New polynomial time methods for whole genome phylogeny reconstruction

Li-San Wang and Tandy Warnow
Department of Computer Sciences, University of Texas at Austin

Abstract

Introduction. Evolution operates on whole genomes by operations that change the order and strandedness of genes within the genomes. This type of data presents new opportunities for discoveries about deep evolutionary events, provided that sufficiently accurate methods can be developed to reconstruct evolutionary trees in these models [2, 5, 6, 4, 7, 9]. A necessary component of any such method is the ability to accurately estimate *true evolutionary distances*, which is the number of events that took place in the evolutionary history between a given pair of genomes. In this paper, we present a very general analytical technique for estimating true evolutionary distances between genomes, whether signed or unsigned, circular or linear, and for any relative probabilities of event classes. This simulation study shows that these distance estimation techniques improve the accuracy of the phylogenetic trees reconstructed by neighbor joining [1, 10].

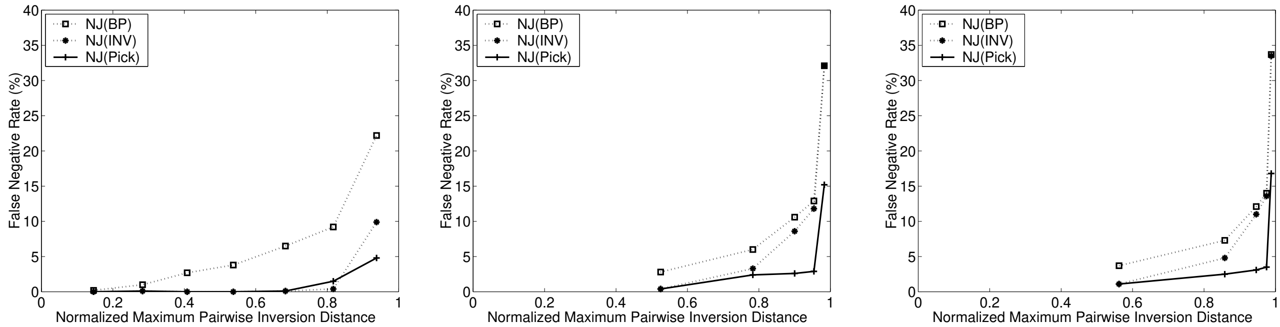
The Nadeau-Taylor Model of Genome Evolution. In the *Nadeau-Taylor* model [8], rearrangement events are divided into *rearrangement classes*. Rearrangement events in the same class occur with equal probability, and the number of events on each edge obeys a Poisson distribution. We consider the following three classes of rearrangements: inversions, transpositions, and inverted transpositions. Each model tree has a triplet of parameters $\gamma_I, \gamma_T, \gamma_{IT}$, where $\gamma_I + \gamma_T + \gamma_{IT} = 1$. These parameters indicate the probability that a random event is an inversion, transposition, and inverted transposition, respectively. The number of events on each edge e is Poisson distributed with mean λ_e. This process produces a set of signed gene orders at the leaves of the model tree.

True Evolutionary Distance (*t.e.d.*) Estimation. Let T be the true tree on which the two genomes, G_i and G_j, evolved. Let P_{ij} between G_i and G_j in T, and let k_e be the actual number of events along edge $e \in P_{ij}$. The *true evolutionary distance* (*t.e.d.*) between G_i and G_j is $k_{ij} = \sum_{e \in P_{ij}} k_e$. If we can estimate all k_{ij} sufficiently accurately, we can reconstruct the tree T using very simple methods, and in particular, using the neighbor joining method (NJ) [10]. Estimates of pairwise distances that are close to the true evolutionary distances will in general be more useful for phylogenetic (i.e. evolutionary) tree reconstruction than edit distances, because edit distances *underestimate* true distances significantly as number of events increases.

The importance of obtaining good *t.e.d.*-estimates when analyzing DNA sequences (under stochastic models of DNA sequence evolution) is understood, and well-studied [12]. However, very little work has been done on obtaining *t.e.d.*-estimates between whole genomes; indeed, only the special case of inversion distances has been solved [3, 11].

In this paper we presents a general analytical technique for estimating the expected number of breakpoints produced by a sequence of random events in the Nadeau-Taylor model with arbitrary γ_I, γ_T, and γ_{IT}. The technique applies to datasets that are either signed or unsigned, circular or linear, and for many other event classes. These techniques have two steps:

- **Step 1:** Given a pair of genomes G and G', compute the breakpoint distance.
- **Step 2:** Estimate k, the number of events along the simple path P in the evolutionary tree between G and G'.



(a) Inversions only

(b) Transpositions and inverted transpositions equally likely

(c) Inversions, Transpositions, and inverted transpositions equally likely

Figure 1: Neighbor joining performance on distance corrections. The number of genes is 120.

Computing the number of breakpoints is easy, and can be done in linear time; finding k can be done by numerical methods such as bisection methods.

