal indicative of insertion (DDUMBA) in one of the sequence in their similar regions

Dot Plot: Repeats:
CHAKAMCHAKAMCHAKA



Dottup alignment http://www.bioinformatics.nl/cgi-bin/emboss/dottup: Identity between repetitive sequence CHAKAMCHAKAMCHAKA repeats appear on several subsidiary diagonals parallel to the main one

Dot Plot: Palindromes



Credit Lesk A.M , Introduction to Bioinformatics 2002

Regions in DNA recognized by **transcriptional regulators** or **restriction enzymes**, or **endonucleases** have sequences related to palindromes, crossing from one strand to the other:

- EcoRI recognition sites
  GAATCC

  CCTAAG
- **CRISPR Cas** systems recognize **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeat sites

CC BY NC SA

Dot Plot/ Matrix Considerations

- There are many dot plot flavors/programs to pick from

- If using computer program, default output might need to be refined by adjusting various settings e.g. sliding window size (comparisons of residues in immediate vicinity)
  - ✧ Long windows make clean dot plots (i.e. more stringent)
  - ✧ Shorter windows more sensitive but come with noise..can help in scenarios of distantly related proteins
  - ✧ Start large then progressively reduce until signal in question appears.

Dot Plot/ Matrix Protein analysis Variations

- Variations can me made to improve protein analysis:

  ✧ Chemical similarity of R groups/ or other feature as basis for sequence similarity considerations

  ✧ Amino acid scoring matrix (see later) could be incorporated into the match assessments.

  ✧ Several different matrices can be made , each with a different scoring system, and diagonal scores averaged. Most significant diagonals then identified

CC BY NC SA

Dot Matrix Sequence Comparison

- Dot plots not enough for detailed examination:
  - Most Don't really produce alignment (simply give generic indications)- Alignments followed up using other methods

- Alignments needed for in-depth pair-wise sequence comparisons

CC BY NC SA

Dot Matrix: Advantages

- Easy identification of the regions of greatest similarity
- Identification of sequence repeat regions based on the presence of parallel diagonals
- Useful in identifying chromosomal repeats and in comparing gene order conservation between two closely related genomes

CC BY NC SA

Dot Matrix: Disadvantages

- Lack statistical rigor in assessing quality of alignment
- Dot plots not enough for detailed examination:
  - Most Don't really produce alignment (simply give generic indications)- Alignments followed up using other methods
- Difficult to scale up to multiple alignments

cc BY NC SA