Sequence alignment

Substitution matrices

Outline

- How do we determine the substitution matrices for sequence alignment?
- Probabilistic models of related/unrelated sequences
- Computing substitution matrices from data

Probabilistic Model of Alignments

- We'll focus on protein alignments without gaps
- given an alignment, we can consider two possibilities

  R: the sequences are related by evolution

  U: the sequences are unrelated

- How can we distinguish these possibilities?
- How is this view related to amino-acid substitution matrices?

Model for *Unrelated* Sequences

- We'll assume that each position in the alignment is sampled randomly from some distribution of amino acids
- We'll assume that amino acids at each position are **independent** of each other

- let q_a be the probability of amino acid a

- the probability of an n-character alignment of x and y is given by

$$\Pr(x, y \mid U) = \prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{n} q_{y_i}$$

"likelihood"

Model for *Related* Sequences

- We'll assume that each pair of aligned amino acids evolved from a common ancestor
- We'll assume each pair is **independent** of the other pairs

- let p_{ab} be the probability that evolution gave rise to amino acid *a* in one sequence and *b* in another sequence

- the probability of an alignment of *x* and *y* is given by

$$\Pr(x, y \mid R) = \prod_{i=1}^{n} p_{x_i y_i}$$

Probabilistic Model of Alignments

- How can we decide which possibility (U or R) is more likely?
- one principled way is to consider the relative likelihood of the two possibilities

$$\frac{\Pr(x, y \mid R)}{\Pr(x, y \mid U)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}}$$

- taking the log, we get

$$\log \frac{\Pr(x, y \mid R)}{\Pr(x, y \mid U)} = \sum_i \log \left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right)$$

- This is the *log-odds ratio* (or *log likelihood ratio)*

Probabilistic Model of Alignments

- If we let the substitution matrix score for the pair *a, b* be:

$$s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

- Then the score of an ungapped alignment is the log likelihood ratio:

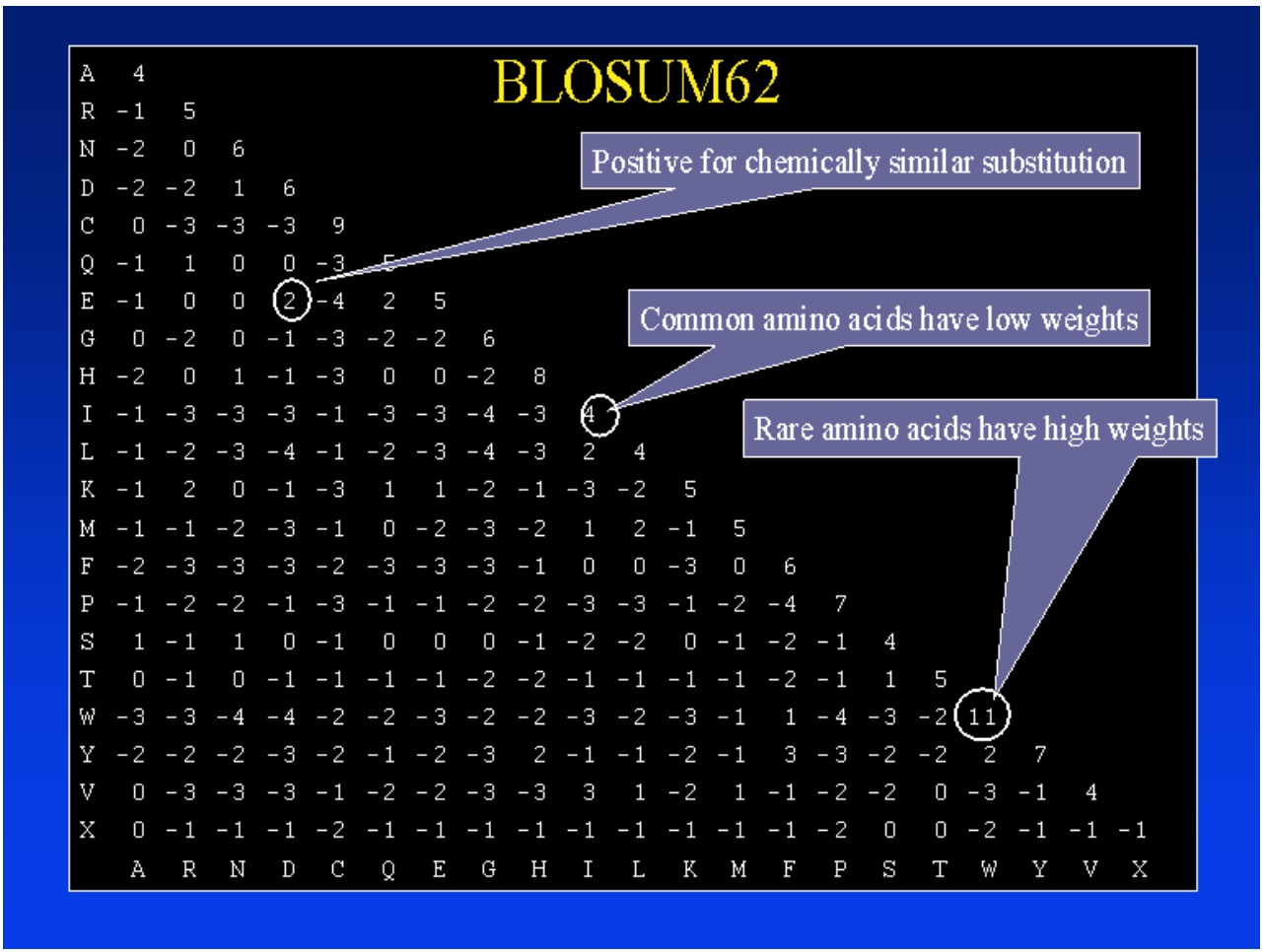
$$S = \sum_i s(x_i, y_i) = \log \frac{\Pr(x, y \mid R)}{\Pr(x, y \mid U)}$$

