The statistics of pairwise alignment

BMI/CS 576
www.biostat.wisc.edu/bmi576/
Colin Dewey
colin.dewey@wisc.edu
Fall 2016
Issues in scoring pairwise alignments

• How do we determine the substitution and gap scores for alignment?
• How do we determine whether the score of the best alignment is indicative of truly related sequences?
• These issues are related and addressed via statistical models
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

\[
\Pr(x, y | U) = \prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{n} q_{y_i}
\]
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 relative substitutability of amino acid pairs in the context of evolution
Blosum 62 Matrix

Positive for chemically similar substitution
Common amino acids have low weights
Rare amino acids have high weights
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
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)
  – 50% identical (BLOSUM-50)
  – 62% identical (BLOSUM-62)
  – etc.
BLOSUM Matrices

- given: a set of sequences in a block
- fill in matrix $A$ with number of observed substitutions (we won’t worry about details of some normalization that happens here)

\[
p_{ab} = \frac{A_{ab}}{\sum_{c,d} A_{cd}} \quad q_a = \frac{\sum_b A_{ab}}{\sum_{c,d} A_{cd}}
\]
Assessing significance of the alignment score

• There are two ways to do this
  – Bayesian approach
  – Classical approach
Bayesian approach

- Compute probability of Related model using Bayes rule
- Requires prior probability of R and U

\[
\Pr(R \mid x, y) = \frac{\Pr(x, y \mid R) \Pr(R)}{P(x, y)} = \frac{\Pr(x, y \mid R) \Pr(R)}{\Pr(x, y \mid R) \Pr(R) + \Pr(x, y \mid U) \Pr(U)} = \frac{\Pr(x, y \mid R) \Pr(R) / \Pr(x, y \mid U) \Pr(U)}{\Pr(x, y \mid R) \Pr(R) / \Pr(x, y \mid U) \Pr(U) + 1}
\]
Classical approach

Determine how likely it is that such an alignment score would result from chance.

3 ways to calculate chance; look at alignment scores for
- real but non-homologous sequences
- real sequences shuffled to preserve compositional properties
- sequences generated randomly based upon a DNA/protein sequence model
Scores from Random Alignments

• suppose we assume
  • sequence lengths $m$ and $n$
  • a particular substitution matrix and amino-acid frequencies

• and we consider generating random sequences of lengths $m$ and $n$ and finding the best alignment of these sequences

• this will give us a distribution over alignment scores for random pairs of sequences
Statistics of Alignment Scores: The Extreme Value Distribution

- in particular, we get an *extreme value distribution*
Distribution of Scores

- the expected number of alignments, $E$, with score at least $S$ is given by:

$$E(S) = K m n e^{-\lambda S}$$

- $S$ is a given score threshold
- $m$ and $n$ are the lengths of the sequences under consideration
- $K$ and $\lambda$ are constants that can be calculated from
  - the substitution matrix
  - the frequencies of the individual amino acids
Statistics of Alignment Scores

- to generalize this to searching a database, have \( n \) represent the summed length of the sequences in the DB (adjusting for edge effects)
- the NCBI BLAST server does just this
- theory for *gapped* alignments not as well developed
- computational experiments suggest this analysis holds for gapped alignments (but \( K \) and \( \lambda \) must be estimated from data)