Scoring Matrices

Shifra Ben-Dor

Irit Orr

Scoring matrices

- Sequence alignment and database searching programs compare sequences to each other as a series of characters.
- All algorithms (programs) for comparison rely on some scoring scheme for that.
- Scoring matrices are used to assign a score to each comparison of a pair of characters.

Scoring matrices

- The scores in the matrix are integer values.
- In most cases a positive score is given to identical or similar character pairs, and a negative or zero score to dissimilar character pairs.

Different types of matrices

- Identity scoring - the simplest scoring scheme, where characters are classified as: identical (scores 1) , or non-identical (scores 0). This scoring scheme is not much used.
- DNA scoring - consider changes as transitions and transversions. This matrix scores identical bp 3, transitions 2, and transversions 0.

Different types of matrices

- Chemical similarity scoring (for proteins) - this matrix gives greater weight to amino acids with similar chemical properties (e.g size, shape or charge of the aa).
- Observed matrices for proteins - most commonly used by all programs. These matrices are constructed by analyzing the substitution frequencies seen in the alignments of known families of proteins.

Observed Scoring Matrices

- Every possible identity and substitution is assigned a score.
- This score is based on the observed frequencies of such occurrences in alignments of evolutionary related proteins.
- This score will also reflect the frequency that a particular amino acid occurs in nature, as some amino acids are more abundant than others.

Observed Scoring Matrices

- Identities are assigned the most positive scores
- Frequently observed substitutions also receive positive scores
- Mismatches , or matches that are unlikely to have been a result of evolution, are given negative scores.

- Each matrix entry gives the ratio of the observed frequency of substitution between each possible pair of amino acids in related proteins to that expected by chance, given the frequencies of amino acids in proteins.
- These ratios are called odds scores.
- These ratios are transformed to logarithms of odds scores called log odds scores.
- Odds scores and log odds scores are used to score protein alignments

Different types of matrices

- Observed Scoring Matrices are superior to simple identity scores, or scores based solely on chemical propensities of the amino
- The most frequently used observed log odds matrices used are the PAM and BLOSUM matrices.

PAM Matrices

- Developed by Margaret Dayhoff and co-workers.
- Derived from global alignments of very similar sequences (at least 85% identity), so that there would be little likelihood of an observed change being the result of several successive mutations, but it should reflect one mutation only.
- PAM - Point Accepted Mutations.

- An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes:

- the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein
- the second is the acceptance of the mutation by the species as the new predominant form. To be accepted, the new amino acid usually must function in a way similar to the old one: chemical and physical similarities are found between the amino acids that are observed to interchange frequently.

- Dayhoff estimated mutation rates from substitutions observed in closely related proteins and extrapolated those rates to model distant relationships.
- PAM gives the probability that a given amino acid will be replaced by any other particular amino acid after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids.

- When used for protein comparison, the mutation probability (odds) matrix is normalized and the logarithm is taken. (this lets us add the scores along a protein instead of multiplying the probabilities)
- The resulting matrix is the “log-odds” matrix, known as the PAM matrix.

PAM#=Point Accepted Mutations / 100 bases

- The number with the matrix (PAM120, PAM90), refers to the evolutionary distance. Greater numbers are greater distances.
- To derive PAM250 you multiply PAM1 250 times itself
- PAM250 is the matrix derived of sequences with 250 PAMs.

PAM250

- At this evolutionary distance, only one amino acid in five remains unchanged.
- However, the amino acids vary greatly in their mutability; 55% of the tryptophans. 52% of the cysteines and 27% of the glycines would still be unchanged, but only 6% of the highly mutable asparagines would remain. Several other amino acids, particularly alanine, aspartic acid, glutamic acid, glycine, lysine, and serine are more likely to occur in place of an original asparagine than asparagine itself at this evolutionary distance!

