

# Introduction to the theory of sequence alignment

Yves Moreau
Master of Bioinformatics
Katholieke Universiteit Leuven
2001-2002

# References


- R. Durbin, A. Krogh, S. Eddy, G. Mitchinson, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Oxford University Press, 1998

- S.F. Altschul *et al.*, Basic Local Alignment Search Tool, *J. Mol. Biol.*, No. 215, pp. 403-410, 1990

- S.F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, Vol. 25, No. 17, pp. 3389-3402, 1997

# Overview

- Alignment of two sequences
  - DNA
  - Proteins
- Similarity vs. homology
  - Similarity
  - Homology
    - Orthology
    - Paralogy
- Elements of an alignment
- Dynamic programming

# Overview

- Global alignment
  - Needleman-Wunsch algorithm
- Local alignment
  - Smith-Waterman algorithm
  - Affine gap penalty
- Substitution matrices
  - PAM
  - BLOSUM
  - Gap penalty
- Significance evaluation
- BLAST

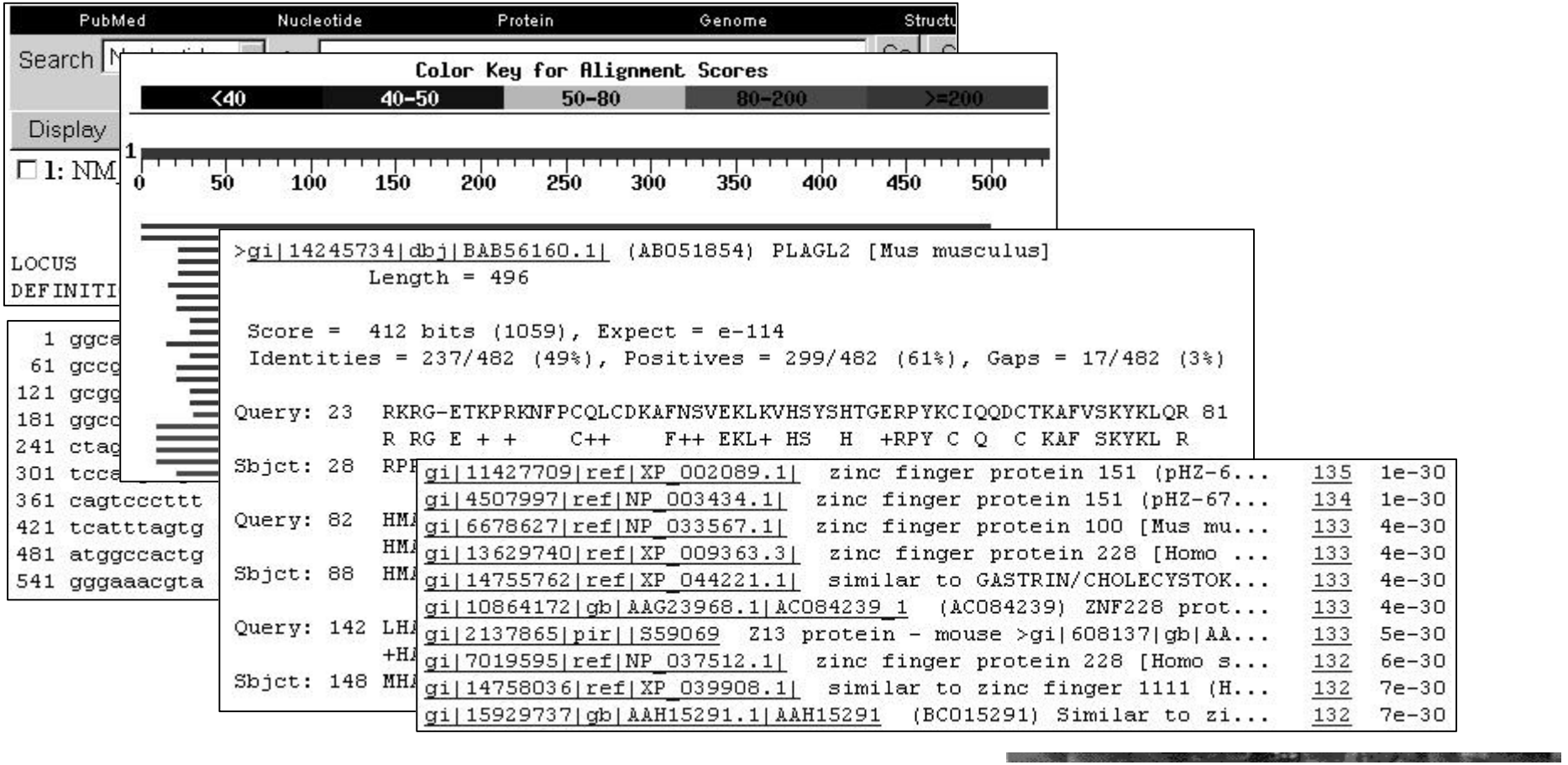
# Overview

- PHI-BLAST

- Multiple alignment
- PSI-BLAST
- CLUSTALW

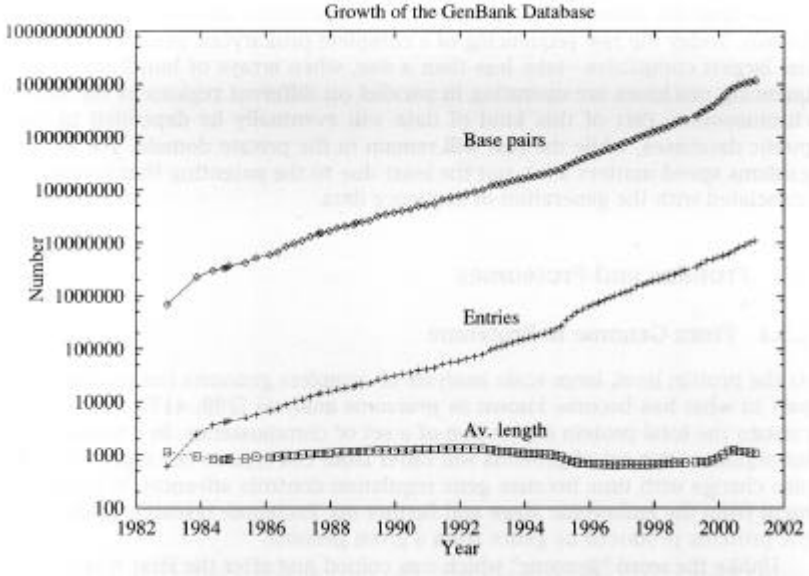
# Biological basis for alignment

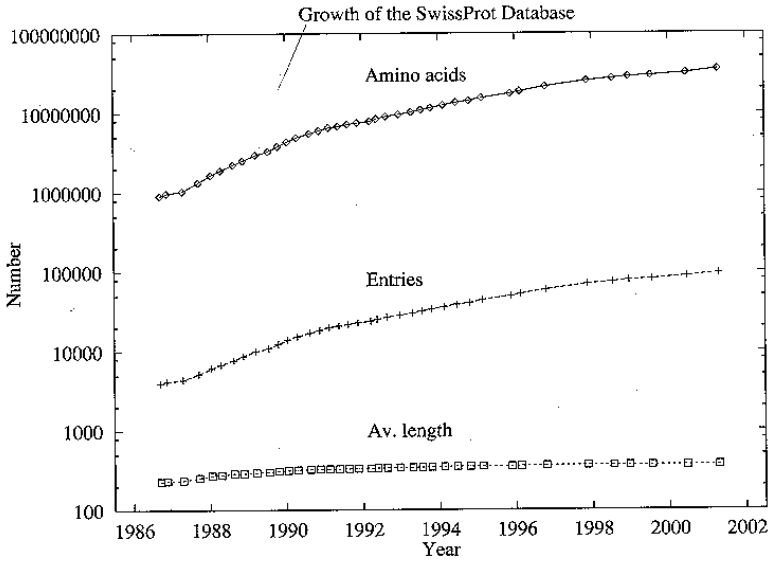
# BLAST for discovery



# Evolution of sequence databases

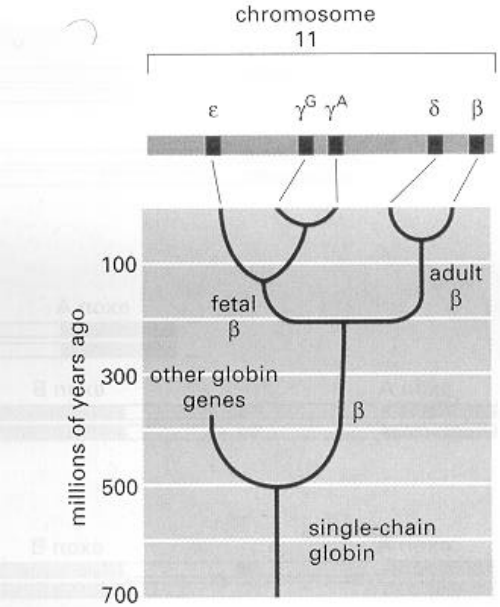
- Genbank
- SWISSProt



# Molecular evolution

- Genomes through imperfect replication and natural selection
- Gene duplications create gene families



