Sequence Similarity Search: Statistics

by Ahmet Sacan

Similarity

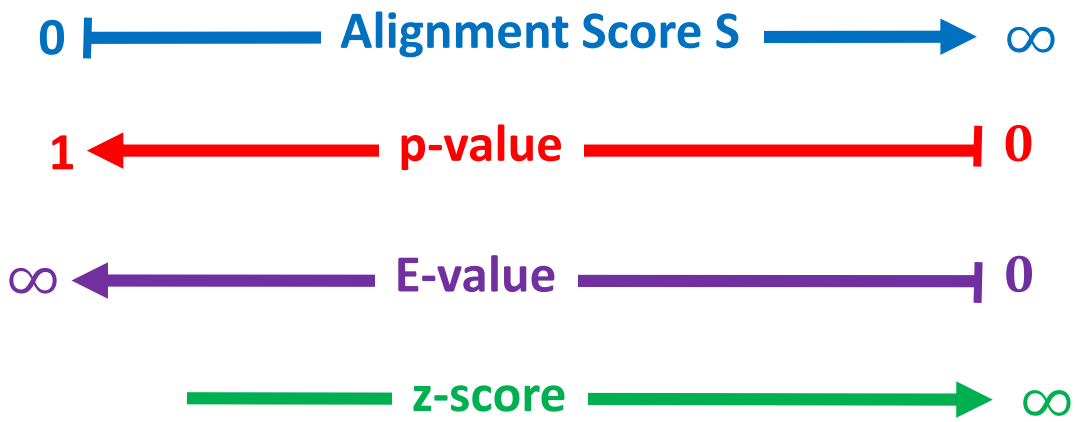
- Measures of similarity
  - Percent identity
  - Alignment Score
- Possible causes for similarity:
  - Common ancestry
  - By chance
- How similar do the sequences need to be to infer homology?

Statistical significance

- Need a model for random similarity
- Homology/True-similarity will be assessed by how different an alignment score is from random similarity

Statistical measures

- Let's consider a homology search result, having an alignment score of S with the query.
- p-value: probability that at least one sequence will produce the same or better score by chance
- E-value: expected number of sequences that will produce same or better score by chance
- z-score: number of standard deviations from (above) the mean of the score distribution



Expected value

- $E[X] = \sum x_i p_i$
- What is the expected value of a die roll?

Statistical significance of alignments

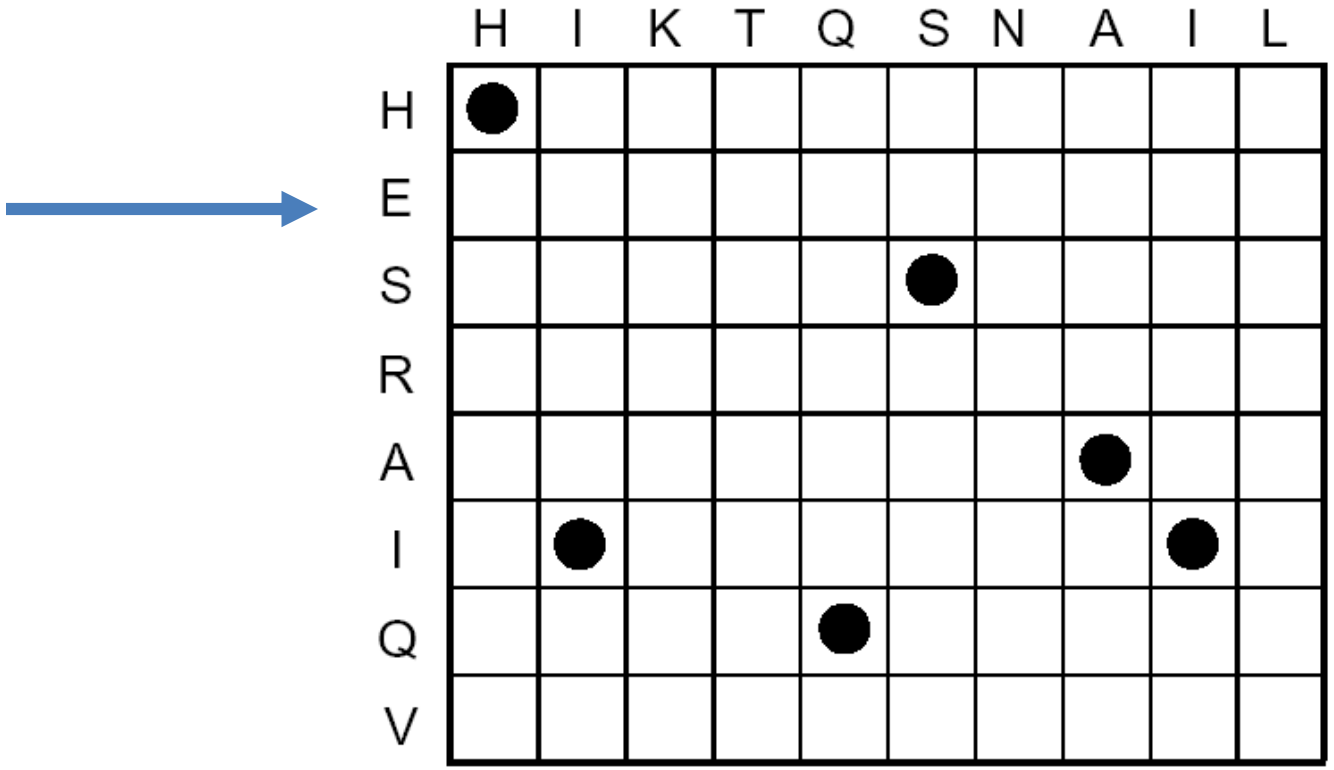
- Match-run
  - Erdos-Renyi
- Local alignments without gaps
  - Karlin-Altschul

