4.3 Analysis of multiple biological sequences

Perhaps the most important aspect of the analysis of biological sequences is the search for sequence similarity, as this may give insight into the evolutionary relationships (homologies) between sequences, or into the biological function associated with a particular genomic region or protein sequence. Given the probabilistic framework that we have developed, we can now study sequence similarity by studying aspects of the probability distributions of potentially related sequences. The most straightforward type of analysis is based on **frequency** ("first-order") comparison (of nucleotides/amino acid residues), whereas more sophisticated analyses will concentrate on studying series **dependence** or **association** within the sequences (in a "second-order" comparison).

4.3.1 Marginal frequency analysis

The most rudimentary assessment of sequence similarity that can be carried out is one based on the relative marginal frequencies of the nucleotides or amino acids within each sequence. This can be done within the context of a **Goodness-of-Fit test.** Consider the table of observed frequencies corresponding to two nucleotide sequences

	Nucleotide				
	A	C	G	T	Total
Sequence 1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
Sequence 2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

We wish to test the hypothesis H_0 that the two sequences have the same (marginal) nucleotide probabilities, $p_A = p_1, p_C = p_2, p_G = p_3$ and $p_T = p_4$. To carry out the test, we first compute the estimates of these nucleotide probabilities (using **maximum likelihood estimation**) under the assumption that H_0 **is true**; it turns out that

$$\hat{p}_1 = \frac{n_{.1}}{n} \qquad \hat{p}_2 = \frac{n_{.2}}{n} \qquad \hat{p}_3 = \frac{n_{.3}}{n} \qquad \hat{p}_4 = \frac{n_{.4}}{n}$$

and hence the **expected** or **fitted values** if H_0 **is true** are given by

$$\hat{n}_{ij} = n_{i.}\hat{p}_j = \frac{n_{i.}n_{.j}}{n} \qquad i = 1, 2,\ j = 1, 2, 3, 4$$

There are two test statistics that may be used; the first is the **Chi-squared statistic**

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

and the second is the **Likelihood Ratio** (**LR**) statistic

$$LR = 2\sum_{i=1}^{2}\sum_{j=1}^{4} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}}$$

Both of these statistics have an approximate **Chi-squared distribution** $\chi^2_{(r-1)(c-1)} = \chi^2_3$ distribution, again given that H_0 **is true.** Typically, a significance level of $\alpha = 0.05$ is used for this test, and the critical value in a one-tailed test of the hypothesis is at the 0.95 point of this distribution, that is, at 7.81.

EXAMPLE

Consider the two sequences detailed in the following table:

	Nucleotide				
	A	C	G	T	Total
Sequence 1	273	233	258	236	1000
Sequence 2	281	246	244	229	1000
Total	554	479	502	465	2000

For these sequences, the test statistics are computed to be

$$\chi^2 = 0.964 \qquad LR = 0.964$$

and thus there is **no evidence to reject** H_0 that the nucleotide probabilities are the same for the two sequences.

4.3.2 Probabilistic Assessement of Alignments

In the search for, for example, sequence homology, a key objective is to assess the degree of similarity between two aligned biological sequences. As an illustration, it is important to be able to assess the similarity between two DNA sequences

$$GGAGACTGTAGACAGCTAATGCTATA$$
$$GAACGCCCTAGCCACGAGCCCTTATC$$

that match exactly in positions 1,3,6,9,10,11,13,14,23,24 and 25, allowing for the fact that evolutionary forces (substitution/mutation, deletion and insertion) will disrupt the exact matching of truly homologous sequences. In fact, is important to allow for partially aligned sequences

$$CGGGTA--TCCAA$$
$$CCC-TAGGTCCCA$$

where the symbol $-$ indicates a wild-card position. Alignments are categorized as

- **global -** when entire sequences are aligned
- **local** - when only subsequences are aligned
- **gapped/ungapped** - when the wild-cards $-$ are allowed or otherwise
- **pairwise/multiple** - when two or more than two sequences are processed.

Global Ungapped Pairwise Alignment

For a probabilistic assessment of the degree of global alignment of two sequences, we use the theory of maximal runs in binary sequences developed in section 4.2.1. For the two ungapped sequences, write a 1 ("success") if there is a match at a given position, and a 0 ("failure") otherwise, so that the entire match binary sequence is

$$00100100111011000000001110$$

with runs of successes of lengths 1, 1, 3, 2 and 3, with zero run lengths otherwise. Applying the previous theory, if we consider hypotheses H_0 that the sequences are both randomly generated, the collection of run-length random variables, $X_1, ..., X_n$ derived from the two sequences are independent and identically distributed as *Geometric*(p) random variables, where

$$p_{MATCH} = p_A^2 + p_C^2 + p_G^2 + p_T^2$$

Again, considering the random variable $Y_n = \max\{X_1, ..., X_n\}$ as the test statistic, then under H_0

$$F_{Y_n}(y) = \mathrm{P}[Y_n \leq y] = (1 - p_{MATCH}^y)^n \qquad \mathrm{P}[Y_n \geq y] = 1 - (1 - p_{MATCH}^y)^n$$

For a formal significance test of the hypothesis H_0, we may again use the maximum run length as the test statistic, and compute a p-value p

$$p = \mathrm{P}[Y_n \geq y_n] = 1 - (1 - p_{MATCH}^{y_n})^n$$

For two sequences of length N, we have that is $n \approx (1 - p_{MATCH})N$, giving that

$$p \approx 1 - (1 - p_{MATCH}^{y_n})^{(1-p_{MATCH})N} \approx 1 - \exp\{-(1 - p_{MATCH})N p_{MATCH}^{y_n}\}$$

For a test at the $\alpha = 0.05$ significance level, we must check whether the computed p-value is less than α. If it is, we reject the hypothesis H_0 that the sequence is random and independent.