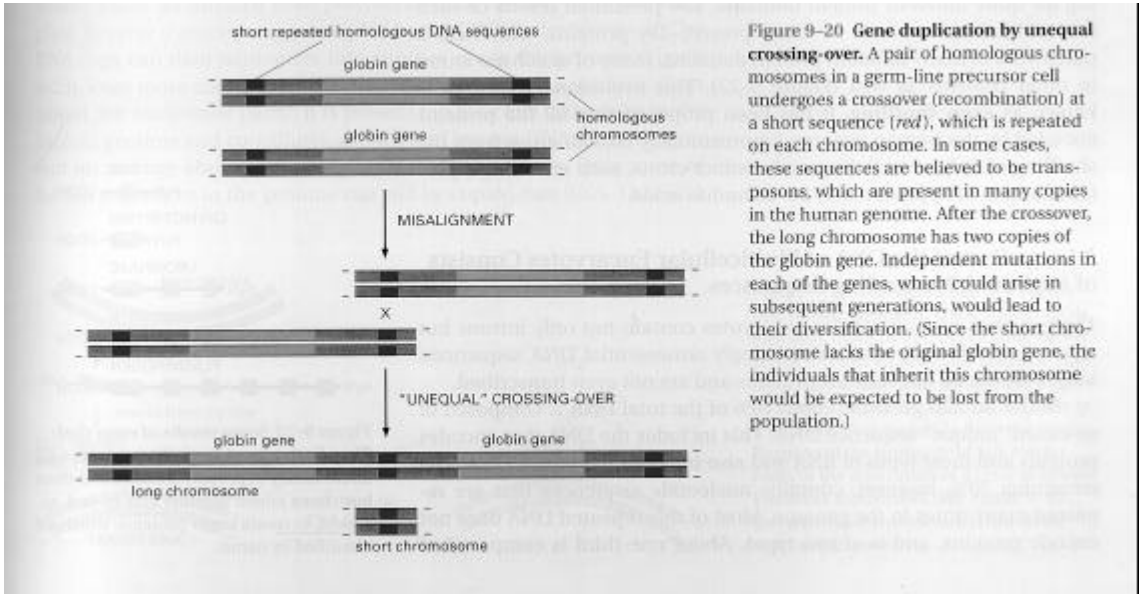



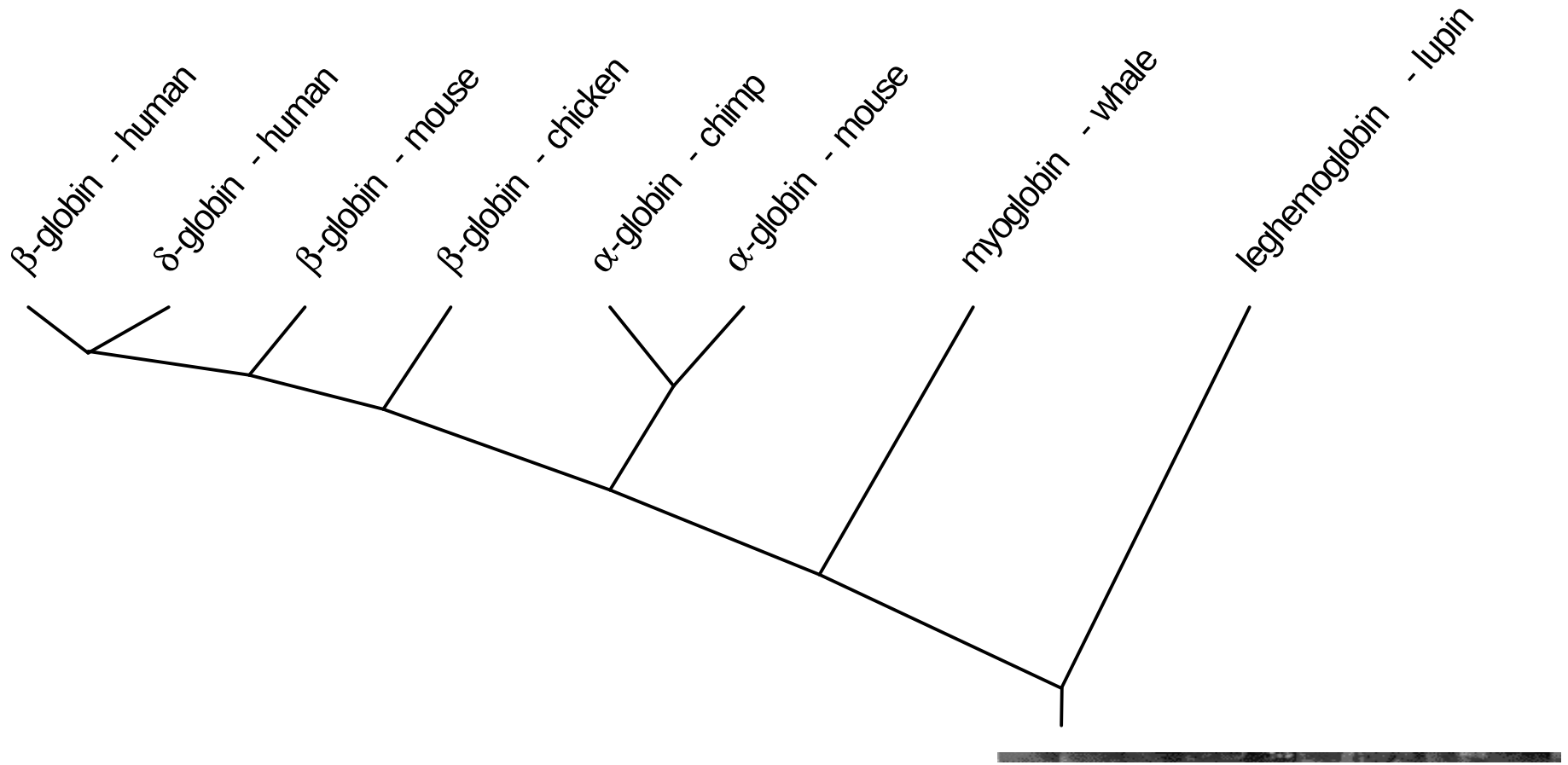

Figure 9–20 **Gene duplication by unequal crossing-over.** A pair of homologous chromosomes in a germ-line precursor cell undergoes a crossover (recombination) at a short sequence (*red*), which is repeated on each chromosome. In some cases, these sequences are believed to be transposons, which are present in many copies in the human genome. After the crossover, the long chromosome has two copies of the globin gene. Independent mutations in each of the genes, which could arise in subsequent generations, would lead to their diversification. (Since the short chromosome lacks the original globin gene, the individuals that inherit this chromosome would be expected to be lost from the population.)

# Similarity vs. homology

- Sequences are *similar* if they are sufficiently resembling at the sequence level (DNA, protein, …)
- Similarity can arise from
  - Homology
  - Convergence (functional constraints)
  - Chance

- Sequences are *homologous* if they arise from a common ancestor
  - Homologous sequences are *paralogous* if their differences involve a gene duplication event
  - Homologous sequences are *orthologous* iftheir differences are not related to a gene duplication

# Orthology vs. paralogy



# Phylogeny

- Relationships between genes and proteins can be inferred on the basis of their sequences
- Reconstruction of molecular evolution = phylogeny