- Local alignments with gaps
- Global alignments

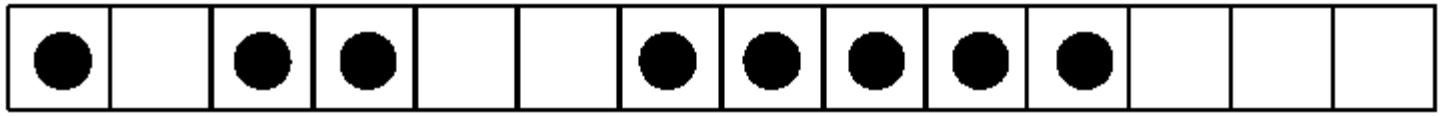
Match-run model: consecutive identical matches

HESRAIQV

HIKTQSNAIL



Analysis of coin tosses

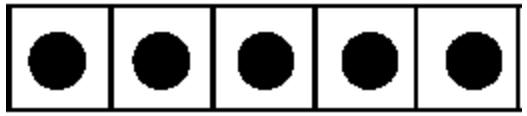


- Let black dots indicate heads.
- Let p be the probability of a head

$$p = \frac{1}{2} = 0.5$$

Probability of L=5 heads in a row

If you toss a coin 5 times, what is the probability of 5 heads in a row?



Expected number of L=5 heads

- What is the expected number of times 5 consecutive heads occur in 14 coin tosses?

Analysis of coin tosses

- Probability of 5 heads in a row:

$$p^5 = \frac{1}{2^5} = 0.031$$

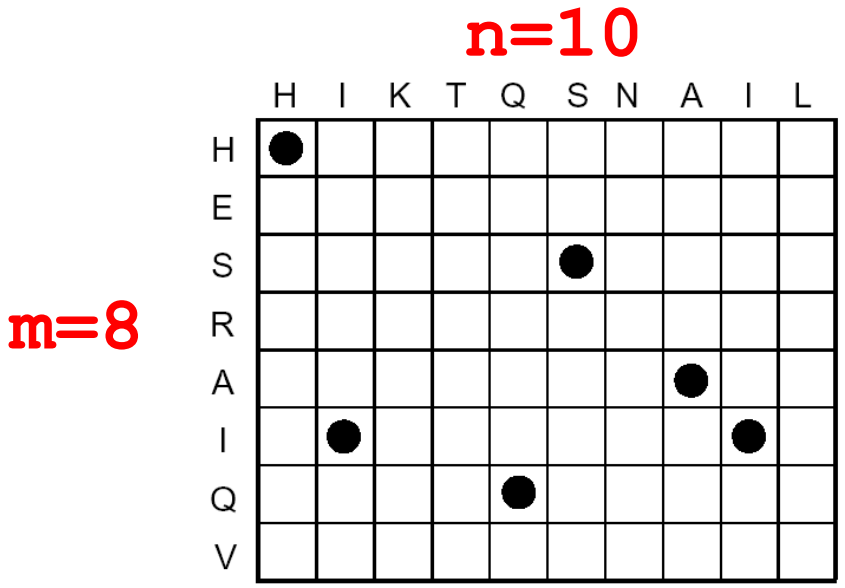
- Expected number of times 5 heads occur in 14 tosses:

$$10 * p^5 = 10 * \frac{1}{2^5} = 0.31$$

Analysis of coin tosses

- Expected number of times a length L run of heads in n tosses
  - $E(L) = (n - L + 1)p^L$
  - If $\mathrm{n} \gg \mathrm{L}$, $E(L) \cong np^L$
- What is the expected length R of the longest run of heads in n tosses?
  - $1 \cong np^R$
  - $\mathrm{R} \cong \log_{1/p}(n)$
- Expected length of longest run of heads for 14 tosses:
  - $\mathrm{R} \cong \log_{1/.5}(14) = 3.8$

Analysis of alignment length



- Probability of an individual match (assume proteins)
  - p = 1/20 = 0.05
- Expected number of matches:
  - E(L=1) = mnp = 8 x 10 x 0.05 = 4
- Expected number of two successive matches
  - $E(L) \cong mnp^L$ = **10** x **8** x 0.05^2 = 0.2

Analysis of alignment length

- Successive matches correspond to a run of matches along the diagonal.
- There are (m-L+1)x(n-L+1) places for a match

$$E(L) = (m - L + 1)(n - L + 1)p^L$$

- assume m>>L, n>>L:

$$E(L) \cong mnp^L$$

- Expected length of longest match:

$$1 \cong mnp^R$$

$$\mathrm{R} \cong \log_{1/p}(mn)$$

Example: Significance of an Alignment

- We perform similarity search of a DNA of length m=1,000 against a genome of length n=1,000,000.
- We find a local stretch of identical alignment with L=10.

$$p = \frac{1}{4}$$

$$E(L) \cong mnp^L = 1000 * 1000000 * \frac{1}{4^{10}} = 953.7$$

Example: Significance of an Alignment

- We perform similarity search of a DNA of length m=1,000 against a genome of length n=1,000,000.
- We find a local stretch of identical alignment with **L=25**.

$$p = \frac{1}{4}$$

$$E(L) \cong mnp^L = 1000 * 1000000 * \frac{1}{4^{25}} = 8.8E^{-7}$$

E-Value with a substitution matrix

- The expected number of alignments with score x or higher is given by:

$$E(S \geq x) = Kmnp^{-\lambda x}$$

- $K < 1$ is a proportionality constant that corrects for the fact that there are not really mn independent places that could have produced $S \geq x$
- λ is related to the substitution matrix, and accounts for the fact that the matrix contains a scaled/transformed version of co-occurrence probabilities.

λ scaling factor for substitution matrix

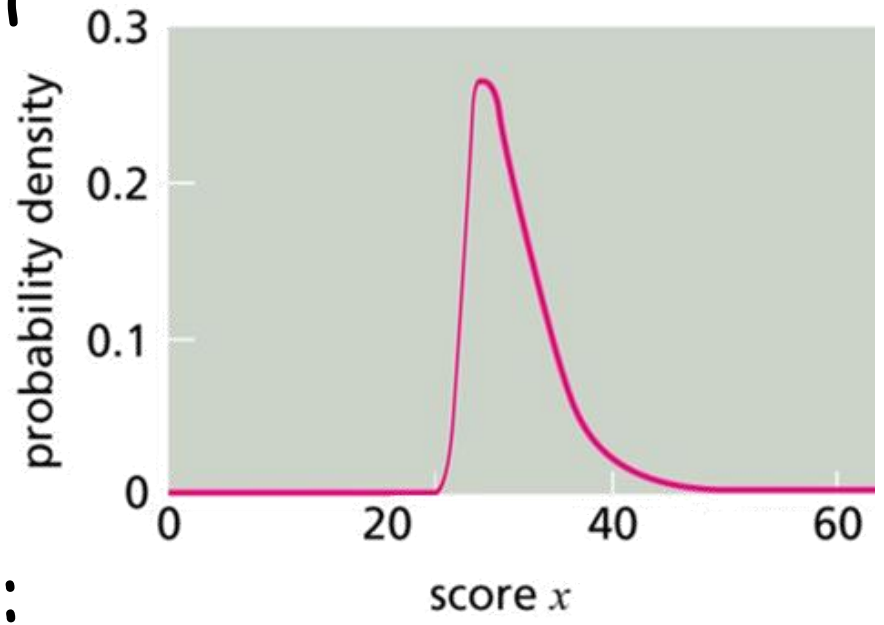
- The substitution matrix (e.g. BLOSUM62) is derived using:
  - $\lambda\, s_{ij} = \log(\frac{q_{ij}}{p_i p_j})$
  - λ is a constant factor to get an integer matrix.
  - $q_{ij} = p_i p_j e^{\lambda s_{ij}}$
- Sum of observed co-occurrence frequencies is 1.
  - $\sum_i \sum_j^i q_{ij} = 1$
- λ can be estimated from the substitution matrix:
  - $\sum_i \sum_j^i p_i p_j e^{\lambda s_{ij}} = 1$

