Maximum Likelihood Estimation for Targeted Homology Search

Peter Menzel1,2, Jan Gorodkin1, Peter F. Stadler2

Abstract

Modelling the characteristic and conserved motifs of genes is in many cases still a manual task that requires expertise and constrains large scale genome annotations by homology search. We suggest an approach for creating models which are suitable for searching in a particular phylogenetic branch by calculating residue probabilities based on a multiple sequence alignment from the seed sequences.

Targeted homology search

Typical sequence models for homology search do not take phylogeny into account. To increase the specificity of search patterns, we suggest an approach for building models designated to be used in one particular phylogenetic branch by taking the relative position of the target species (X) to the species with known sequences (1 ... 5) into account.

Estimating PSSMs by Maximum Likelihood

We employ a maximum likelihood algorithm, which, given a phylogenetic tree and a multiple sequence alignment, calculates the residue probabilities at each alignment position for the target species. These probabilities can be converted into PSSMs for homology search tools, e.g. fragrep. Given a multiple alignment M with m sequences and a phylogenetic tree T with m + 1 leaves, our approach follows two steps: First we use M and T, X to numerically estimate a relative substitution rate $\mu_X$ for each alignment column $i$, so that $\hat{\mu}_i = \arg \max_{\mu} L_{\text{root}}(\mu)$. The computation of the likelihood $L_{\text{root}}(\mu)$ of the tree follows Felsenstein’s pruning algorithm, where the likelihood of a residue $s_{ijk}$ at the interior node k is obtained from the likelihoods at the two child nodes i and j, which have distances to k of $t_i$ and $t_j$, respectively:

$$L_X(\mu) = \left( \sum_{s_i} P_{s_{s_i}(t_i, \mu)} L_{s_i}(\mu) \right) \times \left( \sum_{s_j} P_{s_{s_j}(t_j, \mu)} L_{s_j}(\mu) \right)$$

The transition matrix $P$ contains probabilities $P_{xy}(t_i, \mu) = [\exp(Q_{xy})]^{t_i}$ for changing from state y to state x over time t and a rate $\mu$. The instantaneous rate matrix $Q$ represents a nucleotide substitution model, e.g. HKY85. Model parameters are estimated by standard software like PAML. In the second step, we root the tree T to the target X and use the estimated $\hat{\mu}_i$ to compute the likelihoods $L_X(\mu)$ for T and eventually obtain the residue probabilities for each alignment column in the target species. If the target is in close proximity to one or more other species, then high probabilities will be assigned to the residues from those neighbors. With increasing distance the probabilities will converge to an uninformative equilibrium distribution.

Eventualy, we can compute the information content $I(i) = 2 - H(i)$ for each alignment column i from the Shannon entropy $H(i) = -\sum_{s_i} f_i(s) \log_2 f_i(s)$ and build a search pattern from windows of a certain length that yield a user defined minimum average information content. Alignment columns with high variability (\(\mu\)) can be excluded from the search pattern.

Example for PSSM calculation

- **top**: Target sequence in the 5’ region of the 75K RNA of D. persimilis.
- **middle**: ML estimated nucleotide probabilities for this region.
- **bottom**: Nucleotide frequencies of 11 other Drosophila sequences.

The performance of the ML method was evaluated on a collection of genomic multiz alignments from the drosophila 12 genomes project (http://flybase.org). Two data sets of gap-less alignments containing sequences from all 12 drosophila species were obtained: Set contains 56 alignments with 76.1% average pairwise sequence identity and Set2 has 45 alignments with 67.1% identity. We removed one sequence at a time from each alignment and computed the residue probabilities for this sequence with our ML algorithm from the remaining 11 sequences using the phylogenetic tree below. For comparison, position frequency matrices from the same 11 species were derived. From each alignment we randomly draw 10 windows of different size and computed the MATCH scores of both PSSMs and the corresponding 12th aligned sequence that was excluded from the training set. Comparing the matching scores of both PSSMs, we find that in most cases the ML matrices perform significantly better than the frequency matrices.

Contact: ptr@genome.ku.dk

1 Division of Genetics and Bioinformatics, IBHV, University of Copenhagen
2 Bioinformatics Group and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.