PAM250 MATRIX

```
# This matrix was produced by "pam" Version 1.0.6 [28-Jul-93]
#
# PAM 250 substitution matrix, scale = ln(2)/3 = 0.231049
#
# Expected score = -0.844, Entropy = 0.354 bits
#
# Lowest score = -8, Highest score = 17
#
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  2 -2  0  0 -2  0  0  1 -1 -1 -2 -1 -1 -3  1  1  1 -6 -3  0  0  0  0 -8
R -2  6  0 -1 -4  1 -1 -3  2 -2 -3  3  0 -4  0  0 -1  2 -4 -2 -1  0 -1 -8
N  0  0  2  2 -4  1  1  0  2 -2 -3  1 -2 -3  0  1  0 -4 -2 -2  2  1  0 -8
D  0 -1  2  4 -5  2  3  1  1 -2 -4  0 -3 -6 -1  0  0 -7 -4 -2  3  3 -1 -8
C -2 -4 -4 -5 12 -5 -5 -3 -3 -2 -6 -5 -5 -4 -3  0 -2 -8  0 -2 -4 -5 -3 -8
Q  0  1  1  2 -5  4  2 -1  3 -2 -2  1 -1 -5  0 -1 -1 -5 -4 -2  1  3 -1 -8
E  0 -1  1  3 -5  2  4  0  1 -2 -3  0 -2 -5 -1  0  0 -7 -4 -2  3  3 -1 -8
G  1 -3  0  1 -3 -1  0  5 -2 -3 -4 -2 -3 -5  0  1  0 -7 -5 -1  0  0 -1 -8
H -1  2  2  1 -3  3  1 -2  6 -2 -2  0 -2 -2  0 -1 -1 -3  0 -2  1  2 -1 -8
I -1 -2 -2 -2 -2 -2 -2 -3 -2  5  2 -2  2  1 -2 -1  0 -5 -1  4 -2 -2 -1 -8
L -2 -3 -3 -4 -6 -2 -3 -4 -2  2  6 -3  4  2 -3 -3 -2 -2 -1  2 -3 -3 -1 -8
K -1  3  1  0 -5  1  0 -2  0 -2 -3  5  0 -5 -1  0  0 -3 -4 -2  1  0 -1 -8
M -1  0 -2 -3 -5 -1 -2 -3 -2  2  4  0  6  0 -2 -2 -1 -4 -2  2 -2 -2 -1 -8
F -3 -4 -3 -6 -4 -5 -5 -5 -2  1  2 -5  0  9 -5 -3 -3  0  7 -1 -4 -5 -2 -8
P  1  0  0 -1 -3  0 -1  0  0 -2 -3 -1 -2 -5  6  1  0 -6 -5 -1 -1  0 -1 -8
S  1  0  1  0  0 -1  0  1 -1 -1 -3  0 -2 -3  1  2  1 -2 -3 -1  0  0  0 -8
T  1 -1  0  0 -2 -1  0  0 -1  0 -2  0 -1 -3  0  1  3 -5 -3  0  0 -1  0 -8
```

Pet91 - an updated Dayhoff matrix

- Since the family of PAM matrices were derived from a comparatively small number of families, many of the possible mutations were not observed.
- Jones et al. have derived an updated matrix by examining a very large number of families, and created the PET91 scoring matrix.

Gonnet Matrices

- Another improvement on the PAM matrices
- Done when the database was larger
- All-against-all pairwise comparison to see all possible substitutions
- Optimized to find sequences that are further apart (as opposed to PAM, that started with very similar sequences)

BLOSUM Matrices

- Created by Henikoff & Henikoff, based on local multiple alignments of more distantly related sequences.
- First, multiple alignments of short regions (without gaps) of related sequences were gathered.
- In each alignment the sequences similar at some threshold value of percent identity were clustered into groups and averaged.

BLOSUM Matrices

- Substitution frequencies for all pairs of amino acids were calculated between the groups, this was used to create the log-odds BLOSUM (Block Substitution Matrix).

BLOSUM# - where # is the threshold identity percentage of the sequences clustered in those blocks.

- Thus, BLOSUM62 means that the sequences clustered in this block are at least 62% identical.
- This allows detection of more distantly related sequences, as it downplays the role of the more related sequences in the block when building the matrix.

BLOSUM62 MATRIX

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
```

So which observed matrix to use???

- For global alignments use PAM matrices.
  - Lower PAM matrices tend to find short alignments of highly similar regions.
  - Higher PAM matrices will find weaker, longer alignments.
- For local alignments use BLOSUM matrices.
  - BLOSUM matrices with HIGH number, are better for similar sequences.
  - BLOSUM matrices with LOW number, are better for distant sequences.

Tips...

- When doing global alignment (and database scanning) of related (similar) sequences use PAM200 or PAM250.
- If you don't know what to expect (e.g. for database scanning) use PAM120.
- For local database scanning (e.g blast), or for ungapped, local alignments, use BLOSUM62 (recommended for proteins).

Tips….

- In all cases it is recommended to use more than one matrix for any database scanning.