For the sequences above, $y_n = 3$, and if one of the assumptions of H_0 is that $p_A = p_C = p_G = p_T = \frac{1}{4}$ so that $p_{MATCH} = \frac{1}{2}$, then the approximate p-value is

$$1 - \left(1 - \left(\frac{1}{2}\right)^3\right)^{26/2} = 0.824$$

which is **not** less than $\alpha = 0.05$, and so there is **no evidence to reject** H_0, and hence there is no evidence of alignment (or homology).

Local Ungapped Pairwise Alignment

For a probabilistic assessment of the degree of local alignment of two sequences, we carry out a similar hypothesis test, using match statistics as test statistics. Instead of considering runs of matches ("successes"), we consider runs of partial matches. as follows. Let random variable X be defined as the length of a subsequence of the original sequence that contains **at most k mismatches** ("failures"). The length X of the subsequence id the number of positions up to but not including the $(k+1)$st mis-match; under H_0, the probability distribution of such a variable is the **Negative Binomial** (generalized geometric) distribution , and the probability distribution (cdf) of X is essentially given by the negative binomial mass function as

$$F_X(x) = \mathrm{P}\left[X \leq x\right] = \sum_{j=k}^{x} \binom{j}{k} p_{MATCH}^{j-k}(1-p_{MATCH})^{k+1} \qquad x = k, k+1, k+2, ...$$

For a test statistic, we will again consider a maximum order statistic $Y_n = \max\{X_1, ..., X_n\}$ derived from the analysis of n subsequences.

Unfortunately, the usual order statistics computations and approximations cannot be applied in general here, as the analysis would involve the study of **overlapping** sequences, giving variables $X_1, X_2, X_3...$ etc that are **not independent.** In fact, the simplest way to proceed in the calculation of p-values is to use **simulation** to generate a look up table of probabilities with which the observed maximum run-length can be compared. Broadly, the simulation proceeds as follows:

Monte Carlo simulation

(1) Generate two sequences of length N under the model implied by H_0, that is, two independent sequences with probabilities p_A, p_C, p_G, p_T for each nucleotide.

(2) For different values of k ($k = 0, 1, 2, 3, ...$) trawl the sequences to find the longest (contiguous) subsequence which contains at most k mismatches. Record the length y of this longest subsequence.

(3) Return to (1) and repeat a large number (1000000 say) of times and form a two-way table of the frequencies with which the longest subsequence containing k mismatches was of length y ($y = k, k+1, k+2, ...$).

(4) Convert the frequencies into probabilities by dividing by the number of times the simulation is repeated.

This **Monte Carlo simulation** procedure often gives a probability distribution of sufficient accuracy to allow a p-value to be computed.

4.3.3 Alignment Algorithms

It is possible to align a pair of sequences without explicit reference to probabilistic rules. The general approach to deterministic alignment is to utilize a **scoring rule** that quantifies the degree of alignment between the sequences. One simple scoring rule scores the alignment by computing the difference between the number of matches and the number of mismatches and wild-cards. More sophisticated algorithms score matches according to the similarity between the characters at a putative match position.

Two common "Dynamic Programming" algorithms are used to perform deterministic alignments; they are

- The Needleman - Wunsch algorithm
- The Smith - Waterman algorithm

Both algorithms can be modified to incorporate gaps in alignments and to penalize them appropriately. The Smith-Waterman algorithm is regarded as a superior algorithm as it computes the optimal alignment between two sequences in a shorter computational time. Neither algorithm is explicitly probabilistic in nature, but can be regarded (loosely) as a form of maximum probability/maximum likelihood procedure.

4.3.4 Protein Sequences and Substitution Matrices

For the alignment of protein sequences, special scoring rules (in the sense described above) that correctly quantify the degree of mismatch between two amino acid residues that may lie in corresponding positions in the two sequences. The scoring rule used **is** probabilistic in nature; it quantifies and utilizes the relative probabilities that two residues x and y say appear in corresponding positions due to actual homology and substitution compared with random chance positioning.

Let p_x and p_y denote, respectively, the probabilities that residues x and y appear in a given position in a protein sequence, and let p_{xy} denote the probability that x and y appear in that position due to homology and substitution. Consider the quantity

$$e_{xy} = \begin{cases} \dfrac{2p_x p_y}{p_{xy}} & x \neq y \\[2ex] \dfrac{p_x p_y}{p_{xy}} & x = y \end{cases}$$

that computes the **relative likelihood ("likelihood ratio")** of chance alignment over homology. Finally, let

$$s_{xy} = -2 \log_2 e_{xy}$$

then s_{xy} is a scoring function that is positive if homology is more likely than chance alignment, and is negative if the converse is true.

The scores s_{xy} can be used as the basis for protein sequence alignment. Typically, the probabilities p_x, p_y and p_{xy} that appear in the equations above are either estimated using **maximum likelihood** from protein sequence data, or derived from look-up tables such as the **Point Accepted Mutation (PAM)** that is computed using maximum likelihood in a **Markov Chain** model of the evolutionary processes involved, or **Block Substitution Matrices (BLOSUM)** The BLOSUM50 matrix is given below and corresponds to the values s_{xy} rounded to integer values.

THE BLOSUM50 SUBSTITUTION MATRIX