* 3 purposes of a score
  - compare different solutions of same input.
  - compare different inputs.
  - significance

* likelihood ratio.

$$\log \frac{Pr(\text{alignment} \mid \text{homology})}{Pr(\text{alignment} \mid \text{random})}$$

* BLOSUM.

$$2 \cdot \log \frac{P(a,b)}{2P(a) \cdot P(b)} \qquad \log \frac{P(a,a)}{P(a) \cdot P(a)}$$

$d(s,t)$

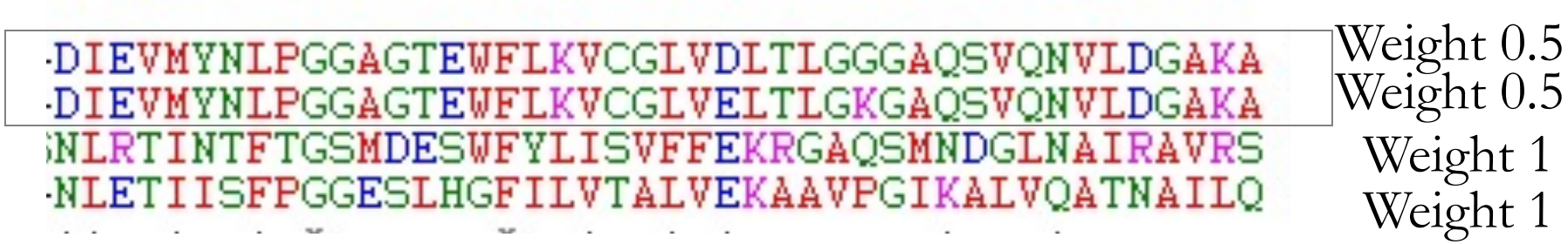
a	b
b	a

$d(ss, tt)$

Deal with Sample Bias

- Some protein families are more well studied so they are over-represented in the database.
- Such bias is caused by the studies, not reflecting what's going on during evolution.
- To remove this bias in statistics, those "redundant" proteins are classified together before BLOSUM calculation.

BLOSUM 62



- The sequences that are 62% or above similarity are grouped together and given total weight 1.
- This way, the AA pairs are counted between groups that are 62% similar or below.
- The lower this number is, the better is the matrix suitable to distant homology search.
- The original BLOSUM paper found out 62 is best at the time the paper was prepared.

A quick review

- Where does a score come from?
- Homology model: Two sequences are homologous using the alignment/evolutionary history.
- Random model: Two sequences are irrelevant.
- Which model better explains the alignment?
- Log likelihood ratio
- If greater than zero, supports Homolog, else, supports Random.

- Statistics should be done carefully to avoid oversampling.

Does the alignment indicate a homology?

PURPOSE 3

Significance

- Consider that BLAST matches a query sequence with all database sequences, and return the highest scoring local alignments.
- The best local alignment score is 100. Does it mean good or bad?
- The answer depends on the score scheme you use so we need to standardize it for effective communication.
- Intuitively, we would ask "can this happen randomly"?
- This is formalized as the p-value.