Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit

3. Results

- The method is illustrated on human α globin and β globin, that are 141 and 146 amino acids long, respectively.

- The results are summarized in the next table.

3.1 Computational Results

- Maximum likelihood alignment is more expensive than parsimony/similarity.

- There are 3 important factors that affect the time of performing a likelihood alignment:

  - (1) the number of entries in the matrix to be calculated;

  - (2) the number of evaluations needed to find the maximum likelihood estimate; and

  - (3) the time used in calculating the basic recursion.

- To determine the maximum likelihood alignment, recall that the likelihood calculations were confined to a band close to the similarity based alignment. This significantly reduces the number of matrix entries that need to be calculated.

- To find the maximum likelihood estimate of the parameters, we need the following:

  - (i) an initial guess of parameters;

  - (ii) an algorithm searching for the minimum; and

  - (iii) a stopping criteria.

• Finding the maximum likelihood estimate can be reduced to a two-variable optimization problem by observing that the expected length of a sequence is $\lambda/(\mu - \lambda)$. This fixes the ratio of λ/μ to:

$$\frac{\lambda}{\mu} = \frac{l_{\text{ave}}}{l_{\text{ave}} + 1}$$

• To find initial estimates of the parameters, we obtain some information from the similarity based alignment. We calculate how many gap signs, matches and mismatches would be expected and choose the parameters that give this expectations. Using this, a guess of $\mu = 0.0316$ and $s = 0.9500$ is obtained from the α globin versus β globin similarity alignment.

• Several numerical optimization methods were tried, but given a good initial guess, the best was BFGS (Broyden-Fletcher-Goldfarb-Shanno).

• Each iteration needs a series of likelihood evaluations to determine first and second derivatives of the likelihood function. The total number of likelihood evaluations is typically less than 50.

• The stopping condition used was the following: When an iteration produced changes in the likelihood estimates that was less than 10^{-3}, the iterations were stopped.

• The likelihood recursions of TKF91 are a bit more complicated than the recursion in the similarity based alignment. Comparison between the two indicated that the likelihood recursion was 50-70 times slower than the optimization recursion.

• To summarize, the above improvements yield a method that is significantly faster than the one described by Thorne *et al.* (1991). In absolute terms, two proteins of length 1500, can be analyzed in less than 5 seconds on a fast desktop computer (Silicon Graphics Octane with a 300 MHz R12000 processor).

3.2 Homology Test

3.2.1 Parsimony/similarity alignment based test

- In the *shuffle test*, the two sequences are aligned and a score is obtained (Doolittle, 1986). The significance of this score is evaluated by permuting the order of the amino acids and aligning the permuted proteins. This is done many times and the real score is compared to the score of the permuted proteins. If the score for the real sequences are much better, the proteins are assumed homologous.

- Drawbacks of this approach: (1) It is dependent on having the correct alignment. (2) It is hard to introduce more realistic models of sequence evolution in this test.

3.2.2 Statistical alignment based test

• Two competing hypotheses: Are they independent sequences from the equilibrium distribution on the set of sequences? Or are they related by a tree with a root finitely far back in time?

• The test will be parametric bootstrap as described by Cox (1962):

- (1) All parameters, $(\lambda_{\text{real}}, \mu_{\text{real}}, s_{\text{real}})$ are estimated by maximum likelihood for the given sequences.

- (2) Pairs of independent sequences, $(s^{(1)}, s^{(2)})_i$, are simulated using these parameters.

- (3) These simulated sequences are analyzed using statistical alignment and parameters are re-estimated, (λ_i, μ_i, s_i).

• The following statistic is calculated for the real sequences and for the simulated sequences:

$$U = -2 \ln \frac{P(s^{(1)}, s^{(2)})}{P(s^{(1)})P(s^{(2)})}$$

Now the value for the real sequences can be compared to the distribution of the values for the simulated sequences. If the value for the real sequences is extreme for the distribution, the sequences are homologous.

• Drawbacks of this method: (1) It is much slower than database scanning programs; (2) The TKF91 process is not a realistic model for sequence evolution.

3.2.3 Goodness of fit-testing the TKF process

- Since the new aspect of the TKF91 process is the indel process, we will focus on testing whether this aspect of sequence evolution is well-modelled.

- Recall that the TKF91 model assumes that insertions and deletions occur in steps of one.

- Goodness of fit testing indel lengths in the TKF model suggests that longer indels occurred.

4. Discussion

• This article has highlighted the following advantages of a more statistical approach to alignment:

- (1) It is explicitly founded on a description of molecular evolution.

- (2) Parameters are estimated and biologically meaningful.

- (3) Different evolutionary events can be assigned probabilities.

- Possible extensions and improvement for the present model:

  - (1) Incorporate indels longer than one nucleotide or amino acid into the model.

  - (2) Generalize the TKF91 dynamic programming algorithm, calculating the likelihood for a set of homologous sequences.

  - (3) Model substitutions and indels that are unevenly distributed along the sequences.

  - (4) Scrap the idea of using an immortal link. That is, we should let a sequence be born from a given equilibrium distribution and to be killed according to some process.

  - (5) Model not only the biological sequence but also the probability that a sequence with one function obtains another function.