Substitution Matrices

David Bernick
Motivation

• we are interested in establishing the likelihood on the hypothesis that two observations in each of two molecules has descended from a common ancestral character.

• we can not, typically, observe the ancestral character

• we instead assume that the characters provide the same “purpose” and ask how similar these two characters are to each other
Mutational Model

• Mutations that produce structural change should be less probable than those that do not
• Mutable amino acids: Asn, Asp, Glu, Ser
• Least mutable: Cys, Trp
Mutational Model - 2

- Model should capture frequency of $AA_i \rightarrow AA_j$, $P(Aa_{ij})$
- Model should consider mutational time – $t$
  – Molecular clock
- Model should be a comparison of 2 models
  – $M$ substitutions in evolutionarily related sequences
  – $R$ “alignments” made by chance
    – Given by the frequency of occurrence in all proteins
- Should model be reversible?
  – $P(Aa_{ij} \mid M,t) = (PAA_{ji} \mid M,t)$
Scoring Models

• Our scores will compare 2 models:
  – Probability of 2 AA deriving from a common ancestor
    • $M = P(\text{related} \mid \text{AA}_i, \text{AA}_j)$
  – Probability of 2 AA being aligned by chance
    • $R = P(\text{random} \mid \text{AA}_i, \text{AA}_j)$
  – We make a log-odds score
    • $s(i,j) = \log_2(M/R)$
  – We can then scale and round the scores by some convenient factor (to make nice integers)
|   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
| 5 | -2|-1|-2|-1|-1|0|-2|-1|-2|-1|-1|-3|-1|1|0|-3|-2|0|-2|-1|-1|-5 |   |
| -2| 7 |-1|-2|-4|1|0|-3|0|-4|-3|3|-2|-3|-3|-1|-1|-3|-1|-3|1|0|-1|-5 |   |
| -1| -1| 7 |2|-2|0|0|0|1|-3|-4|0|-2|-4|-2|1|0|-4|-2|-3|4|0|-1|-5 |   |
| -2| -2| 2 |8|-4|0|2|-1|-1|-4|-4|-1|-4|-5|-1|0|-1|-5|-3|4|5|1|-1|-5 |   |
| -1| -4|-2|-4|13|-3|-3|-3|-3|-2|-2|-2|-2|-4|-1|-1|-5|-3|-1|-3|-3|-2|-5 |   |
| -1| 1 |0 |0|-3|7|2|-2|1|-3|-2|2|0|-4|-1|0|-1|-1|-1|-3|0|4|-1|-5 |   |
| -1| 0 |0 |2|-3|2|6|-3|0|-4|-3|1|-2|-3|-1|-1|-3|-2|-3|1|5|1|-5 |   |
| 0 |-3|0|-1|-3|-2|-3|8|-2|-4|-4|-2|-3|-4|-2|0|-2|-3|-3|-4|-1|-2|-5 |   |
| -2| 0 |1 |-1|-3|1|0|-2|10|-4|-3|0|-1|-1|-2|-1|-2|-3|2|-4|0|0|-1|-5 |   |
| -1| -4|-3|-4|-2|-3|-4|-4|-4|5|2|-3|2|0|-3|-3|-1|-3|-1|4|-4|-3|-1|-5 |   |
| -2| -3|-4|-4|-2|-2|-3|-4|-3|2|5|-3|3|1|-4|-3|-1|-2|-1|1|-4|-3|-1|-5 |   |
| -1| 3 |0|1|-3|2|1|2|0|-3|-3|6|-2|-4|-1|0|-1|-3|-2|-3|0|1|-1|-5 |   |
| -1| -2|-2|-4|-2|0|-2|-3|-1|2|3|-2|7|0|-3|-2|-1|1|0|1|-3|-1|-1|-5 |   |
| -3| -3|-4|-5|-2|-4|-3|-4|-1|0|1|-4|0|8|-4|-3|-2|1|4|-1|-4|-4|-2|-5 |   |
| -1| -3|-2|-1|-4|-1|-1|-2|-2|-3|-4|-1|-3|-4|10|-1|-1|-4|-3|-3|-2|-1|-2|-5 |   |
| 1 | 1 |1 |0|-1|0|1|0|-1|-3|-3|0|-2|-3|-1|5|2|-4|-2|-2|0|0|-1|-5 |   |
| 0 | 1 |0|0|-1|-1|-1|-2|-2|-1|-1|-1|-2|-1|2|5|3|-2|0|0|-1|0|-5 |   |
| -3| -3|-4|-5|-5|-5|-1|-3|-3|-3|-3|-2|-3|-1|1|-4|-4|-3|15|2|-3|-5|-2|-3|-5 |   |
| -2| -1|-2|-3|-3|-1|-2|-3|2|-1|-1|-2|0|4|-3|-2|-2|2|8|-1|-3|-2|-1|-5 |   |
| 0 | 3 |-3|-4|-1|-3|-3|-4|-4|4|1|-3|1|-1|-3|-2|0|-3|-1|5|-4|-3|-1|-5 |   |
| -2| -1|4|5|-3|0|1|-1|0|-4|-4|0|-3|-4|-2|0|0|-5|-3|-4|5|2|-1|-5 |   |
| -1| 0 |0|1|-3|4|5|-2|0|-3|-3|1|-1|-4|-1|0|-1|-2|-2|-3|2|5|-1|-5 |   |
| -1| -1|-1|-1|-2|-1|-1|-2|-1|-1|-1|-2|-2|-1|0|-3|-1|-1|-1|-1|-1|-1|-5 |   |
| -5| -5|5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|-5|5|1  |   |
PAM matrices

