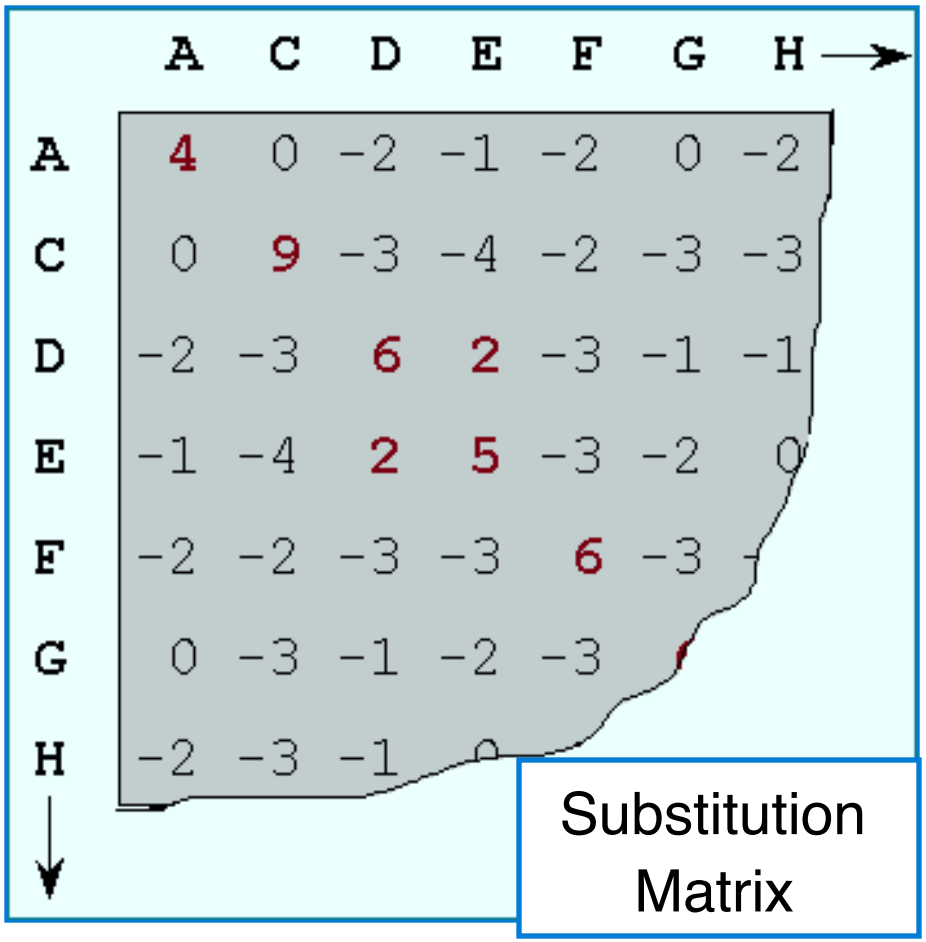
# Log-odds into Matrix




- **Two most used matrices are:**
  - **–PAM (Percent Accepted Mutation)**
    - **–Based on explicit evolutionary model**
    - **–Represents a specific evolutionary distance**
  - **–BLOSUM (BLOck SUbstitution Matrices)**
    - **–Based on empirical frequencies**
    - **–Always a blend of distances as seen in protein databases**

# PAM (Percent Accepted Mutations)
# - (# of mutations out of 100 AAs)

  - Based on explicit evolutionary model
  - PAM-1 is a scoring system for sequences in which 1% of the residues have undergone mutation: 1 pair in a 100 residue segment
  - PAM-250 represents 250% mutation, i.e., an average of 2.5 accepted mutation per residue (multiple mutations per pair) a very distant relationship
- Important to remember:
  - A value less than 0 or greater than 0 indicates that the frequency is less than or greater than that expected by chance, respectively
  - Smaller-numbered matrices correspond to closely related sequences
  - Larger-numbered matrices correspond to more distantly related sequences

# Scoring Matrices: PAM (Point Accepted Mutation)

§ PAM method: Align pairs of sequences with high identity (~99%)

- Examine large number of pairs of sequences from many protein families

§ Count the number of times amino acid ***i*** substitutes for amino acid ***j*** and vice-versa

