

BIOL 495S/ CS 490B/ MATH 490B/ STAT 490B

Spring 2002, Feb. 4 and 6 lectures

Reference:

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, 1998 “Biological sequence analysis: Probabilistic models of proteins and nucleic acids”, Section 2.2, 2.8 and 3.2.
The textbook “A primer of genome science”, p 95-97.

Defining substitution matrices for sequence alignment

We need score terms for each aligned residue pair. A biologist with a good intuition for proteins could invent a set of 210 scoring terms for all possible pairs of amino acids, but it is extremely useful to have a guiding theory for what the scores mean. We will derive substitution scores from probabilistic model.

Let us establish some notation. We will be considering a pair of sequences, x and y , of lengths n and m , respectively. Let x_i be the i th symbol in x and y_j be the j th symbol of y . These symbols will come from some alphabet \mathcal{A} ; in the case of DNA this will be the four bases $\{A, G, C, T\}$, and in the case of proteins the twenty amino acids. We denote symbols from this alphabet by lower-case letters like a, b . Let us first consider ungapped global pairwise alignments, that is, two completely aligned equal-length sequences.

The unrelated or random model R gives the probability

$$P(x, y | R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

with the assumption that letter a occurs independently with frequency q_a . In the alternative match model, aligned pairs of residues occur with a joint probability p_{ab} . This value p_{ab} can be thought of as the probability that the residues a and b have been derived from a common ancestor. The probability for the whole alignment is

$$P(x, y | M) = \prod_i p_{x_i y_i}.$$

The ratio of these two likelihoods is known as the odds ratio:

$$\frac{P(x, y | M)}{P(x, y | R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}.$$

We take the logarithm of this ratio (log-odds ratio) to obtain scores of aligning the pair of sequences

$$S = \sum_i s(x_i, y_i),$$

where

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

is the log likelihood ratio of the residue pair (a, b) occurring as an aligned pair, as opposed to an unaligned pair.

Arranging $s(a,b)$ scores in a matrix, we get the substitution matrix, for instance, the BLOSUM50 matrix. The score matrix is obtained from probabilities. Then it comes the issue of how to estimate the probabilities. A simple and obvious approach would be to count the frequencies of aligned residue pairs in confirmed alignments, and to set the probabilities p_{ab}, q_a to the normalized frequencies. The set of BLOSUM matrices are derived by this approach. In detail, they were derived from a set of aligned, ungapped regions from protein families called the BLOCKS database. The sequences from each block were clustered, putting two sequences into the same cluster whenever their percentage of identical residues exceeded some level $L\%$. Then we calculate the frequencies A_{ab} of observing residue a in one cluster aligned against residue b in another cluster, correcting for the size of the clusters by weighting each occurrence by $1/(n_1 n_2)$, where n_1 and n_2 are the respective cluster sizes.

From A_{ab} , the probabilities are estimated by

$$q_a = \sum_b A_{ab} / \sum_{cd} A_{cd}$$

i.e. the fraction of pairings that include an a , and

$$p_{ab} = A_{ab} / \sum_{cd} A_{cd}$$

i.e. the fraction of pairings between a and b out of all observed pairings. Then the score matrix entries were derived using

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

The resulting log-odds score matrices were scaled and rounded to the nearest integer value. For $L = 62$ and $L = 50$ we get BLOSUM62 and BLOSUM50 substitution matrices respectively. BLOSUM62 is standard for ungapped matching, and BLOSUM50 for alignment with gaps. Note that lower L values correspond to more distantly related sequences, and are applicable for less similarity searches.

Hidden Markov models

In the example of CpG island, we want to build a single model that incorporates both Markov chains, model “+” and model “-”. A motivation of this is to answer a question like: How do we find the CpG island in a long unannotated sequence? We want to have both the Markov chains present in the same model, with a small probability of switching from one chain to the other at each transition point. We have to introduce two states corresponding to each nucleotide symbol. We now have A_+, C_+, G_+, T_+ which emit A, C, G, T respectively in CpG island regions, and A_-, C_-, G_-, T_- correspondingly in non-island regions. The transition probabilities in this model are set so that within each group they are close to the transition probabilities of the original component model, but there is also a small but finite chance of switching into the other component. Overall there is more chance of switching from “+” to “-” than vice versa, so if left to run free, the model will spend more of its time in the “-” non-island states than in the island states.

The relabelling of the states is the critical step. The essential difference between a Markov chain and a hidden Markov model is that for a hidden Markov model there is not a one-to-one correspondence between the states and the symbols. It is no longer possible

to tell what state the model was in when x_i was generated just by looking at x_i . In our example there is no way to tell by looking at a single symbol C in the isolation whether it was emitted by state C_+ or state C_- .