Substitution Matrices

- two popular sets of matrices for protein sequences
  - PAM matrices [Dayhoff *et al.,* 1978]
  - BLOSUM matrices [Henikoff & Henikoff, 1992]

- both try to capture the the relative substitutability of amino acid pairs in the context of evolution

Blosum 62 Matrix



Substitution Matrices

- the substitution matrix score for the pair *a*, *b* is given by:

$$s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

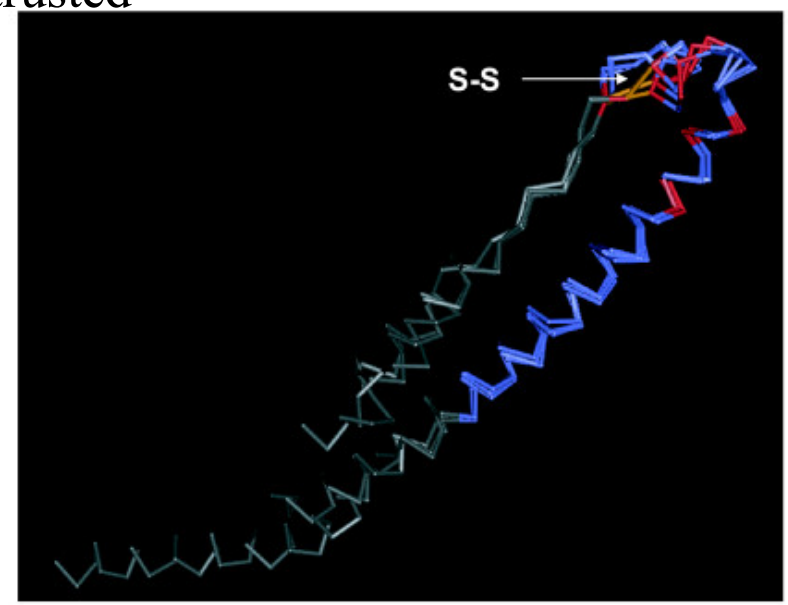
- but how do we get values for p_{ab} (probability of *a* and *b* given that they are derived from a common ancestor)?
- it depends on how long ago sequences diverged

  diverged recently: $p_{ab} \approx 0$ for $a \neq b$

  diverged long ago: $p_{ab} \approx q_a q_b$

Substitution Matrices

- key idea: trusted alignments of related sequences provide information about biologically permissible mutations
- protein structure similarity provides the gold standard for which alignments are trusted




```
D2/CEL/OR  sdvqAISSTIQDLQDQVDSLAEVVLQNRRGLDLLTAEQGGICLALQEKCcfyank
MMLV       ddlrEVEKSISNLEKSLTSLSEVVLQNRRGLDLLFLKEGGLCAALKEECafyad~
HTLV-1     kdisQLTQAIVKNHKNLLKIAQYAAQNRRGLDLLFWEQGGLCKALQEQCcflnit
Ebola      qlanETTQALQLFLRATTELRTFSILNRKAIDFLLQRWGGTCHILGPDCriephd
```

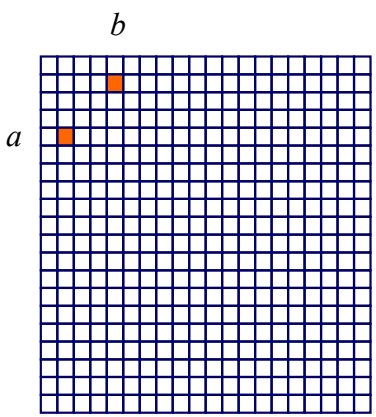
BLOSUM Matrices

- [Henikoff & Henikoff, *PNAS* 1992]
- probabilities estimated from "blocks" of sequence fragments that represent *structurally* conserved regions in proteins
- transition frequencies observed directly by counting pairs of characters between clusters in the blocks. Sequences within blocks are clustered at various levels:
  - 45% identical (BLOSUM-45) distantly related
  - 50% identical (BLOSUM-50)
  - 62% identical (BLOSUM-62) more recently
  - etc.

BLOSUM Matrices

- given: a set of sequences in a block clustered at X% identity
- fill in matrix *A* with number of observed substitutions between pairs of sequences in different clusters of the block
- (we won't worry about details of some normalization that happens here)

a paired with *b* →



$$p_{ab} = \frac{A_{ab}}{\sum_{c,d} A_{cd}} \qquad q_a = \frac{\sum_b A_{ab}}{\sum_{c,d} A_{cd}}$$

PAM matrices

- Use amino acid pair counts from closely related sequences only
  - Initially gives a substitution matrix for closely related sequences
- Matrices for more distantly related sequences are derived from the initial matrix via extrapolation (essentially matrix multiplication)

Summary

- Derivation of substitution scores as log ratios of related vs. unrelated model likelihoods
- Estimation of parameters for substitution scores via trusted alignments