Technical Development. For any genome X, define the function $B_i(X)$ as follows: $B_i(X) = 1$ if there is a breakpoint between genes i, $i+1$ in genome X, $B_i(X) = 0$ otherwise. Let X_k be the genome by applying a random sequence of k events on X_0. All X_k's have n genes; no events change the gene content. We are unable to obtain an exact calculation of $E[BP(X_k, X_0)]$, and so instead we produce an estimate $\mathcal{F}_k$. The following theorem shows $\mathcal{F}_k$ has low error:

Theorem 1 *Let $\mathcal{E}$ be a class of rearrangement events such that for any genome X, $\rho \in \mathcal{E}$, and breakpoint position i, $\Pr(B_i(\rho X) = 1 \mid B_i(X) = 0)$ is independent of X. Then $|\mathcal{F}_k(\mathcal{E}) - E[BP(X_k, X_0)]| = O(1)$.*

Denote $\mathcal{H}$ for the class of rearrangement events of Nadeau-Taylor model of evolution, associated with any setting of γ_I, γ_T, and γ_{IT}. We have a stronger result.

Theorem 2 $|\mathcal{F}_k(\mathcal{H}) - E[BP(X_k, X_0)]| \leq 1 + \frac{1}{n-1}$ *for all $k \geq 0$.*

NJ performance simulation study. In Figure 1 we present the results of our simulation study in which we explored the topological accuracy of NJ trees obtained using the following three techniques for estimating distances: (1) BP, the breakpoint distance, (2) INV, the inversion distance, and (3) Pick: pick the distance matrix of INV if $\max\{\text{INV}(i,j)\} \leq 0.8n$, otherwise use the distance matrix returned by the *t*.e.d. estimator on $\mathcal{F}_k$, where n is the number of genes. The number of genes in each genome is 120, and the number of genomes in each dataset is 10, 20, 40, or 80. We measure error in terms of false negative rates, which is the percentage of edges in the model tree missing from the reconstructed tree.

Note that neighbor joining obtains highly accurate trees using the estimates of the actual number of events, even when the maximum normalized edit distance is close to saturation (i.e. close to 100%). By comparison, neighbor joining obtains much poorer estimates of the true tree using breakpoint distances, and demonstrates a sharp decline in accuracy as the input approaches saturation. The performance of neighbor joining under minimum inversion distances is interesting: it is better, in some cases, than our theoretically obtained estimates of the actual number of events, but not when the datasets are close to saturation. For these datasets, neighbor joining does best using our estimates of pairwise distances.

References

[1] Atteson, K., "The performance of the neighbor-joining methods of phylogenetic reconstruction," *Algorithmica* **25**, 2/3 (1999), 251–278.

[2] Blanchette, M., Bourque, G., & Sankoff, D., "Breakpoint phylogenies," in *Genome Informatics 1997*, Miyano, S., & Takagi, T., eds., Univ. Academy Press, Tokyo, 25–34.

[3] Caprara, A., & Lancia, G., "Experimental and statistical analysis of sorting by reversals," in *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds., Kluwer Academic Publishers (2000), 171–184.

[4] Cosner, M. E., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., Wang, L.-S., Warnow, T., and Wyman, S., "A new fast heuristic for computing the breakpoint phylogeny and a phylogenetic analysis of a group of highly rearranged chloroplast genomes," in *Proc. 8th Intl. Conf. on Intel. Sys. for Mol. Bio. ISMB 2000.*

[5] *DCAF: Gene Order Dynamics, Comparative Maps, and Multigene Families*, workshop proceedings available as *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds., Kluwer Academic Publishers (2000).

[6] Downie, S.R., & Palmer, J.D., "Use of chloroplast DNA rearrangements in reconstructing plant phylogeny," in *Plant Molecular Systematics*, Soltis, P., Soltis, D., and Doyle, J.J., eds., Chapman & Hall, NY (1992), 14–35.

[7] Moret, B.M.E., Wyman, S.K., Bader, D.A., Warnow, T., & Yan, M., "A new implementation and detailed study of breakpoint analysis," *Proc. 6th Pacific Symp. Biocomputing PSB 2001*, Hawaii (2001)

[8] Nadeau, J.H., & Taylor, B.A., "Lengths of chromosome segments conserved since divergence of man and mouse," *Proc. Nat'l Acad. Sci. USA* **81** (1984), 814–818.

[9] Olmstead, R.G., & Palmer, J.D., "Chloroplast DNA systematics: a review of methods and data analysis," *Amer. J. Bot.* **81** (1994), 1205–1224.

[10] Saitou, N., & Nei, M., "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Mol. Biol. & Evol.* **4** (1987), 406–425.

[11] Sankoff, D., & Blanchette, M., "Probability Models for Genome Rearrangements and Linear Invariants for Phylogenetic Inference," *Proc. 3rd Int'l Conf. on Comput. Mol. Bio. RECOMB99*, ACM Press, NY (1999), 302–309.

[12] Swofford, D., Olson, G., Waddell, P., and Hillis, D., "Phylogenetic inference," in *Molecular Systematics* (2nd ed.), D. Hillis, C. Moritz, and B. Mable, eds., Sinauer Associates Inc., Sunderland, Mass. (1996), chap. 11.