§ Calculate the log odds ratio of the number of times ($M_{ij}$) amino acid ***j*** was found in place of ***i***, divided by the frequency ($f_j$) of ***j*** in the protein set:

$$\log(\text{odds ratio}) = \log(M_{ij}/f_j)$$

- $M_{ij}/f_j = 1$: Same as chance, log odds = 0
- $M_{ij}/f_j > 1$: Higher than chance, log odds is positive
- $M_{ij}/f_j < 1$: Lower than chance, log odds is negative

§ Typically assume that ***i*** to ***j*** is the same as ***j*** to ***i***, so data is merged and only half of matrix is presented

# Problems with PAM

- **PAM** model assumes all residues are equally mutable **(mutation for A←→G,T←→C are more likely than A←→T, G←→C )**
- Model devised using the most mutable positions rather than the most conserved positions, i.e., those that reflect chemical and structural properties of importance
- Derived from a biased set of sequences: small globular proteins available in the database in 1978

# BLOSUM (# -- use seqs less than #% identical)

- Important to remember:
  - matrices constructed using multiple alignments of evolutionarily divergent but highly conserved proteins
  - Every possible identity or substitution is assigned a score based on its observed frequences in the alignment of related proteins.
  - Pairs that are more likely than chance will have positive scores, and those less likely will have negative scores
  - Larger-numbered matrices correspond to more recent (less) divergence
  - Smaller-numbered matrices correspond to more distantly related sequences
  - By default, BLOSUM62 is often used

# Scoring Matrices: BLOSUM

§ The BLOSUM matrices were constructed in a way similar to PAM (Henikoff and Henikoff, 1992)

§ The difference was in the method used to estimate the substitution frequencies [BLOSUM = Block Substitution Matrix]

§ For BLOSUM matrices, the data for the substitution frequencies were taken from the BLOCKS database (highly conserved regions of proteins)

§ BLOSUM matrices generated from direct observation, not extrapolation

§ The matrix number refers to the minimum level of identity the sequences may have and still contribute independently to the model

- BLOSUM62 means that all sequences have 62% sequence identity

# Differences between PAM and BLOSUM

- 1. PAM matrices are based on an explicit evolutionary model (i.e. replacements are counted on the branches of a phylogenetic tree), whereas the BLOSUM matrices are based on an implicit model of evolution.

- 2. The PAM matrices are based on mutations observed throughout a global alignment, this includes both highly conserved and highly mutable regions. The BLOSUM matrices are based only on highly conserved regions in series of alignments forbidden to contain gaps.

# Differences between PAM and BLOSUM

- 3. The method used to count the replacements is different: unlike the PAM matrix, the BLOSUM procedure uses groups of sequences within which not all mutations are counted the same.

- 4. Higher numbers in the PAM matrix naming scheme denote larger evolutionary distance, while larger numbers in the BLOSUM matrix naming scheme denote higher sequence similarity and therefore smaller evolutionary distance. Example: PAM150 is used for more distant sequences than PAM100; BLOSUM62 is used for closer sequences than Blosum50.

# Two major scoring matrices

- PAM = accepted point mutation
  - 71 trees with 1572 accepted mutations, sequences with >85% identity
  - PAM1 means average of 1% change over all amino acids
  - 1 PAM = 10mer evolutionary distance

- BLOSUM = Blocks substitution matrices
  - Based on BLOCKS database (Henikoff & Henikoff, 1992) of over 2000 conserved amino acid patterns in over 500 proteins

# Example: PAM Matrices

- Point Accepted Mutations
  - For Amino Acid alignment
  - Biased by differing rates of mutation in different protein families
- Train on evolutionary alignments
- Develop Mutation Probability Matrix (MPM) that:
  - has likelihood of sequence b being replaced by a on off-diagonal
  - Has no residue change on diagonal

# Example PAM Substitution matrix

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

# Example Substitution Matrix for scoring: BLOSUM Matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

Figure 9: BLOSUM62 substitution matrix

# Alignment Summary

- Distance rewards matches and penalizes substitutions/gaps
- Calculate all paths that create an alignment
- Find the optimal path for the alignment
- Applications:  Find similar sequences (whether different, shortened elongated)

  Phylogenetic Trees (evolution)