Reconstructing Histories of Complex Gene Clusters on a Phylogeny (Supplementary Materials)

Tomáš Vinař[1], Broňa Brejová[1], Giltae Song[2], and Adam Siepel[3]

[1] Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská Dolina, 842 48 Bratislava, Slovakia

[2] Center for Comparative Genomics and Bioinformatics, 506B Wartik Lab, Penn State University, University Park, PA 16802, USA

[3] Dept. of Biological Statistics and Comp. Biology, Cornell University, Ithaca, NY 14853, USA

A Appendix: Proposal distribution

As described above, the proposal distribution for histories is defined by a sequential sampling procedure that selects groups of atom pairs to merge in each step. The goal is to define this distribution so that the overall proposal distribution is as close as possible to the actual conditional distribution $p(H'_i \mid H_{i-1})$, making the acceptance probability as close as possible to one. Directly characterizing $p(H'_i \mid H_{i-1})$ appears to be difficult, so we settle for a heuristic weighting function in the proposal distribution for merges that is designed to produce reasonably good proposed histories. The Metropolis-Hastings algorithm will ensure that the retained samples will accurately reflect the posterior distribution, once the Markov chain reaches stationarity.

In each step, we consider all possible duplications consistent with the current set of guide trees, as well as selected deletion and speciation events. Deletions do not leave observable sequence traces in extant species, and thus it is impossible to date them precisely; instead, in the proposal algorithm we associate deletions with the speciation or duplication events that occurred before them. We allow a single deletion following a duplication. We consider only deletions completely inside the source or target sequence of the duplication.

A speciation event is represented as a copy of all atomic segments from one species to a previously empty sequence of another species, possibly followed by several deletions in both species. We only allow speciations in the partial order imposed by the species tree. Additionally, we propose only speciations that maximize the total sequence length of matched atomic segments between the two species. As in the case of duplications, only segments that are currently cherries in the corresponding guide trees can be matched. For example, if we have sequences in two species $S_1 = a_1b_1c_1$ and $S_2 = a_2b_2c_2$, and b_1 and b_2 are not cherries in the segment tree, we have to propose speciation from an ancestral sequence ab_1b_2c or ab_2b_1c, followed by one deletion in species S_1 and one deletion in species S_2. Proposals that obey these constraints can be easily generated by a simple dynamic programming algorithm, and in the case of many possible speciation proposals, we only keep 20 highest weight candidates. Note that it is always possible to propose at least one event until we reach an ancestral sequence of unique atoms.

We characterize each proposed event by a feature vector $f_1, \ldots, f_k$ and the probability of choosing the event will be proportional to $\exp(\sum_i w_i f_i)$ for some fixed set of weights w_i. In the rest of this section we briefly describe these features and their weights.

Target length. The basis of the overall score is the length ℓ of duplication or speciation, i.e. how much sequence is removed by unwinding the event. We set $f_1 = \ln(\ell)$ and $w_1 = 1$.

Previously seen event. To keep the newly proposed history similar to the previous sample H_{i-1}, we add bonus to events seen in H_{i-1}. This is achieved by a binary indicator feature f_2 and weight $w_2 = \ln(10)$. Some events may not be possible in the new history due to changes in the guide trees.

Branch length mean and variance. For a given duplication consistent with the guide tree set, we can compute the mean distance μ of corresponding cherries in the guide tree (weighted by the lengths of atoms in nucleotides), and also variance on such distance σ. The lower μ indicates likely more recent events, while large variance σ would indicate that we are merging two or more events that happened at different times. We set $f_3 = \mu$, $f_4 = \sigma$, $w_3 = -10$, $w_4 = -1$.

Partial duplication penalty. If the proposed duplication is a subset of a larger duplication, we set indicator $f_5 = 1$ and use $w_5 = -\ln(100)$.

