Maximum Likelihood Estimation for Targeted Homology Search

Peter Menzel[1,2], Jan Gorodkin[1], Peter F. Stadler[2]

FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN

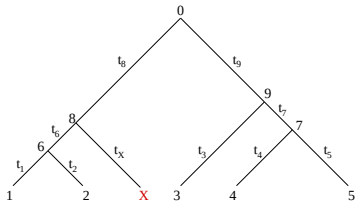
UNIVERSITÄT LEIPZIG 1409 2009 ALMA MATER LIPSIENSIS

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.

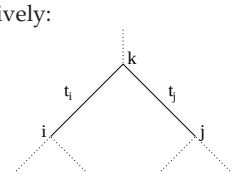


Homology search in species X using known sequences from species 1 … 5.

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 PSSM patterns 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 $T \setminus X$ to numerically estimate a relative substitution rate $\hat{\mu}_i$ for each alignment column i, so that $\hat{\mu}_i = \operatorname{argmax}_\mu L_{\text{root}}(\mu)$. The computation of the likelihood L_{root} of the tree follows Felsenstein's pruning algorithm, where the likelihood of a residue s_k 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:



$$L_{s_k}(\mu) = \Big(\sum_{s_i} P_{s_k s_i}(t_i,\mu) L_{s_i}(\mu)\Big) \times \Big(\sum_{s_j} P_{s_k s_j}(t_j,\mu) L_{s_j}(\mu)\Big)$$

The transition matrix $\mathbf{P}$ contains probabilities $P_{xy}(t,\mu) = [e^{t\mu\mathbf{Q}}]_{xy}$ for changing from state y to state x over time t and a rate μ. 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 reroot the tree T to the target X and use the estimated $\hat{\mu}_i$ to compute the likelihoods $L_X(\hat{\mu}_i)$ 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.

Eventually, we can compute the information content $I(i) = 2 - H(i)$ for each alignment column i from the Shannon entropy $H(i) = -\sum_s f_i(s) \log_2 f_i(s)$ 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 ($\hat{\mu}$) can be excluded from the search pattern.

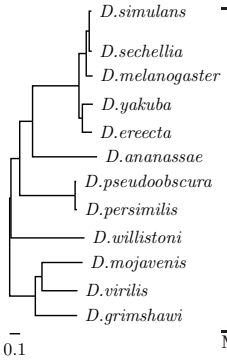
Example for PSSM calculation

top: Target sequence in the 5′ region of the 7SK RNA of *D. persimilis*.
middle: ML estimated nucleotide probabilities for this region
bottom: Nucleotide frequencies of 11 other Drosophila sequences.

Performance Evaluation

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 drosophilid species were obtained: *Set* contains 56 alignments with 76.1% average pairwise sequence identity and *Set2* has 45 alignments with 67.1% identity.

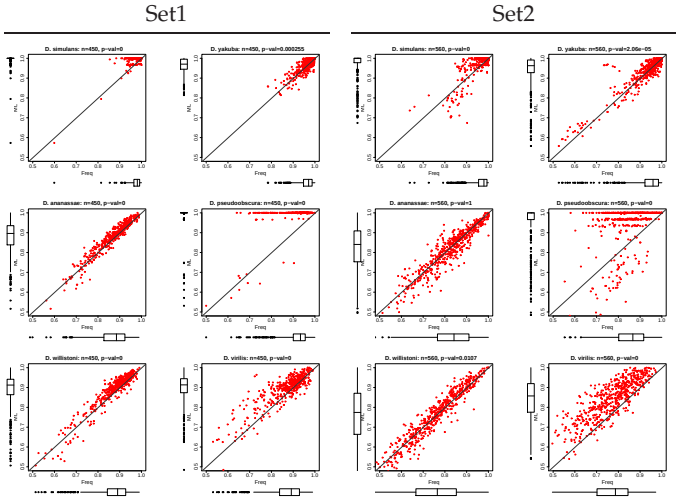
We removed one sequence at a time from each alignment and computed the residue probabilities for this sequence with our ML algorithm from the remaining 11 sequences using the phylogenetic tree below. For comparison, position frequency matrices from the same 11 species were derived. From each alignment 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 match scores of both PSSMs, we find that in most cases the ML matrices perform significantly better than the frequency matrices.



Species	Data set 1			Data set 2		
	ML	Freq	Δ	ML	Freq	Δ
D. sim.	1.000	0.981	0.019	1.000	0.980	0.020
D. sec.	1.000	0.981	0.019	1.000	0.975	0.025
D. mel.	0.986	0.979	0.007	0.970	0.972	-0.002
D. yak.	0.970	0.971	-0.001	0.963	0.959	0.003
D. ere.	0.971	0.972	-0.001	0.959	0.959	0.000
D. ana.	0.896	0.885	0.011	0.841	0.842	-0.001
D. pse.	1.000	0.933	0.067	1.000	0.867	0.133
D. per.	1.000	0.928	0.072	1.000	0.865	0.135
D. wil.	0.912	0.890	0.022	0.774	0.765	0.009
D. moj.	0.912	0.882	0.030	0.838	0.772	0.066
D. vir.	0.913	0.891	0.022	0.858	0.787	0.071
D. gri.	0.877	0.864	0.013	0.824	0.759	0.065

Median `MATCH` scores of the ML PSSMs and frequency PSSMs

The evaluation of the method on the test data set shows a significant gain of specificity of the PSSMs for the target species, even for randomly drawn samples. This improvement highly depends on the phylogenetic proximity of a known species' sequence. If the target species is evolutionary distant in the tree, it is still possible to only use those sites in the alignment, which have a high information content and the specificity is better or same compared to frequency based search patterns.



`MATCH` scores of the ML and frequency PSSMs for randomly drawn windows of length 30nt.

Contact: ptr@genome.ku.dk

[1] Division of Genetics and Bioinformatics, IBHV, University of Copenhagen
Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark
[2] Bioinformatics Group and Interdisciplinary Center for Bioinformatics,
University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.