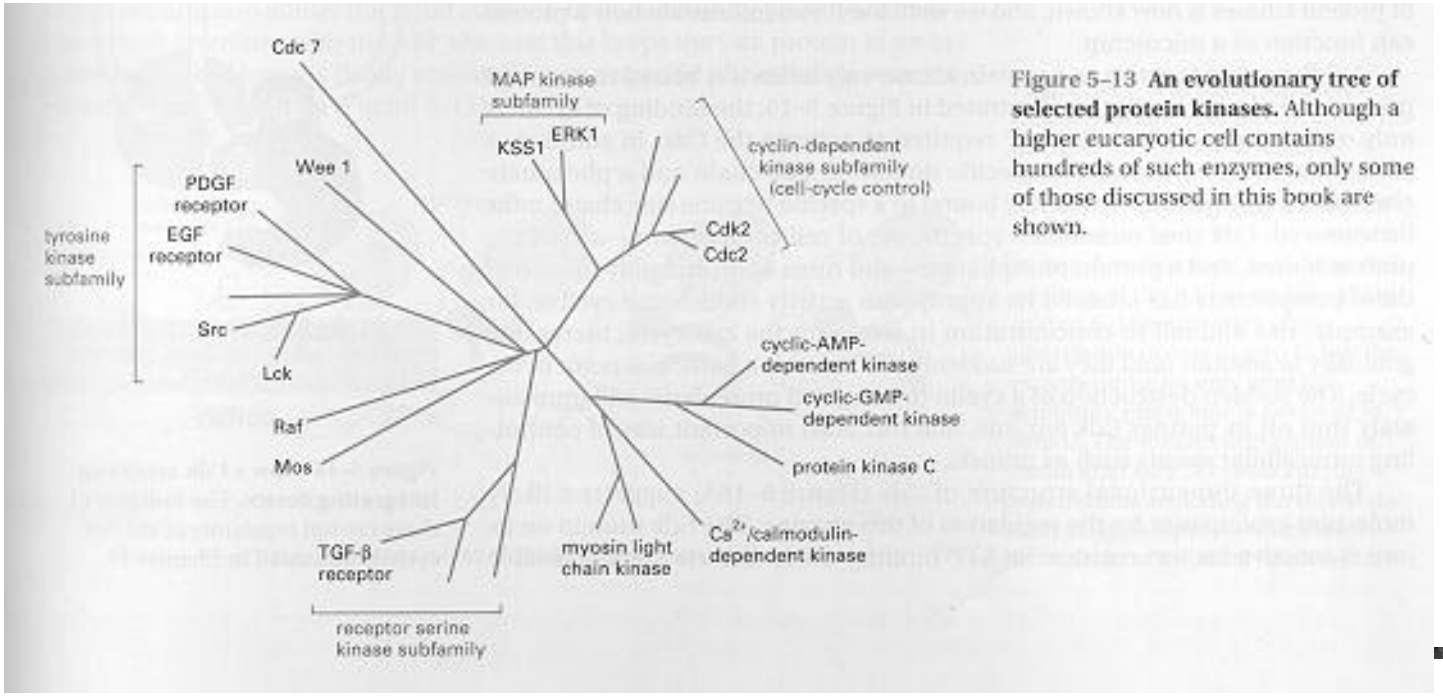



Figure 5–13 **An evolutionary tree of selected protein kinases.** Although a higher eucaryotic cell contains hundreds of such enzymes, only some of those discussed in this book are shown.

# Homology for the prediction of structure and function

- Homologous proteins have comparable structures
- Homologous proteins potentially have similar functions (ortholog: similar cellular role; paralog: similar biochemical function)

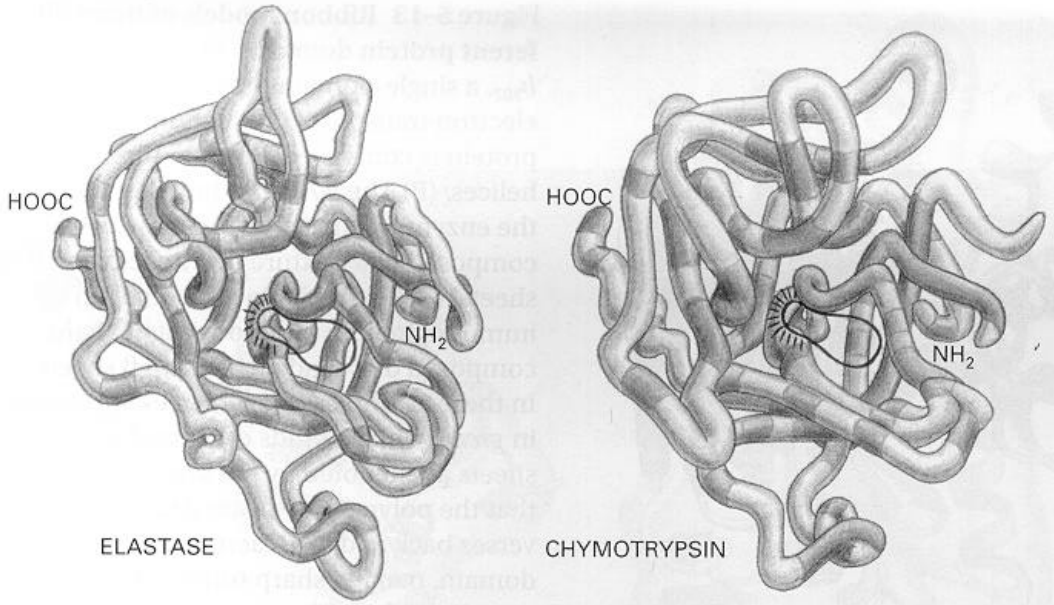



**Figure 5–14 Comparison of the conformations of two serine proteases.** The backbone conformations of elastase and chymotrypsin. Although only those amino acids in the polypeptide chain shaded in *green* are the same in the two proteins, the two conformations are very similar nearly everywhere. The active site of each enzyme is circled in *red;* this is where the peptide bonds of the proteins that serve as substrates are bound and cleaved by hydrolysis. The serine proteases derive their name from the amino acid serine, whose side chain is part of the active site of each enzyme and directly participates in the cleavage reaction.

# Homology for prediction with DNA

- Conserved regions arise from evolutionary pressure and are therefore functionally important
  - Genes
  - Control regions
- Comparative genomics
- Genes can be predicted by comparing genomes at an appropriate evolutionary distance (e.g., mouse and human)

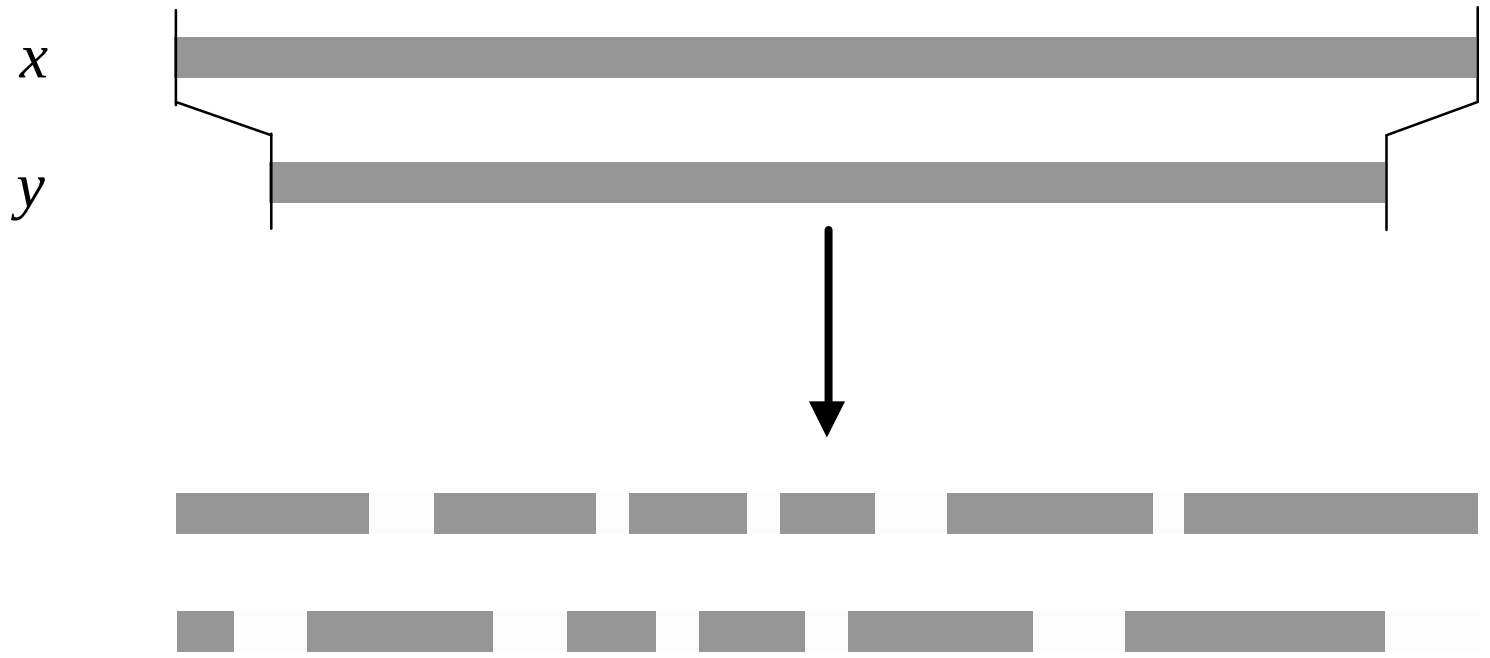
# Principles of pairwise alignment

# Elements of an alignment

- Types of  alignments
  - DNA vs. protein
  - Pairwise va. multiple alignment
  - Global alignment
  - Local alignment
- Scoring model for alignments
  - Substitutions
  - Gaps (insertions, deletions)
  - Substitution matrix and gap penalty
- Algorithm
  - Dynamic programming
  - Heuristic
- Significance evaluation

```
HEAGAWGHE-E
--P-AW-HEAE
```

# Global alignment



# Global alignment

- Alignment of ‘human alpha globin’ against ‘human beta globin’, ‘lupin leghemoglobin’ and ‘glutathionine S-transferase homolog F11G11.2’ (‘+’ for good substitutions)

