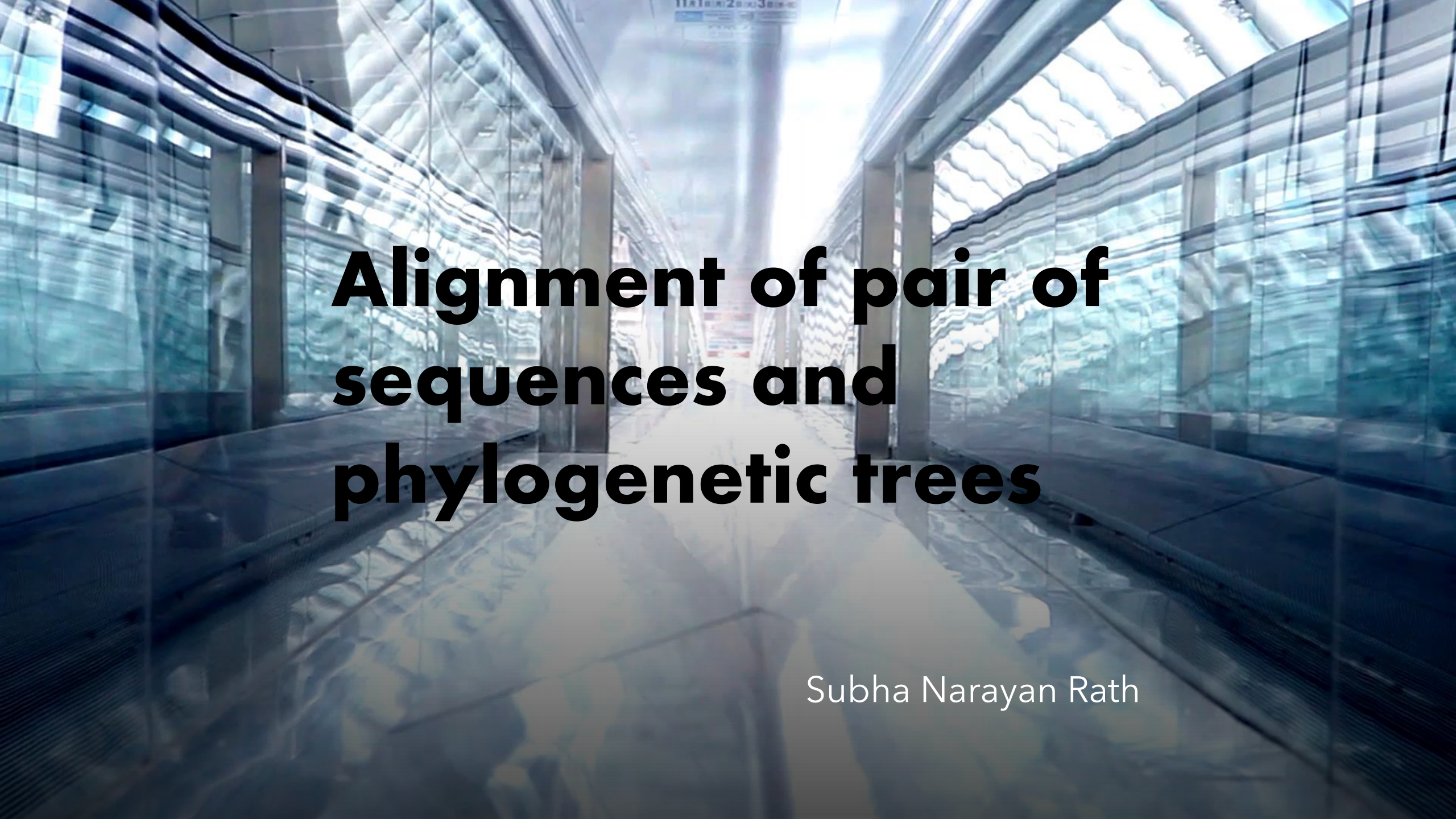


# Alignment of pair of sequences and phylogenetic trees

Subha Narayan Rath

# Intro to sequence alignment

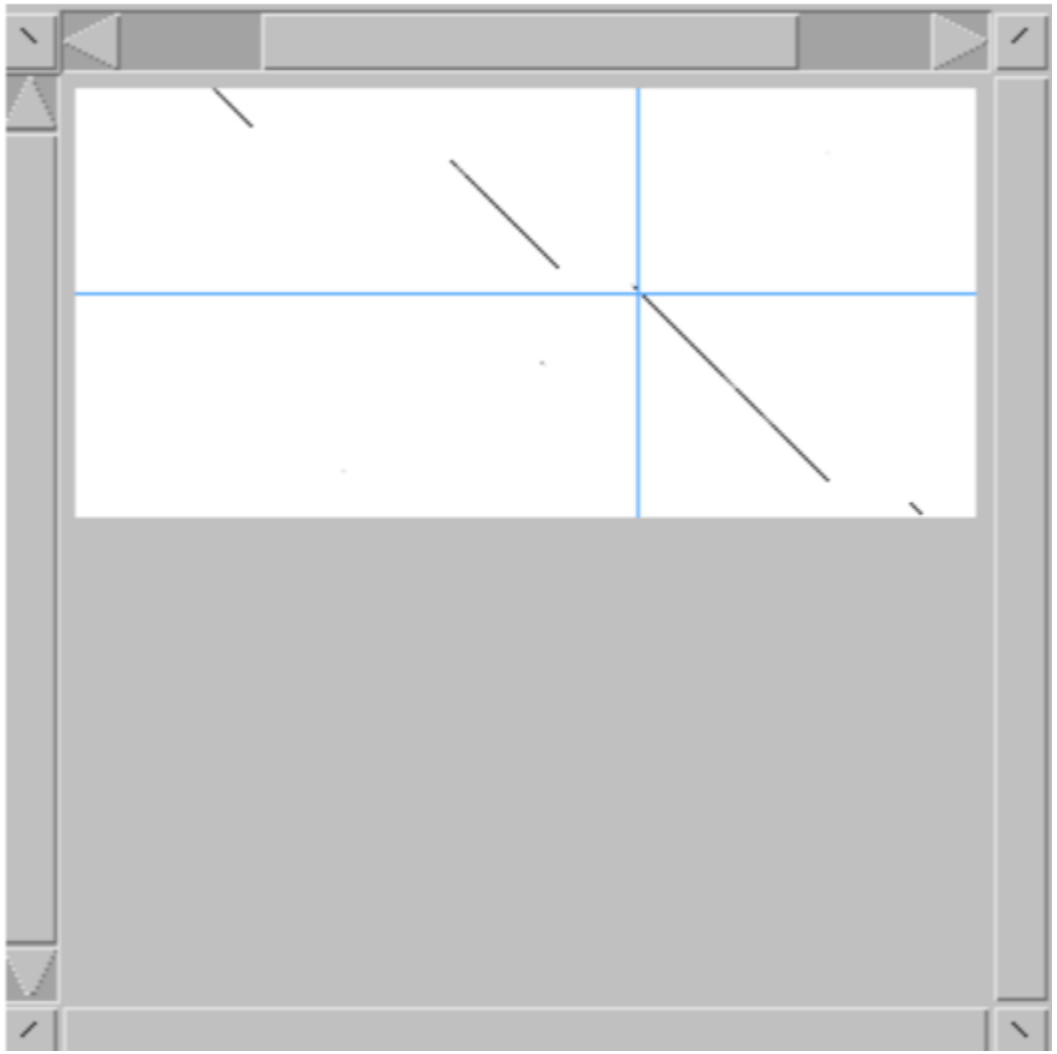
- Given 2 or more sequences:
  - Measure their similarity
  - Determine the residue-residue correspondence
  - Observe the patterns of conservation and variability
  - Infer evolutionary relationships
- Sequence alignment is the identification of residue-residue correspondences.
- A mutual alignment of more than 2 sequences is called multiple sequence alignment

# Dotplot

- IT is a tool in bioinformatics to compare two sequences and show the sequences of close similarity in a visually understandable way.
- Sliding window size: noisy vs smooth (default is 10)
  - Window size changes with goal of analysis
    – size of average exon
    – size of average protein structural element
    – size of gene promoter
    – size of enzyme active site
- Cut-off value by statistics and Z score
- If the direction of the movement: diagonal or horizontal or vertical….

