Weight Matrices

Specific Plan:

Given: N DNA sequences, each of length L.
Method:

1. Form an array of numbers, with four rows labeled `A`, `C`, `G` and `T`, and L columns numbered 1 to L. The value in row r of column i is the number of times that nucleotide r appears as the i^{th} entry in one of the sequences.
2. Divide each of the $4L$ numbers by N. Then the sum of entries in any specified column is 1.0. Let $f_i(r)$ denote the value in row r of column i. For example, $f_3(\mathtt{A})$ intuitively estimates the probability that the 3rd entry of the sequence is `A`.
3. Replace each $f_i(r)$ by $\log_2(f_i(r)/p(r))$, where $p(\mathtt{A}) = p(\mathtt{T}) = 0.29$, and $p(\mathtt{C}) = p(\mathtt{G}) = 0.21$. (But what if $f_i(r) = 0$?)

Given any DNA sequence x of length L, we can use this *weight matrix* to associate a number $w(x)$, where $w(x) > 0$ if and only if x looks more like one of the original N sequences than like a typical mammalian N-mer.

General Plan:

Given: Two training sets of DNA sequences, a “positive set” P and a “negative set” Q.
Goal: Discriminate P and Q. That is, develop a method that can decide whether an arbitrary DNA sequence is more P-like or more Q-like.
Method:

1. Select a variety of probabilistic sequence model that is appropriate for P and one appropriate for Q.
2. Estimate model parameters for the two models. For any DNA sequence x, let $\Phi_P(x)$ and $\Phi_Q(x)$ denote the probability that the models for P and Q (respectively) will generate x, assuming that they generate a sequence of the same length as x.
3. For any x, the number $\log(\Phi_P(x)/\Phi_Q(x))$ exceeds 0 if and only if x is more P-like than Q-like.