Constraint on substitution matrix

- Substitution matrix has to satisfy the following constraint:
  - $E(s_{i,j}) = \sum_{i,j} p_i p_j s_{i,j} \leq 0$
  - What happens otherwise?

Probability distribution of alignment scores

- Scores of local ungapped alignments follow an Extreme Value (Gumbel) distribution
  - Location (of peak probability):
$$U = \ln(Kmn) / \lambda$$
  - Scale: $-1/\lambda$
- CDF:
$$P(S < x) = e^{-e^{-\lambda(x-U)}}$$
- Probability of score being at least x:
$$P(S \geq x) = 1 - e^{-e^{-\lambda(x-U)}}$$
$$= 1 - e^{-Kmne^{-\lambda x}}$$



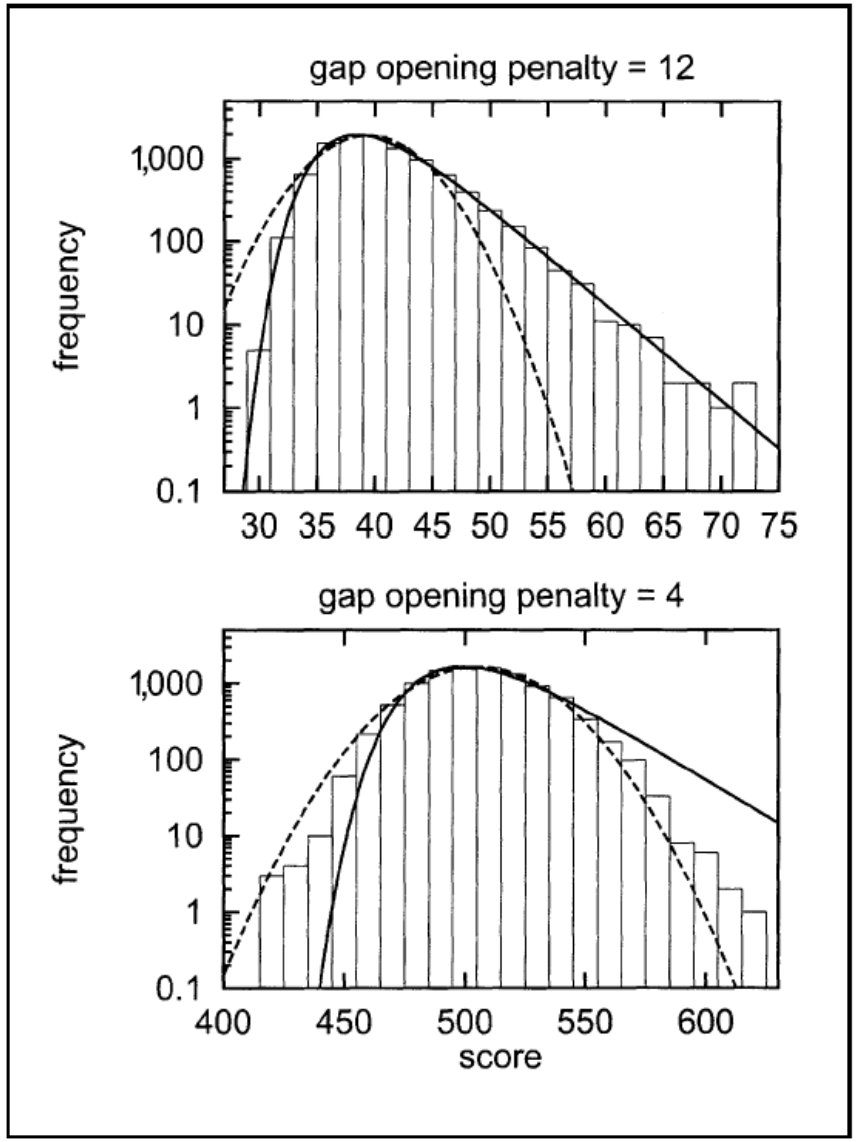
Statistics of alignment score in practice

- BLAST pre-calculates parameters for scoring matrices, gap penalties, and database size.