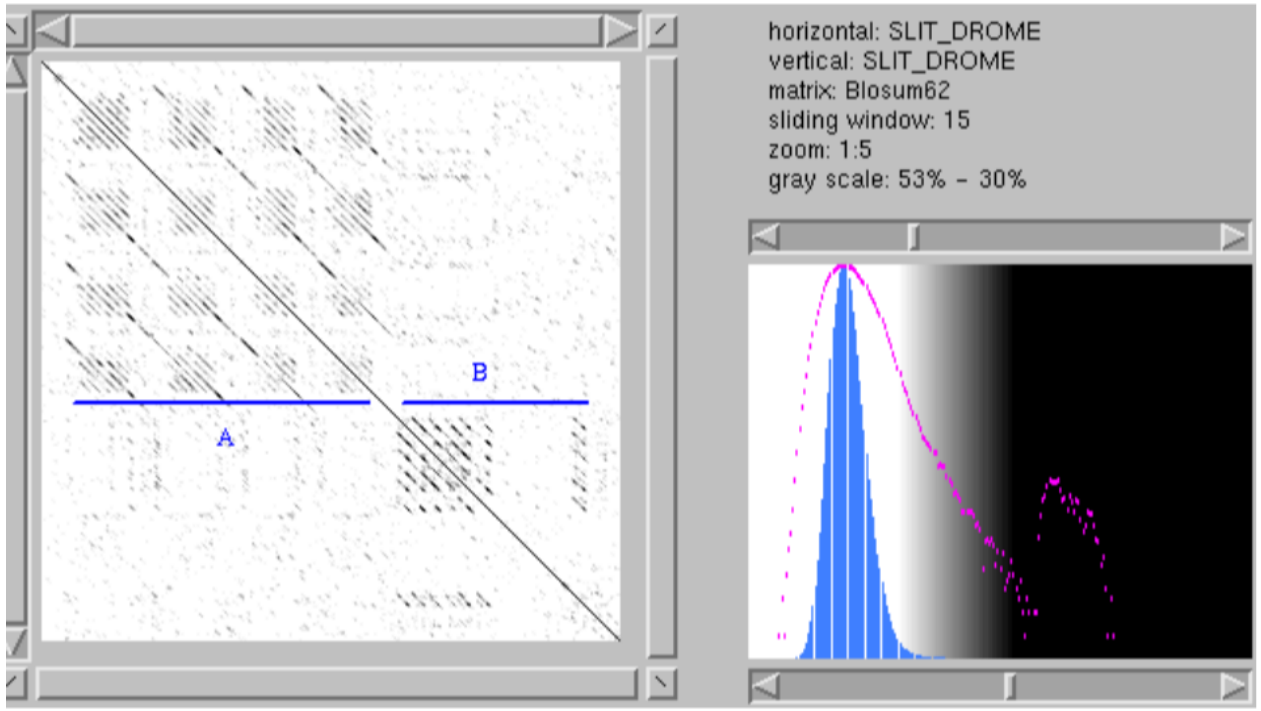
# Dotplot can show

- Repeated domains
- Conserved domains
- **Exons and introns**
- Terminators
- Frameshifts
- Low-complexity regions