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

Strong homology

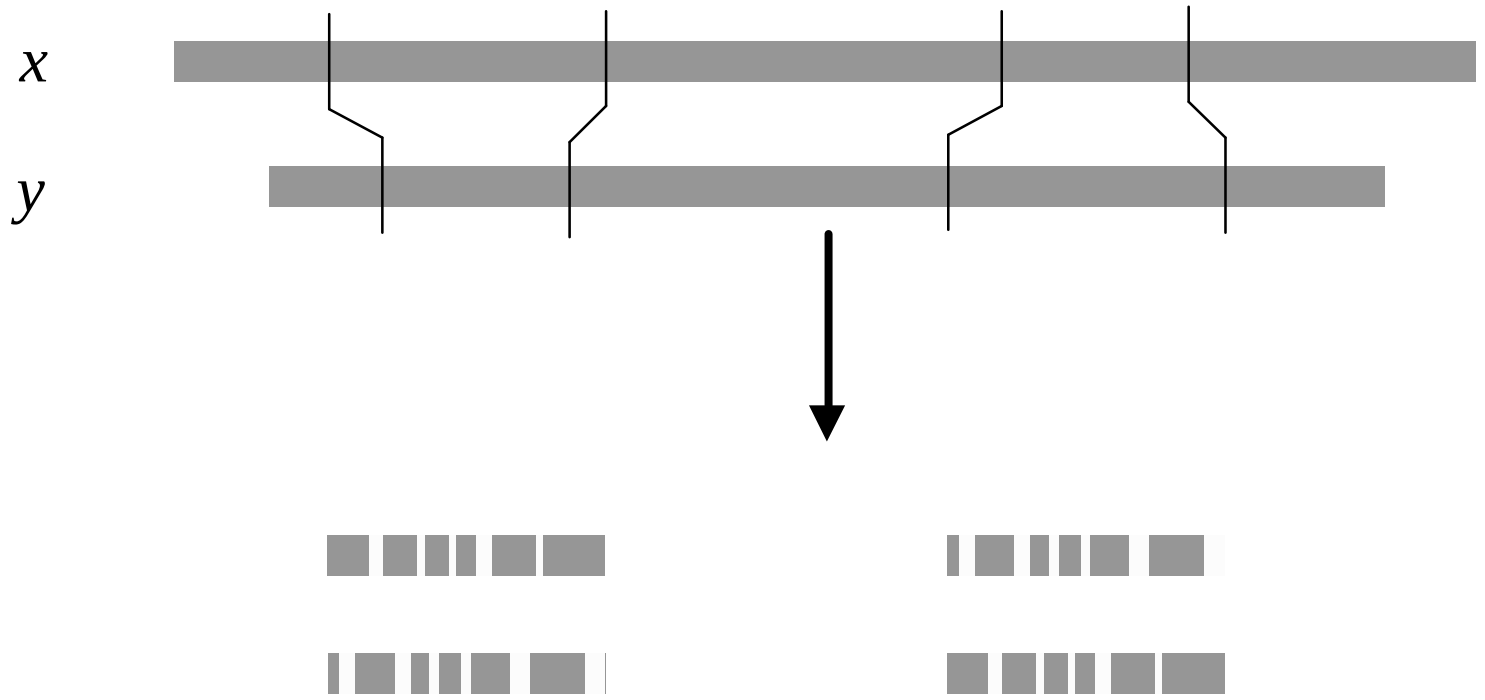
```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++           +L+ L+++H+ K
LGB2_LUPLU  NNPEFQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

Low similarity / structural homology

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +   +D L  ++ H+ D+  A +AL D     ++AH+
F11G11.2    GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Chance similarity

# Local alignment



# Substitution matrix and gap penalty

- The alignment of two residues can be more or less likely
- To compute the quality of an alignment, we assign a gain or a penalty to the alignment of two residues
- Gaps also have a penalty

```
HEAGAWGHE-E
--P-AW-HEAE
```

BLOSUM50 substitution matrix

| | A | R | N | D | C | Q | EGHILKMFPSTWYV |
|---|---|---|---|---|---|---|---|
| A | 5 | -2 | -1 | -2 | -1 | -1 | ..... |
| R | -2 | 7 | -1 | -2 | -4 | 1 | ..... |
| N | -1 | -1 | 7 | 2 | -2 | 0 | ..... |
| D | -2 | -2 | 2 | 8 | -4 | 0 | ..... |
| C | -1 | -4 | -2 | -4 | 13 | -3 | ..... |
| Q | -1 | 1 | 0 | 0 | -3 | 7 | ..... |
| ... | ... | ... | ... | ... | ... | ... | ..... |

# Substitution matrix for DNA

- Standard

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

# Dynamic programming

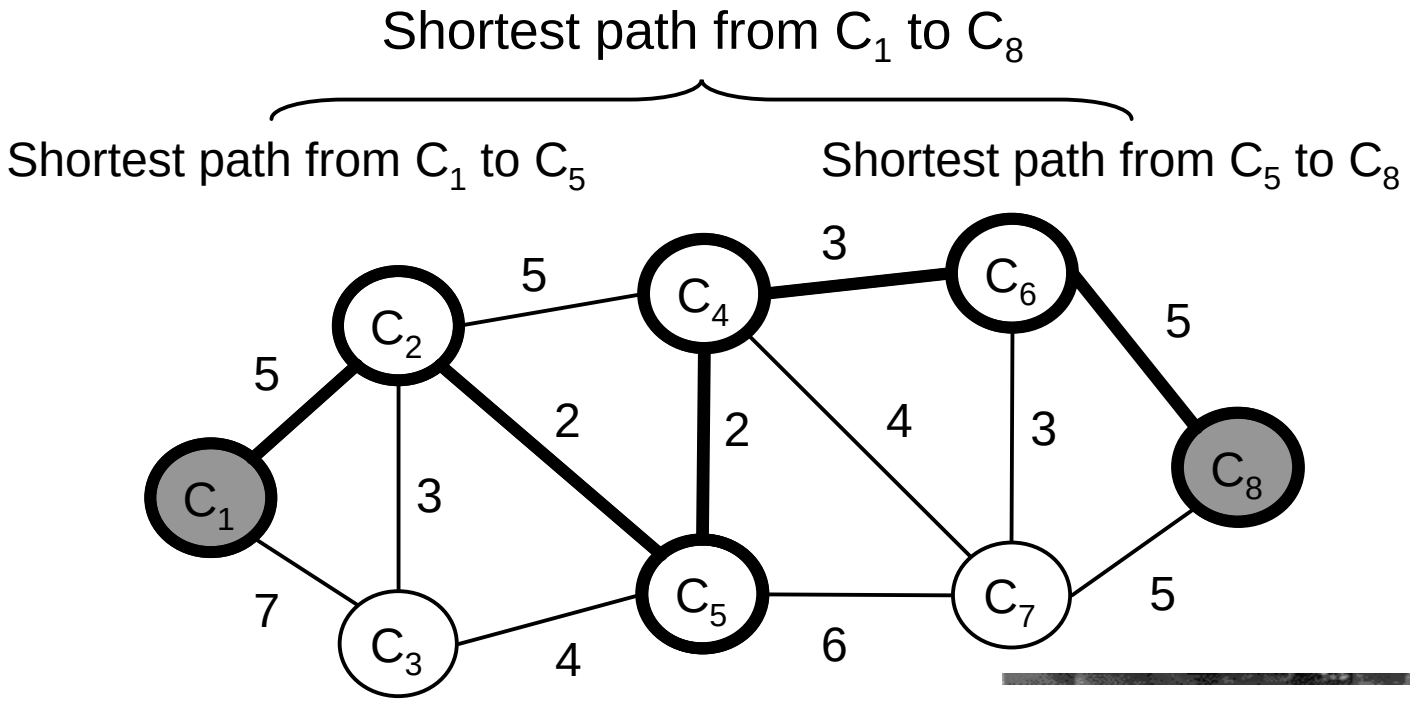
- To align is to find the minimum penalty / maximum score path through the penalty table = DYNAMIC PROGRAMMING
- Substitution matrix = BLOSUM 50
- Gap penalty = -8