- FASTA estimates the probability distribution of alignments during querying.

Alignment statistics can be derived from random/ized databases

- Random databases
  - Take amino acid distributions into account

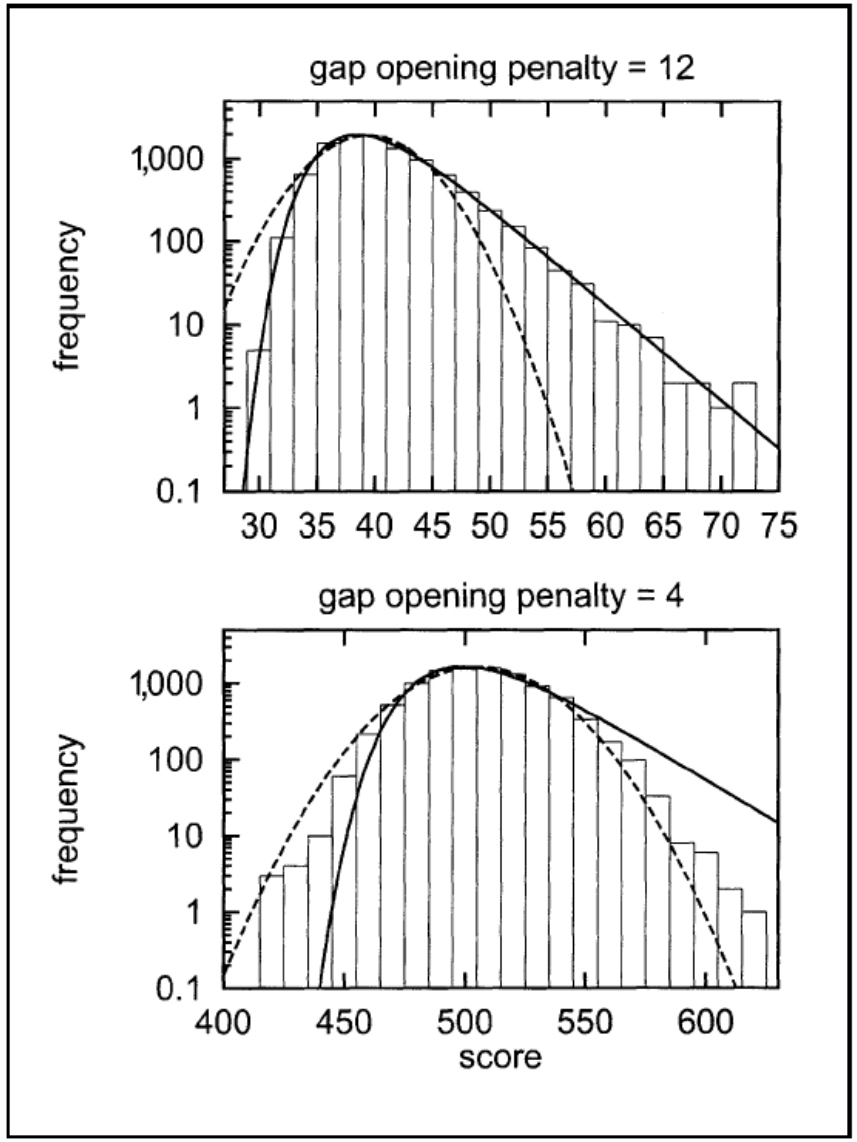


Single-residue statistics

- Leucine is most abundant amino acid (9.3%), followed by Serine (7.2%)
- The rarest amino acid is Tryptophan (1.3%), followed by Cysteine (1.7%)
- Residue compositions can be informative
  - Metallothioneins have ~30% Cysteine, and form metal cages using -SH groups.
  - Some antifreeze proteins have ~50% Alanine, causing hydrophobic interactions with water.

Alignment statistics can be derived from random/ized databases

- Random databases
  - Take amino acid distributions into account
- Randomized databases
  - Scramble
  - Permute
  - Window-permute
  - reverse



Summary

- Significance of a local alignment search result:
  - compare with a random distribution of alignment scores
  - E-value <= E-5
  - p-value <= E-2

- Obtaining the random distribution:
  - Build a statistical/parametric model
    - Match-run
    - Karlin-Altschul: Incorporate substitution matrix and residue frequencies
  - Generate scores from random/-ized databases