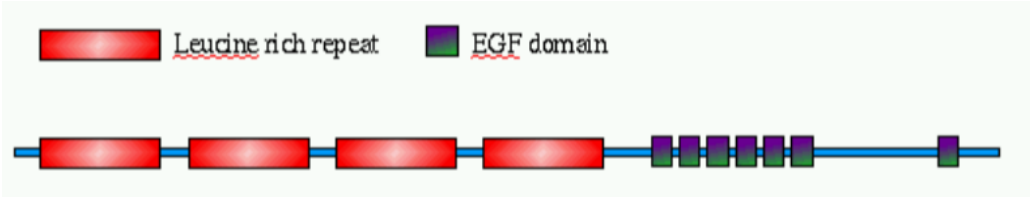


https://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher

# Problem 5.1: Repeated domains in SLIT protein




Drosophila melanogaster SLIT protein against itself




Arrangement of domains as described in Swiss-Plot entry

# Software to create a dotplot

- ANACON – Contact analysis of dot plots.
- D-Genies– Specializes in interactive whole genome dotplots of large genomes
- Dotlet – Provides a program allowing you to construct a dot plot with your own sequences.
- dotmatcher– Web tool to generate dot plots (and part of the EMBOSS suite).
- Dotplot – easy (educational) HTML5 tool to generate dot plots from RNA sequences.
- dotplot – R package to rapidly generate dot plots as either traditional or ggplot graphics.
- Dotter– Stand alone program to generate dot plots.
- JDotter – Java version of Dotter.
- Flexidot – Customizable and ambiguity-aware dotplot suite for aesthetics, batch analyses and printing (implemented in Python).
- Gepard – Dot plot tool suitable for even genome scale.
- Genomdiff – An open source Java dot plot program for viruses.
- LAST for whole-genome “split-alignment”.
- lastz and laj – Programs to prepare and visualize genomic alignments.
- yass – Web-based tool to generate (both forward and reverse complement) dot plots from genomic alignments.
- seqinr – R package to generate dot plots.
- SynMap – An easy to use, web-based tool to generate dotplots for many species with access to an extensive genome database. Offered by the comparative genomics platform CoGe.
- UGENE Dot Plot viewer – Opensource dot plot visualizer.

# Measures of sequence similarity

- Given 2 strings, two measures of the distance between them.
- A. Hamming distance: defined between two strings of equal length, is the number of positions with mismatching characters
- B. Levenshtein or edit, distance: defined between two equal or unequal length, is the minimal number of “edit operations’ required to change one string to the other.
- It could be Deletion, Insertion, Substitution of a single character in either sequence
- A given sequence of edit operations induces a unique alignment, but not vice versa!!!
- For molecular biology certain changes are likely to occur than others e.g. amino acid substitutions of similar sizes, so variable weights to different edit operations
- What about similarity scoring system??
- Transition mutation vs transversion mutation with higher points for former groups.

# Derivation of substitution matrices: PAM matrices

- To measure the relative probability of any particular substitution, first we find the relative frequencies of changes in pairs of aligned homologous sequences and based on that we can make a scoring matrix for substitutions.
- A common change should score HIGHER than a rare one
- A measure of sequence divergence is PAM = 1% Accepted Mutation
- 1 PAM apart of two sequences would have 99% identical residues and collecting statistics of these pair produces 1PAM substitution matrix
- Power of the matrix is used for more divergent sequences

# PAM and % identity

## PAM numbers vs. observed am.ac. mutational rates

| PAM Number | Observed Mutation Rate (%) | Sequence Identity (%) |
|---|---|---|
| 0 | 0 | 100 |
| 1 | 1 | 99 |
| 30 | 25 | 75 |
| 80 | 50 | 50 |
| 110 | 40 | 60 |
| 200 | 75 | 25 |
| 250 | 80 | 20 |

PAM 250 levels
(250% of expected change or
250 substituions per 100 amino acids),
corresponds to 20% overall sequence similarity.

The occurrence of reversions, either directly or via other changes, produces slowdown of mutation rates