```
HEAGAWGHE-E
--P-AW-HEAE
```

| | * | H | E | A | G | A | W | G | H | E | E |
|---|---|---|---|---|---|---|---|---|---|---|---|
| * | 0 | -8 | -8 | -8 | -8 | -8 | -8 | -8 | -8 | -8 | -8 |
| P | -8 | -2 | -1 | -1 | -2 | -1 | -4 | -2 | -2 | -1 | -1 |
| A | -8 | -2 | -1 | 5 | 0 | 5 | -3 | 0 | -2 | -1 | -1 |
| W | -8 | -3 | -3 | -3 | -3 | -3 | 15 | -3 | -3 | -3 | -3 |
| H | -8 | 10 | 0 | -2 | -2 | -2 | -3 | -2 | 10 | 0 | 0 |
| E | -8 | 0 | 6 | -1 | -3 | -1 | -3 | -3 | 0 | 6 | 6 |
| A | -8 | -2 | -1 | 5 | 0 | 5 | -3 | 0 | -2 | -1 | -1 |
| E | -8 | 0 | 6 | -1 | -3 | -1 | -3 | -3 | 0 | 6 | 6 |

# Dynamic programming

- Belman's optimality principle
- Example: finding the shortest train route between two cities



# Alignment algorithms

# Global alignment

- Needleman-Wunsch algorithm
- Progressively complete the table $F(i,j)$ (!!! column, row) that keeps track of the maximum score for the alignment of sequence $x$ up to $x_i$ to sequence $y$ up to $y_j$

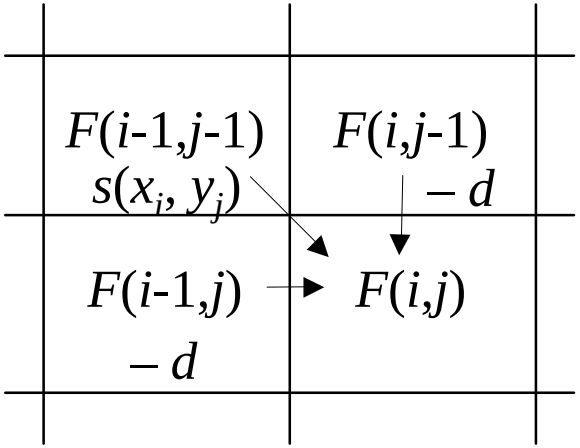
I G A $x_i$ A I G A $x_i$ G A $x_i$ - -

L G V $y_j$ G V $y_j$ - - S L G V $y_j$

- Substitution matrix $s(x, y)$ and gap penalty $d$
- Recurrence

$$F(i,j) = \max \begin{cases} F(i\text{-}1,j\text{-}1) + s(x_i, y_j) & \text{substitution} \\ F(i\text{-}1,j) - d & \text{deletion} \\ F(i,j\text{-}1) - d & \text{insertion} \end{cases}$$

$$F(0,0) = 0$$




- Start from to left
- Complete progressively by recurrence
- Use traceback pointers

$$F(i,j) = \max \begin{cases} F(i\text{-}1,j\text{-}1) + s(x_i, y_j) \\ F(i\text{-}1,j) - d \\ F(i,j\text{-}1) - d \end{cases}$$

| | * | H | E | A | G | A | W | G | H | E | E |
|---|---|---|---|---|---|---|---|---|---|---|---|
| * | 0 | -8 | -16 | -24 | -32 | -40 | -48 | -56 | -64 | -72 | -80 |
| P | -8 | -2 | -9 | -17 | -25 | -33 | -42 | -49 | -57 | -65 | -73 |
| A | -16 | -10 | -3 | -4 | -12 | -20 | -28 | -36 | -44 | -52 | -60 |
| W | -24 | -18 | -11 | -6 | -7 | -15 | -5 | -13 | -21 | -29 | -37 |
| H | -32 | -14 | -18 | -13 | -8 | -9 | -13 | -7 | -3 | -11 | -19 |
| E | -40 | -22 | -8 | -16 | -16 | -9 | -12 | -15 | -7 | 3 | -5 |
| A | -48 | -30 | -16 | -3 | -11 | -11 | -12 | -12 | -15 | -5 | 2 |
| E | -56 | -38 | -24 | -11 | -6 | -12 | -14 | -15 | -12 | -9 | 1 |

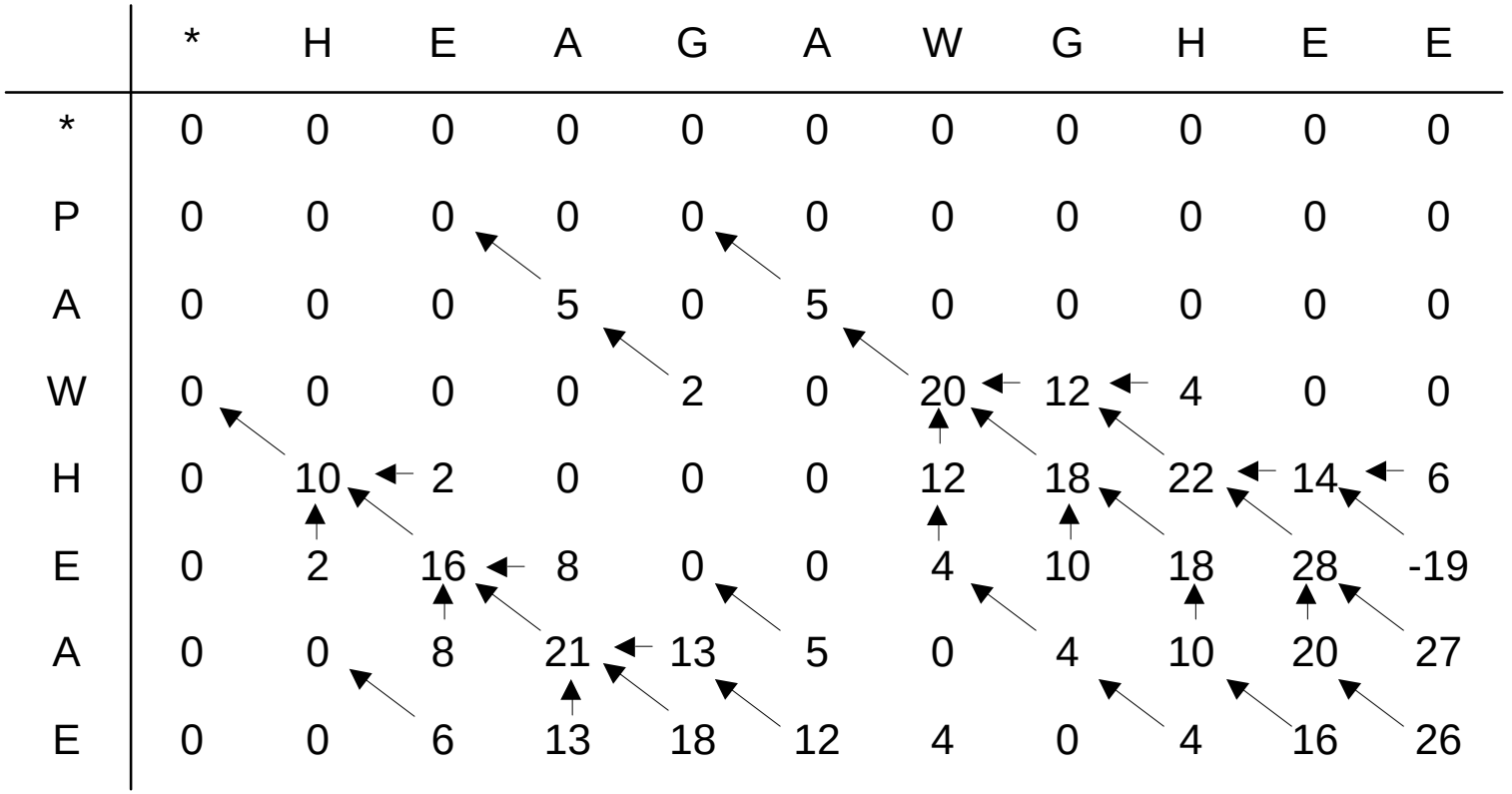
# Local alignment

- Smith-Waterman algorithm
- Best alignment between subsequences of $x$ and $y$
- If the current alignment has a negative score, it is better to start a new alignment

$$
F(i,j) = \max \begin{cases} 0 & \text{restart} \\ F(i\text{-}1,j\text{-}1) + s(x_i, y_j) & \text{substitution} \\ F(i\text{-}1,j) - d & \text{deletion} \\ F(i,j\text{-}1) - d & \text{insertion} \end{cases}
$$

$$
F(0,0) = 0
$$

- Start from top left
- Complete progressively by recurrence
- Traceback from the highest score and stop at zero




AWGHE
AW-HE

# Substitution matrices

- How can we choose a reasonable substitution matrix?
- Look at a set of confirmed alignments (with gaps) and compute the amino acid frequences $q_a$, the substitution frequences $p_{ab}$, and the gap function $f(g)$
- Likelihood model
  - Random sequences: $P(x,y|\mathrm{R}) = \Pi_i q_{xi} \Pi_j q_{yj}$
  - Alignment: $P(x,y|\mathrm{M}) = \Pi_i p_{xixj}$
  - Odds ratios: $P(x,y|\mathrm{M})/P(x,y|\mathrm{R}) = \Pi_i p_{xixj}/(\Pi_i q_{xi} \Pi_j q_{yj})$
  - Log-odds score: $S(x,y) = \Sigma_i s(x_i,y_i)$ with $s(a,b) = \log(p_{ab}/q_a q_b)$
- Substitution matrix $s(a,b) = \log(p_{ab}/q_a q_b)$

# PAM matrix

- Point Accepted Mutations matrix
- Problems
  - Alignments are not independent for related proteins
  - Different alignments correspond to different evolution times
- PAM1 matrix
  - Tree of protein families
  - Estimate ancestral sequences
  - Estimate mutations at short evolutionary distance
  - Scale to a substitution matrix
  - 1% Point Accepted Mutations (PAM1)
- PAM250 is 250% Point Accepted Mutations (~20% similarity) = 250ste power of PAM1

# BLOSUM matrix

- BLOCKS SUbstitution Matrix
- PAM does not work so well at large evolutionary distances
- Ungapped alignments of protein families from the BLOCKS database