Let us formalize the notation for hidden Markov models. We have to distinguish the sequence of states from the sequence of symbols. Let us call the state sequence the path π . The path itself follows a simple Markov chain. The i th state in the path is called π_i . The chain is characterized by parameters

$$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k).$$

Another set of parameters is the emission probabilities $e_k(b)$ for emitting symbol b from state k .

$$e_k(b) = P(x_i = b \mid \pi_i = k)$$

For the CpG island model the emission probabilities are all 0 or 1. The reason for the name emission probabilities is that hidden Markov models generate or emit sequences. The states are hidden though. A sequence can be generated from an HMM as follows: First a state π_1 is chosen according to the starting probabilities, denoted by a_{0i} . In that state an observation x_1 is emitted according to the distribution $e_{\pi_1}(\cdot)$ for that state. Then a new state π_2 is chosen according to the transition probabilities a_{π_i} and so forth. This way a sequence of random, artificial observations are generated. Therefore, we will sometimes say that $P(x)$ is the probability that x was generated by the model.

It is now easy to write down the joint probability of an observed sequence x and a state sequence π :

$$P(x, \pi) = \prod_{i=1}^L a_{\pi_{i-1}\pi_i} e_{\pi_i}(x_i), \quad (4)$$

where we require $\pi_0 = 0$. For example, the probability of sequence CGCG being emitted by the state sequence (C_+, G_-, C_-, G_+) in the model is

$$a_{0,C_+} \times 1 \times a_{C_+,G_-} \times 1 \times a_{G_-,C_-} \times 1 \times a_{C_-,G_+} \times 1.$$

Equation (4) is the HMM analogue of Equation (3) in Markov chain model (Refer to the lecture notes). However, it is not so useful in practice because in general we do not know the path. In the following sections we will study algorithms to estimate the path and parameters for an HMM.

An example of the HMM

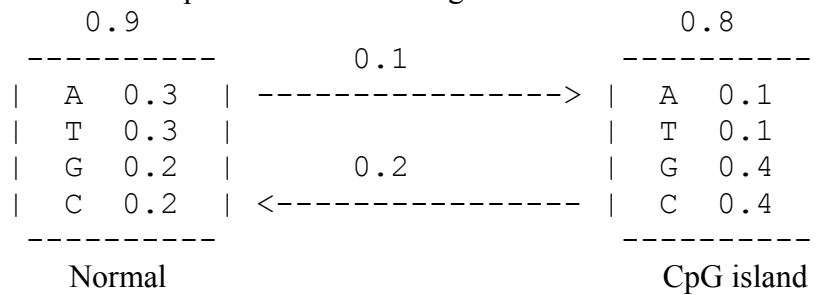
Consider the CpG island example again. We will use another state labeling approach. We now define two hidden states. Every nucleotide belongs to either a normal region (N) or to a CpG island region (R). We use a HMM to describe a long sequence with both normal and CpG island regions. A symbol sequence may look like

TTACTTGACGCCAGAAATCTATATTTGGTAACCCGACGCTAA

With a corresponding state sequence as

NNNNNNNNNNRRRRRNNNNNNNNNNNNNNNNNNRRRRRRRNNNN

The underlined regions are CpG island according to the state sequence. The HMM describing these data is represented as in the figure below.



The states of the HMM are the two categories, N or R. Transition probabilities govern the assignment of states from one position to the next. In the current example, if the present state is N, the following position will be N with probability 0.9, and R with probability 0.1. The four nucleotides in a sequence will appear in each state in accordance to the corresponding emission probabilities.

Consider a simple sequence TGCC. One possible way that this sequence could arise is from the set of hidden states NNNN. Given the hidden states, the probability of generating (emitting) the observed sequence is

$$P(TGCC | NNNN) = 0.3 \times 0.2 \times 0.2 \times 0.2 = 0.0024$$

Whereas, the joint probability of the sequence being emitted by the states NNNN would be

$$P(TGCC, NNNN) = P(TGCC | NNNN)P(NNNN)$$

Assuming the first nucleotide is always in a normal region, then the probability is calculated as

$$\begin{aligned} P(TGCC, NNNN) &= P(TGCC | NNNN)P(NNNN) \\ &= 0.0024 \times 0.9 \times 0.9 \times 0.9 \\ &= 0.00175 \end{aligned}$$

where $P(NNNN)$ is computed by Markov transitions. Summing over all the possible state paths, we obtain the probability of $P(TGCC)$ from the HMM. It will be calculated by

$$\begin{aligned}
P(TGCC) = & P(TGCC | NNNN)P(NNNN) + P(TGCC | NNNR)P(NNNR) \\
& + P(TGCC | NNRN)P(NNRN) + P(TGCC | NRNN)P(NRNN) \\
& + P(TGCC | NNRR)P(NNRR) + P(TGCC | NRNR)P(NRNR) \\
& + P(TGCC | NRRN)P(NRRN) + P(TGCC | NRRR)P(NRRR)
\end{aligned}$$

When we compute the probability of the sequence for all possible paths, we can use the path that contributes to maximum probability as our best estimate of the unknown hidden states. For the sample sequence, one finds that the most probable path is in fact NNNN, which is slightly higher than the path NRRR. If the fifth nucleotide in the series were also a G or C, the path NRRRR would be more likely than NNNNN, providing evidence for the existence of a CpG island.