• Point Accepted Mutations - Margaret Dayhoff

• Construct an $N \times N$ matrix where each column represents the observed transition frequency of substitution for a given AA to each of the others.

• She observed 1572 exchanges in 71 closely related protein families.

• All columns in the Mutation matrix sum to 1.

• She defined 1 PAM unit to yield 1 observed mutation / 100 AA.

• PAM is then a unit of time
PAM Matrices - 2

• Multiple mutations can occur at a single location.

• If sequences are closely related this is less probable
  – Negligible effect at 1 PAM
  – Substantial effect at 250 PAMs
PAM Matrices - 3

• Calculate $P(B|A) = \frac{c(AB)}{c(A^*)}$
  – Only mutations are counted ($A \neq B$)
  – When counting mutations
    • $c(AB)=c(BA)$ (we count both)
  – Scale all $P(B|A)$ such that:
    \[
    \sum_a \sum_b P(A)P(B)P(B|A) = 0.01
    \]
    
    \[\text{where: } P(A) = \frac{\sum c(A^*)}{N}\]
  
  • Using scaling factor: $\sigma$
  – Rescale all off diagonals $P(B|A)$ using $\sigma$
  – Set $P(A|A) = 1-\sigma$
  – This is the $S(1)$ mutation matrix (not sub. Matrix)
PAM Matrices – 4

• We can now compute PAM Matrices for longer PAM times, by:

\[ S(2) = S(1)^2 \]
\[ S(n) = S(1)^n \]

• Entries in \( S(t) \) are then converted to scores by:

\[ s(A, B | t) = \log_2 \frac{P(B | A, t)}{P(B)} \]
PAM Matrices – 5

• and finally..
  – any score matrix can be rescaled with a constant
• PAM250 is scaled to “third bits”
PAM Matrices - 6

• PAM Matrices are compelling, but...
  – $S(1)$ models short timeframe substitutions
    • Dominated by single nucleotide changes
  – $S(250)$ does not capture more remote timeframe substitutions
BLOSUM Matrices

• Henikoff and Henikoff
• Select aligned blocks from Protein families
  – BLOCKS are aligned, ungapped regions or related proteins
• Cluster all sequences that are at or above some threshold in %identity (L)
• Compute Aligned Pair counts between clusters, weighting each count by cluster size
  – $1/n_1 n_2$
  – this removes bias from oversampled sequences
• We now normalize over each column to produce $P(AA_{ij} | AA_i, M)$.
  – transition probabilities from some AA to all possibilities

• Compare this model again R (random model) as a likelihood ratio

• Convert to log space

• Scale
Example:
   cluster sequences above L% ID into their own cluster
   count pairwise alignments(assume each in sep. cluster)
   AA (3)
   AD (3)
   AG (3)
   DG (1)
Compute conditionals:
   P(A|A)=3/9
   P(D|A)=3/9
   P(G|A)=3/9

... 
Note that Matrix is symmetric
Compute s(A,D) = log2(P(D|A)/P(D)) = log2( (3/9) / (4/10) )= -.26
   s(A,A) = log2(P(A|A)/P(A)) = log2( (3/9) / (9/10)) = -1.4