- Group sequences with more than $L\%$ identical amino acids (e.g., BLOSUM62)
- Use the substitution frequency of amino acids between the different groups (with correction for the group size) to derive the substitution matrix

# BLAST

- For large databases, Smith-Waterman local alignment is too slow
- Basic Local Alignment Search Tool (BLAST) is a fast heuristic algorithm for local alignment (http://www.ncbi.nlm.nih.gov/Entrez)
  - BLASTP – protein query on protein database
  - BLASTN – nucleotide query on nucleotide database
  - BLASTX – translated nucleotide query on protein database (translation into the six reading frames)
  - TBLASTN – protein query on translated nucleotide db
  - TBLASTX – translated nucleotide query on translated nucleotide db

# BLASTP

- Step 1: Find all words of length $w$ (e.g., $w=3$) for which there is a match in the query sequence with score at least $T$ (e.g., $T=11$) for the chosen substitution matrix (e.g., BLOSUM62 with gap penalty $10+g$)

- Step 2: Use a finite state automaton to find all matches with the word list in the database (*hits*)

# BLASTP

- Step 3: Check which hits have another hit without overlap within a distance of $A$ (e.g., $A$=40) (the distance must be identical on the query and on the target) (*two-hits*)

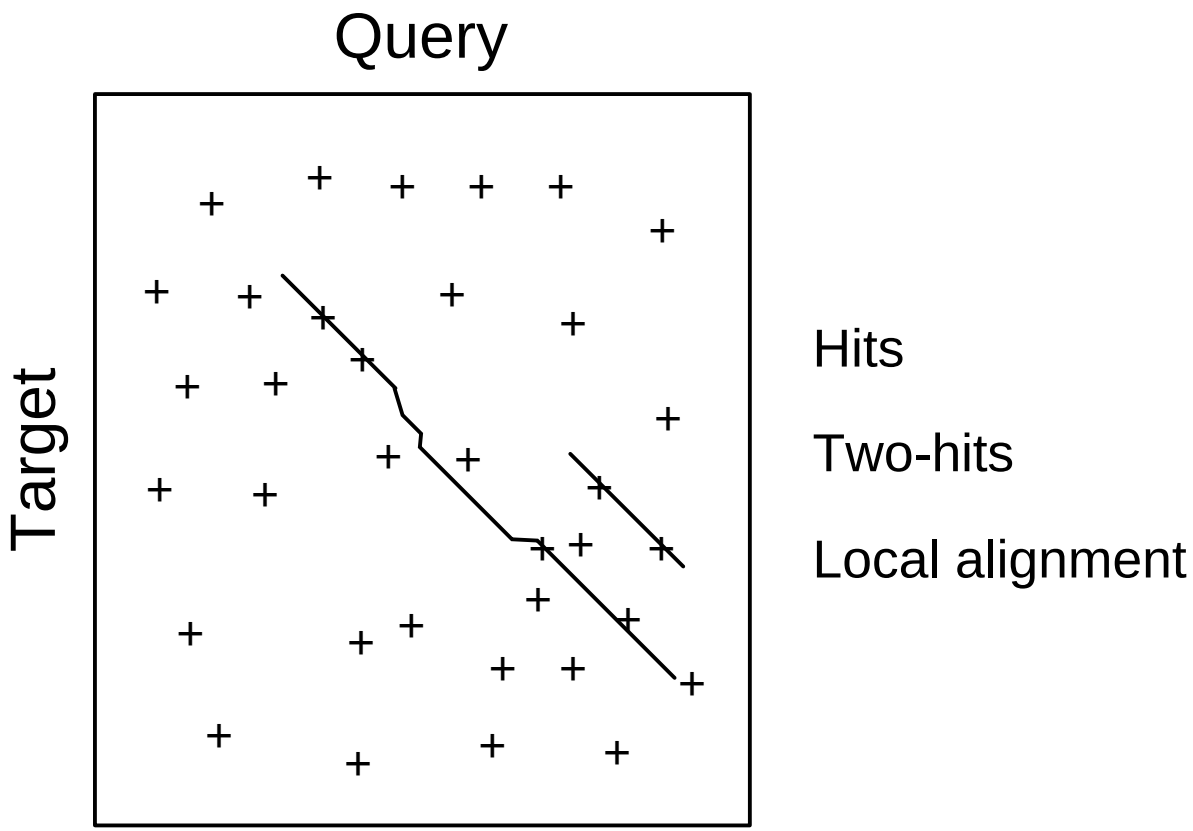
- Step 4: Extend the left hit of the two-hits in both directions by ungapped alignment ; stop the extension when the score drops by $X_g$ (e.g., $X_g$=40) under the best score so far (*high scoring segment pair* HSP)

# BLASTP

- Step 5: Extend the HSPs with normalized score above $S_g$ ($S_g$ =22 bits) by ungapped alignment ; stop the extension when the score drops by $X_g$ (e.g., $X_g$=40) under the best score so far ; select the best gapped local alignment

- Step 6: Compute the significance of the alignments ; for the significant alignments, repeat the gapped alignment with a higher dropoff parameter $X_g$ for more accuracy

# BLASTP



# BLASTP example

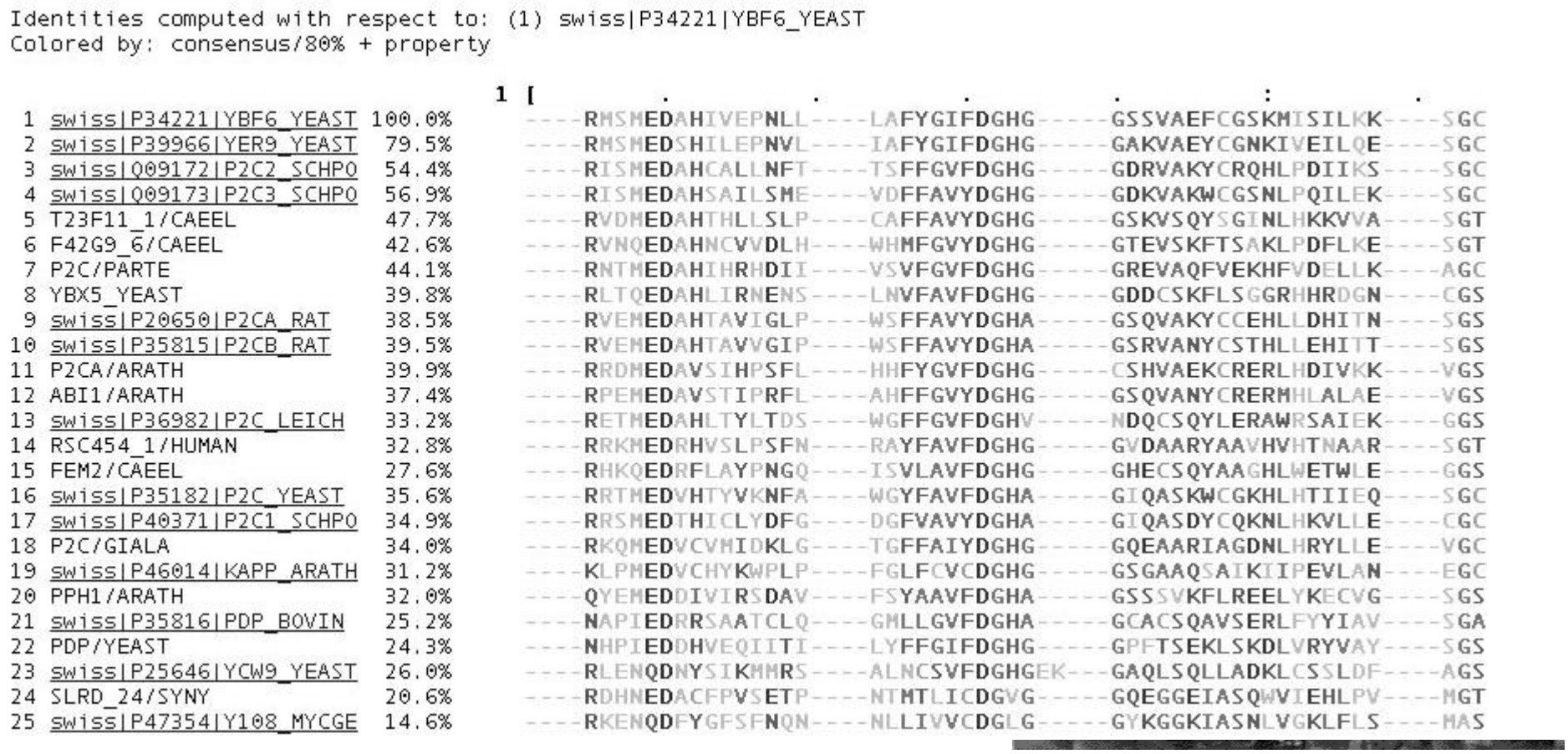
| Protein family | Query (SWISS-PROT) | Smith-Waterman | BLAST (# matches) |
|---|---|---|---|
| Serine protease | P00762 | 275 | 275 |
| Serine protease inhibitor | P01008 | 108 | 108 |
| Ras | P01111 | 255 | 252 |
| Globin | P02232 | 28 | 28 |
| Hemagglutinin | P03435 | 128 | 128 |
| Interferon alpha | P05013 | 53 | 53 |
| Alcohol dehydrogenase | P07327 | 138 | 137 |
| Histocompatibility antigen | P10318 | 262 | 261 |
| Cytochrome P450 | P10635 | 211 | 211 |
| Glutathione transferase | P14942 | 83 | 81 |
| $H^+$-transport ATP synthase | P20705 | 198 | 197 |
| | | | |
| Running time | | 36 | 0.34 |

# Multiple aligment

# Multiple alignment

- What to do if we want to compare together the sequences of several proteins from the same family (not pair by pair)?
- Multiple alignments are, for example, important in the study of the role played by the residues in the 3D structure of the proteins (ideally, aligned residues should have the same 3D coordinates and should correspond evolutionarily)
- Manually constructing a multiple alignment is tedious

# Multiple alignment



# Multiple alignment

- Elements of a multiple alignment method
  - Scoring model
  - Algorithm for finding the best aligment

- Ideally, given a phylogenetic tree for the sequences, the probability of a multiple alignment is the product of the probabilities of all the events needed for this tree (in practice, not enough data will be available for this)

# Scoring model

- A pragmatic scoring function is of the form

$$S(m) = G + \Sigma_i S(m_i)$$

  $G$ gap score, $m_i$ column $i$ of the multiple alignment, $S(m_i)$ score for column $i$

- Column score: sum of pairs score

$$S(m_i) = \Sigma_l \Sigma_{k<l} s(m_i^k, m_i^l)$$

  $s()$ substitution matrix

- Sum of pairs score not statistically well-founded
  - Extension of log-odds score (3 seqs.): $\log(p_{abc}/(q_a q_b q_c))$
  - Sum of pairs: $\log(p_{ab}/(q_a q_b)) + \log(p_{bc}/(q_b q_c)) + \log(p_{ac}/(q_a q_c))$

# PSI-BLAST

- Position-Specific Iterated BLAST