Breakpoint reuse penalty. Although, we allow breakpoint reuse, we favor duplications with fewer breakpoint reuses which seems to be particularly useful for determining correct direction of duplications. We have implemented the three conditions stipulated by Zhang et al. (2008) based on collapsibility of atom pairs on boundaries of the duplicated segments. We set f_6 to the number of violated conditions and $w_6 = -\ln(10)$;

Pair reduction bonus. Consider the number π of distinct pairs of adjacent atom types that occur in the current set of sequences. For input with n atom types, $\pi = n - 1$ when we reach the ancestral sequence, and each duplication reduces π by at most 2. This gives us a lower bound on the number of events necessary to reach the ancestral sequence. We set f_7 to be the reduction of π achieved by the event (f_7 can be negative if π increases) and $w_7 = \ln(10)$.

Deletion penalties. Deletion associated with a duplication is penalized by setting $f_8 = 1$ and $w_8 = -\ln(10)$. In addition, we penalize longer deletions by setting $f_9 = \ln(d/(d+\ell))$ and $w_9 = 3$ where ℓ is the length of the target sequence in the duplication, and d is the length of the deletion. Each deletion associated with a speciation is penalized by setting $f_{10} = 1$ and $w_{10} = -\ln(1000)$.

Heat constants. Finally, in some rounds of the MCMC sampler, we want to explore radically new histories, while in other rounds we want to concentrate on smaller local improvements. Thus, we exponentiate the final event weights to a heat constant, which changes from round to round. In our experiments, we have used cyclic sequence of heats $(0.5, 0.6, 1, 1.2)$.

Bibliography

Zhang, Y., Song, G., Vinar, T., Green, E. D., Siepel, A., and Miller, W. (2008). Reconstructing the Evolutionary History of Complex Human Gene Clusters. In *Research in Computational Molecular Biology (RECOMB)*, pages 29–49.

B Appendix: Supplementary Figures

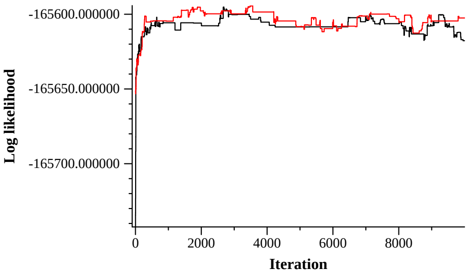

Fig. B1. Convergence of the MCMC sampler. Log likelihood as a function of iteration number for two independent chains with random starting points on a slowly evolving simulated cluster.

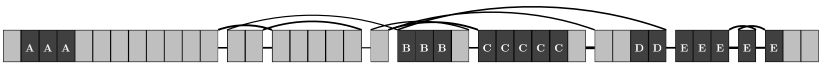

Fig. B2. Ancestral sequence reconstruction for PRAME. The cartoon shows large blocks of consecutive atomic segments, with block size proportional to the number of atoms per block. The blocks are ordered according to the highest posterior ordering and the alternative edges show other possible pairs of adjacent atoms with > 25% posterior probability. The atoms spanning five ancestral genes at 90% similarity are marked A-E.

Table B1. Distribution of events along the individual branches of the phylogeny. The table shows a histogram of the differences between the actual and the expected number of events computed from the MCMC samples.

	rate 200 (slow)					**rate 300 (fast)**				
Branch	< 0	0	1	2	> 2	< 0	0	1	2	> 2
Duplications:										
human	0	20	0	0	0	0	20	0	0	0
hominid	1	19	0	0	0	5	15	0	0	0
chimp	0	20	0	0	0	0	20	0	0	0
macaque	6	13	1	0	0	2	18	0	0	0
root	0	15	5	0	0	0	17	2	1	0
total	3	16	0	1	0	4	16	0	0	0
Deletions:										
human	0	20	0	0	0	0	20	0	0	0
hominid	1	16	3	0	0	0	15	5	0	0
chimp	0	20	0	0	0	0	20	0	0	0
macaque	0	18	2	0	0	1	18	0	1	0
root	0	19	1	0	0	0	20	0	0	0
total	0	17	1	2	0	1	12	6	1	0