**Note** *Think about intermediate "substitution" steps …*

# PAM 250 MATRIX OF M.O. DAYHOFF

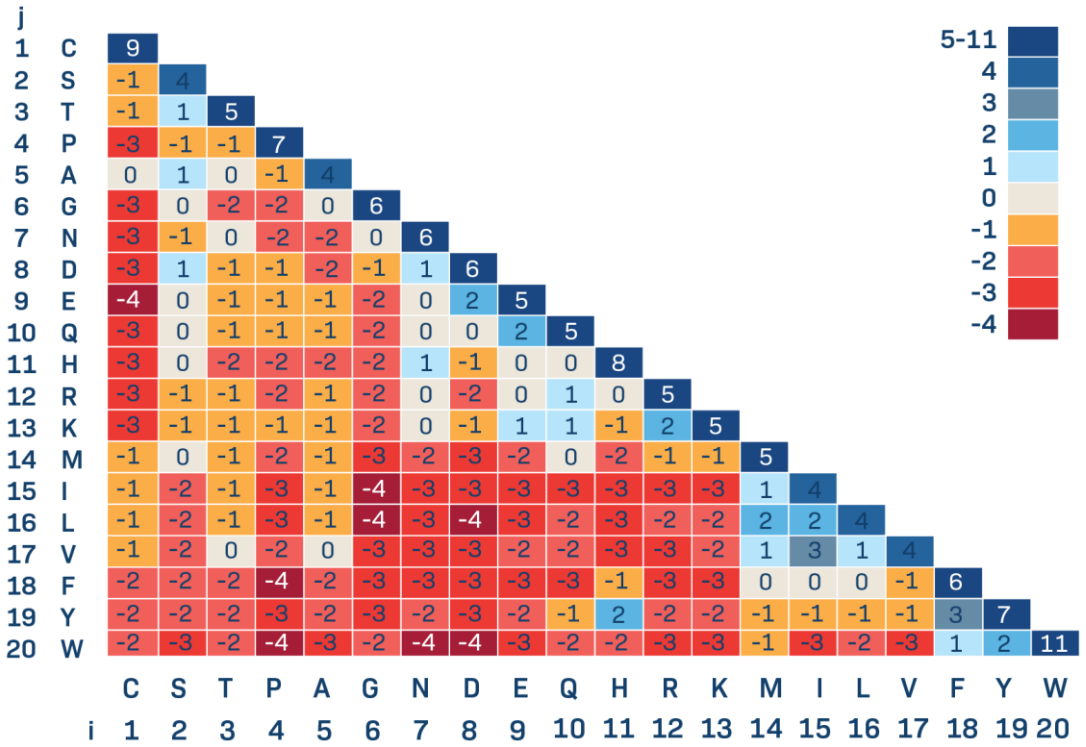
- It expresses scores of log-odds values:
- Score of mutation I to J = $\log_{10}$ (**observed** I to J mutation rate/mutation rate **expected** from amino acid frequencies)
- The numbers are multiplied by 10 to avoid decimals
- The probability of 2 independent mutations is the product of their individual probabilities and hence added.
- Score is +ve: sequences are related or conservative substitution

## The PAM 250 Scoring Matrix

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

# BLOSUM matrix of Henikoff and Henikoff

- Using much larger amount of data available now
- Means BLOcks SUbstittuion Matrix and based on BLOCKS database (representing known protein families) of aligned protein sequences
- From family of closely related proteins alignable without gaps… they calculated the ratio of number of observed pairs of amino acids at any position to the number expected from overall amino acid frequencies
- They have sequence identities higher than a threshold e.g. BLOSUM 62% is commonly used where the matrix built using sequences no more than 62% similarity



# Scoring insertions and deletions (substitution matrix) or gap weighting

- In addition to substituion matrix: there is a way of gap weighting too
- Aligning DNA sequnces: CLUSTAL-W is recommended
- Alignign Protein sequences: BLOSUM 62 is recommended

| | Matrix/ protocol recommended | Gap initiation | Gap extension | Match | Mismatch |
|---|---|---|---|---|---|
| DNA | CLUSTAL-W | 10 | 0.1 | 1 | 0 |
| Protein | BLOSUM62 | 11 | 1 | Matrix | Matrix |

# Computing the alignment of two sequences

- An algorithm used for this: dynamic programming and very imp for molecular biology
- Guarantee: to give an optimal global alignment
- Problem1: many alignment may give the same optimal score
- Problem 2: technical: the time required to align two sequences is proportional to n * m, as it is the size of edit matrix that must be filled in
- Variations of the dynamic programming method:
- 1. entire sequence to entire sequence: global match
- 2. region of one sequence to entire other sequence: local match
- 3. region of one to region of another: motif match
- Typical approximation approach would take a small integer k > all instances of each k-tuple of residues in the probe sequence that is found in database sequences