- Searching the database not with a single sequence but with a score matrix describing a protein family is much more sensitive
- Idea:
  - Run BLAST for the query sequence
  - Gather the significant alignments
  - Build a specific matrix describing the aligned sequences
  - Run BLAST using this specific matrix to recruit more seqs.
  - Build new matrix and iterate until no more seqs. are recruited)

# PSI-BLAST

- Motifs can be found in a protein family; if we search using such motifs we can detect more distant relationships
- It is difficult to delineate multiple motifs automatically in a set of proteins; therefore, PSI-BLAST builds a motif as long as the query ($L$x20 matrix)
- The motif is *not* a substitution matrix but gives *at each position* a specific score for the presence of an amino acid and is thus called a *position-specific scoring matrix* (PSSM) (gap scores are not position specific)

# Position-specific scoring matrix

MFGKRAFVHHYVGEGMEENEFTDARQDLYELEVDYANL
MFKRKGFLHWYTGEGMEPVEFSEAQSDLEDLILEYQQY
MFSRKAFLHWYTGEGMEEGDFAEADNNVSDLLSEYQQY
RGAFL

| | A | B | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | X | Y | Z | * | - |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RNAFL | -56 | -57 | -56 | -58 | 3 | -54 | -61 | -51 | -48 | -47 | -46 | 35 | -56 | -62 | 10 | 22 | -58 | -55 | -51 | -58 | -42 | -58 | 7 | 0 | 0 |
| | 2 | 7 | -36 | -1 | 0 | 24 | -9 | -32 | -38 | -40 | -38 | -38 | 18 | -47 | -38 | -41 | 0 | -37 | -39 | -32 | -19 | 19 | -21 | 0 | 0 |
| RSKRA | 20 | -46 | -39 | -47 | -44 | 10 | -8 | -40 | -45 | 8 | -46 | -43 | -44 | -48 | -41 | 16 | -5 | -41 | -18 | -44 | -22 | 11 | -42 | 0 | 0 |
| | -44 | -45 | -44 | -46 | -42 | 20 | -48 | -38 | -41 | 19 | 0 | -38 | -42 | -50 | -37 | 22 | -45 | -44 | -43 | -41 | -28 | 4 | -39 | 0 | 0 |
| RNAFL | -1 | -6 | -41 | 6 | -36 | -40 | -43 | -37 | -36 | 20 | 8 | 14 | -23 | -45 | -19 | 17 | -40 | -40 | -38 | -47 | -21 | -43 | -27 | 0 | 0 |
| | 21 | 8 | -44 | 3 | 14 | 6 | -14 | -45 | -51 | -46 | -52 | -50 | 13 | -4 | -44 | -26 | -42 | -45 | -47 | -53 | -25 | -48 | -19 | 0 | 0 |
| RDAYM | -48 | -29 | -44 | -51 | -48 | 29 | -53 | -42 | 8 | -47 | -41 | -41 | -2 | -57 | 18 | -49 | 1 | -47 | -42 | -40 | -31 | 13 | -11 | 0 | 0 |
| | -36 | -16 | -32 | 3 | -40 | 17 | -44 | -34 | -22 | -37 | 12 | 14 | -40 | -46 | -38 | 7 | -40 | -35 | 14 | -31 | -20 | 12 | -39 | 0 | 0 |
| ENYKK | -50 | -46 | -53 | -49 | -45 | 8 | -53 | 36 | -54 | 17 | -53 | -49 | -43 | -55 | -1 | 11 | -49 | -50 | -54 | -51 | -37 | -39 | -21 | 0 | 0 |
| | -42 | -45 | -42 | -47 | -43 | 16 | -46 | 5 | -44 | 11 | -43 | -40 | -43 | -50 | 2 | 3 | 1 | -43 | -44 | 42 | -28 | -27 | -18 | 0 | 0 |
| RNAFL | -50 | -48 | -51 | -46 | 18 | -19 | -56 | -41 | -53 | -50 | -54 | -53 | -51 | 7 | -47 | -52 | 1 | -3 | -53 | -41 | -33 | 32 | -18 | 0 | 0 |
| | 9 | -12 | -33 | 7 | -37 | -40 | -40 | -40 | -30 | 8 | -35 | 16 | -37 | -44 | -38 | -37 | -7 | 18 | 14 | -45 | -19 | -43 | -38 | 0 | 0 |
| RNAFM | -41 | -4 | -41 | 5 | -46 | 10 | 22 | -46 | 15 | -47 | -24 | 18 | -15 | -6 | -47 | -48 | -14 | -44 | -35 | -46 | -22 | -41 | -47 | 0 | 0 |
| | 2 | -20 | -51 | 3 | 24 | 22 | 2 | -47 | -52 | -48 | -51 | -51 | -49 | -56 | -44 | -51 | -49 | -51 | -51 | -50 | -32 | -41 | -14 | 0 | 0 |
| MYAKR | -36 | 6 | -43 | 7 | 6 | -48 | 23 | -36 | -50 | 6 | -49 | -43 | 5 | -46 | 15 | -36 | -37 | -40 | -47 | -47 | -25 | -47 | 11 | 0 | 0 |
| | -55 | 14 | -54 | 18 | -52 | 6 | 2 | -54 | -47 | -6 | -45 | 35 | 8 | -63 | -53 | -55 | -56 | -54 | -50 | -57 | -35 | -53 | -52 | 0 | 0 |
| MYSKR | -40 | 22 | -44 | 24 | 17 | -50 | -9 | -36 | -50 | -38 | -52 | -48 | 19 | -47 | -35 | -41 | 0 | -38 | -48 | -53 | -29 | -48 | -12 | 0 | 0 |
| | -52 | -49 | -52 | -45 | 23 | 15 | -59 | -50 | -18 | -51 | 16 | -42 | -54 | -10 | -46 | -53 | -54 | -53 | -47 | -53 | -34 | -46 | -16 | 0 | 0 |
| MYAKR | -35 | 13 | -37 | 18 | 3 | -40 | 13 | -36 | -35 | -37 | -20 | 24 | 8 | -44 | -35 | -39 | -1 | 4 | -11 | -44 | -17 | -43 | -19 | 0 | 0 |
| | -73 | -37 | -82 | -10 | 31 | -82 | -78 | -69 | -78 | -68 | -78 | 13 | -71 | -78 | -62 | -72 | -72 | -74 | -76 | -82 | -62 | -81 | -21 | 0 | 0 |
| MYAKR | -63 | -70 | -55 | -71 | -71 | 34 | -69 | -59 | -50 | -66 | 1 | 24 | -69 | -71 | -65 | -66 | -66 | -62 | -54 | -50 | -51 | -40 | -68 | 0 | 0 |
| | -3 | -2 | -41 | 25 | 3 | -50 | -42 | -42 | -48 | -42 | -51 | -47 | -35 | -48 | -42 | -45 | 10 | 17 | -44 | -54 | -28 | -49 | -22 | 0 | 0 |
| MYAKR | -51 | 3 | -57 | 8 | 27 | -62 | -56 | -49 | -5 | -49 | -61 | -57 | -4 | -58 | -45 | -53 | 9 | -6 | -55 | -64 | -34 | -60 | -13 | 0 | 0 |
| | 25 | -57 | -48 | -58 | -56 | -61 | -51 | -57 | -58 | -56 | -60 | -58 | -55 | -58 | -56 | -58 | 18 | -49 | -17 | -64 | -42 | -61 | -56 | 0 | 0 |
| MYSKR | -49 | -18 | -55 | 3 | 14 | -55 | -53 | -43 | -53 | 4 | -52 | 13 | -46 | -55 | -8 | 28 | -49 | -50 | -52 | -56 | -34 | -54 | 2 | 0 | 0 |
| | 0 | 3 | -44 | 11 | 22 | -50 | -23 | 13 | -50 | -39 | -51 | -47 | -7 | -47 | -2 | -41 | 13 | -38 | -46 | -52 | -24 | -46 | 9 | 0 | 0 |
| | -44 | 18 | -40 | 17 | -41 | -44 | -46 | -41 | 20 | -45 | -39 | -38 | 20 | -52 | -45 | -46 | -42 | 7 | 5 | -52 | -28 | -47 | -43 | 0 | 0 |
| | -41 | -51 | -36 | -53 | -50 | -37 | -52 | -48 | 4 | -48 | 9 | 27 | -50 | -52 | -46 | -48 | -45 | 5 | 21 | -45 | -30 | -44 | -47 | 0 | 0 |
| | 11 | 10 | -39 | 3 | 3 | -43 | -39 | -32 | -44 | -3 | -22 | 3 | 18 | -45 | 23 | -36 | -12 | -37 | -41 | -44 | -18 | 1 | 14 | 0 | 0 |
| | 10 | 17 | 16 | 24 | -5 | -47 | -38 | -37 | -45 | -39 | -48 | -44 | 7 | -45 | 6 | -41 | 8 | -36 | -20 | -50 | -23 | -46 | 1 | 0 | 0 |

# BLAST with PSSMs

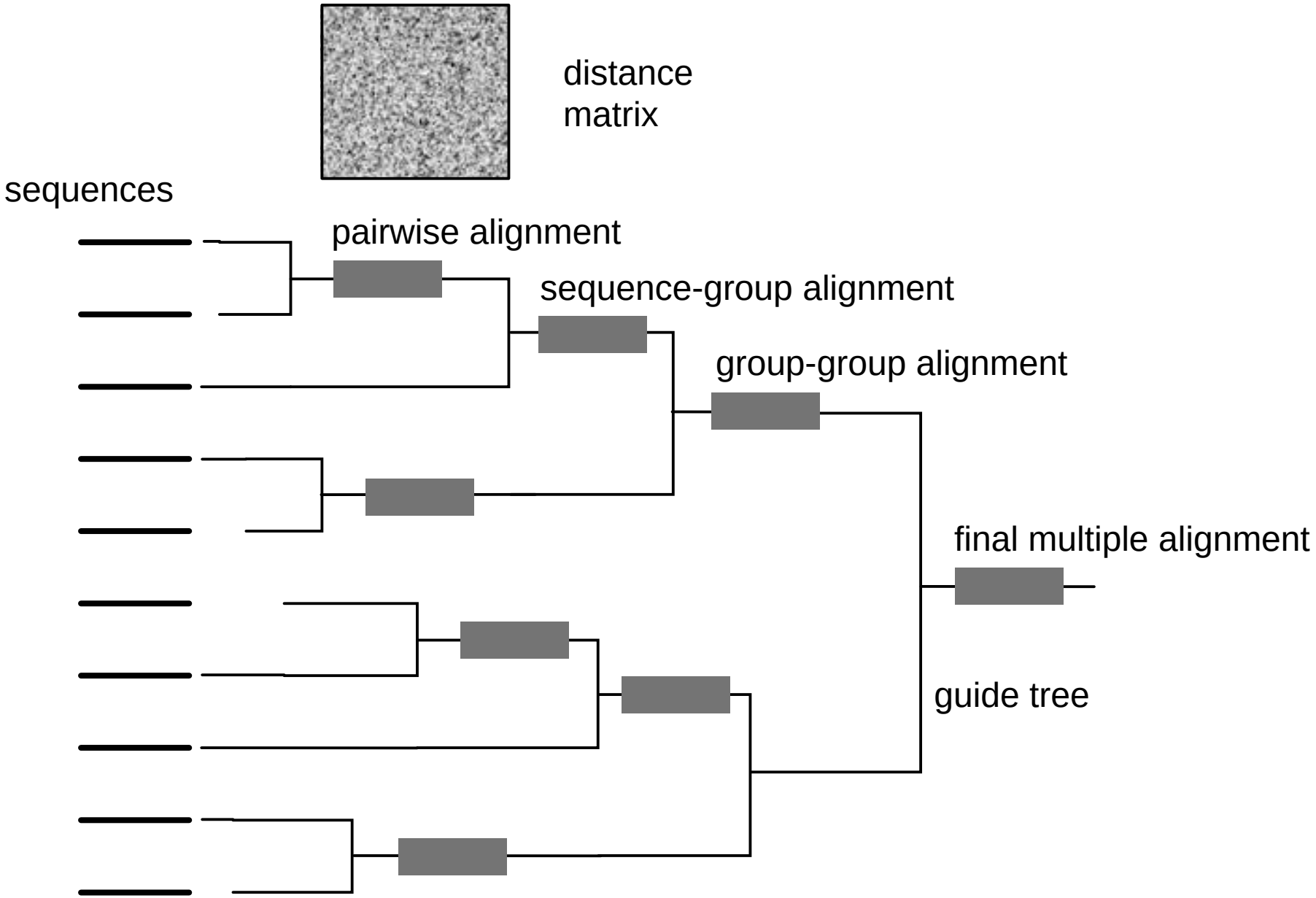
- Only minor modifications necessary to run BLAST with a position-specific scoring matrix (essentially, replace the score provided by the query and the substitution matrix $s(t_i,q_i)$ by the score provided by the PSSM $s(t_i,C)$)
- To avoid having to modify the parameters of BLAST, the scores must be rescaled appropriately

# PSI-BLAST example

| Protein family | Query (SWISS-PROT) | Smith-Waterman | BLAST | PSI-BLAST (# matches) |
|---|---|---|---|---|
| Serine protease | P00762 | 275 | 275 | 286 |
| Serine protease inhibitor | P01008 | 108 | 108 | 111 |
| Ras | P01111 | 255 | 252 | 375 |
| Globin | P02232 | 28 | 28 | 623 |
| Hemagglutinin | P03435 | 128 | 128 | 130 |
| Interferon alpha | P05013 | 53 | 53 | 53 |
| Alcohol dehydrogenase | P07327 | 138 | 137 | 160 |
| Histocompatibility antigen | P10318 | 262 | 261 | 338 |
| Cytochrome P450 | P10635 | 211 | 211 | 224 |
| Glutathione transferase | P14942 | 83 | 81 | 142 |
| $H^+$-transport ATP synthase | P20705 | 198 | 197 | 207 |
| Running time | | 36 | 0.34 | 0.87 |

# CLUSTALW

- Multiple alignment can be seen as multidimensional dynamic programming; however, for $N$ sequences of length $L$, the complexity is $O(L^N)$ and therefore infeasible in practice
- CLUSTALW (Cluster and ALign) is an example of progressive multiple alignment
  - Build a guide tree based on sequence similarity (not a real phylogenetic tree)
  - Do pairwise alignment going back up the guide tree
    - Pairwise sequence alignment
    - Sequence-group alignment
    - Group-group alignment


distance
matrix
sequences
pairwise alignment
sequence-group alignment
group-group alignment
final multiple alignment
guide tree

# CLUSTALW

- Algorithm
  - Construct distance matrix of all pairs
    - Perform pairwise gapped alignment of all pairs
    - Score as percentage differences in aligned sequences (gaps excluded)
  - Build the guide tree
    - Use the neighbor-joining method
    - Place the root on the edge where the mean branch length is equal on both sides
    - Use the three to derive a weight for each sequence
  - Progressively align along the tree (seq.-seq., seq.-prof., prof.-prof.)
    - Lock already aligned sequences together
    - Align two groups using sum-of-pairs score for dynamic programming
  - Use several heuristics

# Summary

- Similarity vs. homology
- Dynamic programming
- Local alignment
  - Needleman – Wunsch
- Global alignment
  - Smith – Waterman
  - Affine gap penalty
- Substitution matrices
  - PAM
  - BLOSUM
- Significance evaluation

- BLAST
- PHI-BLAST

- Multiple alignment
- PSI-BLAST
- CLUSTALW