Conditioned genome reconstruction: how to avoid choosing the conditioning genome

Matthew Spencer[1,2], David Bryant[3,4], Edward Susko[1]

1:Department of Mathematics and Statistics, Dalhousie University
2:Department of Biochemistry and Molecular Biology, Dalhousie University
3:McGill Centre for Bioinformatics, McGill University
4:Department of Mathematics, University of Auckland

GenomeAtlantic

Summary

Whole-genome phylogenies can be reconstructed from patterns of gene presence and absence. Because genomes differ greatly in size, we want to avoid using a stationary model. Logdet distances do not assume stationarity, and therefore seem a good choice. However, an unknown number of genes are absent from every taxon. If we do not take account of these genes, we will get the wrong distances even with infinite data, and might therefore get the wrong topology. It has been suggested that we could use only those genes present in a conditioning genome, which is excluded from the resulting tree. We prove that this approach will give consistent estimation of topology. However, we show that the choice of conditioning genome matters. Finally, we describe methods by which we can avoid choosing a conditioning genome, and show that these work well.

Conditioned logdet distances

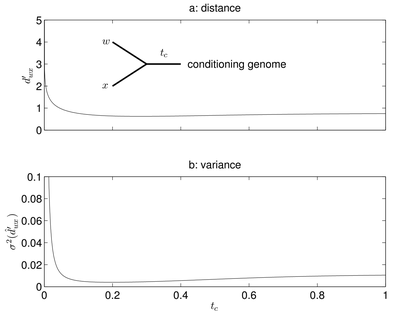
The data for a pair of taxa are the frequencies of patterns 00, 01, 10, 11, where 0 indicates absence and 1 indicates presence of a gene. Stationary models of gene gain and loss are not ideal, because different taxa have very different genome sizes. Logdet distances (Lake, 1994; Lockhart et al., 1994) do not assume stationarity. However, some of the 00 genes will be absent from every taxon, so we will not know that they existed. We will therefore get the wrong distances. Conditioned genome reconstruction (Lake and Rivera, 2004; Rivera and Lake, 2004) has been suggested as a way to overcome this. Logdet distances are calculated using only those genes present in a conditioning genome, that cannot be included in the tree. Does it work, and which conditioning genome should we use?

Theory

Suppose that taxa w and x are both connected to a common ancestor v. A distance measure is tree-additive if $d_{wx} = d_{wv} + d_{vx}$, and is non-negative if $d_{wx} > 0$ if $w \neq x$, and 0 otherwise. If a distance measure is tree-additive and non-negative, most distance methods will recover the correct tree topology.

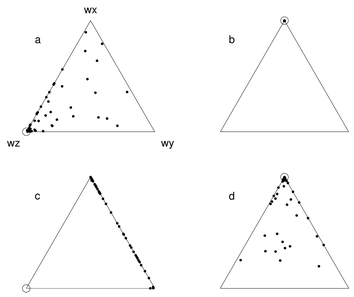
Standard logdet distances (e.g. for nucleotide data) are tree-additive and non-negative, even when different taxa have different nucleotide frequencies. We proved that conditioned logdet distances have the same property, for almost any choice of conditioning genome.

If we had infinite data we would therefore get the correct topology for any choice of conditioning genome. However, the pairwise distances that we estimate depend on the distance from the conditioning genome to the taxa of interest. In the example below, t_c is the distance from the conditioning genome to the path connecting w and x. With small t_c, we estimate a large distance between w and x (a), with a large variance (b). When the variance is large, there may be strong small-sample bias from finite data.



The choice of conditioning genome matters

We analyzed four-taxon subsets of a bacterial genome database. Different conditioning genomes can give strong bootstrap support for different tree topologies. In the figure below, each point is from one choice of conditioning genome, and the vertices represent 100% support for one of the topologies. In (b), we get the right topology no matter what conditioning genome we choose (two of the taxa are from the same genus and we always put them together). In (a) and (d), we sometimes get the right topology, but sometimes get the wrong topology with strong bootstrap support. In (c), we always get the wrong topology (two of the taxa are parasites/endosymbionts, and tend to get grouped together because they have lost similar sets of genes).



Bootstrap proportions of the three topologies wx, wy and wz were estimated using conditioned logdet distances and unweighted least-squares for four-taxon subsets of the 50-taxon, 4873-gene-family bacterial genome database COG (Tatusov et al., 2003). Open circles at the vertices indicate what we think are the correct topologies. The four-taxon data sets were:

(a) w=*Synechocystis* sp., x=*Escherichia coli* K12, y=*Mesorhizobium loti*, z=*Mycoplasma genitalium*;

(b) w=*Bacillus subtilis*, x=*Bacillus halodurans*, y=*Haemophilus influenzae*, z=*Pasteurella multocida*;

(c) w=*Aquifex aeolicus*, x=*Yersinia pestis*, y=*Buchnera* sp. APS, z=*Ureaplasma urealyticum*;

(d) w=*Corynebacterium glutamicum*, x=*Lactococcus lactis*, y=*Salmonella typhimurium* LT2, z=*Campylobacter jejuni*. In each case, each of the remaining 46 bacterial taxa from the COG database was used as a conditioning genome, and 1000 bootstrap replicates were run after conditioning.

How to avoid having to choose

If we have m taxa, we can use each one in turn as the conditioning genome, and get a set of m distance matrices, each with one taxon missing. We developed two supertree methods to get a topology for all m taxa. Both are consistent (will get the correct topology given infinite data). Neither gives us edge lengths on the full topology, because edge lengths from different conditioning genomes are not estimates of the same quantity.

Summing over subtrees

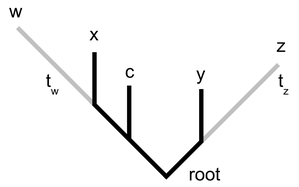
For infinite data, the sum of squares is zero on the true topology of every $(m-1)$-taxon subtree, and every m-taxon tree has a unique set of $(m-1)$-taxon subtrees. We can sum the objective function over all subtrees, and choose the topology for which this sum is minimal. This method is consistent because the sum of sums of squares over all subtrees is zero on the true topology alone.

Modified BIONJ

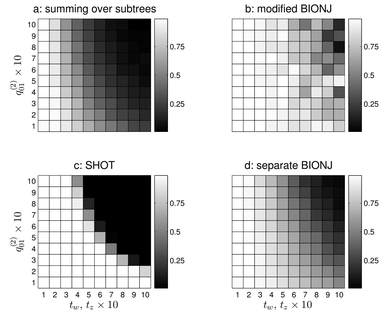
BIONJ (Gascuel, 1997) aggregates pairs of taxa, choosing the pair at each step that minimizes the sum of edge lengths. We first choose a candidate pair of taxa to aggregate from every distance matrix using the BIONJ criterion. We then pick the candidate pair that was chosen most often, and aggregate the subtrees containing this pair in every distance matrix. This method is consistent because BIONJ is consistent, and every candidate pair of taxa will be a correct pair given sufficient data.

Which method is best?

We simulated data on the five-taxon tree below. We used one set of transition rates on the gray edges, and another set everywhere else. We also varied the edge lengths t_w and t_z. As the rate $q_{01}^{(2)}$ of gene gain on t_w and t_z increases and t_w and t_z get larger, w and z tend to have larger genomes.



We compared four methods for recovering the topology: summing over subtrees, modified BIONJ, BIONJ on SHOT distances, and separate BIONJ on each conditioned logdet distance matrix. SHOT distances (Korbel et al., 2002) are not tree-additive, but sometimes do well in practice and were designed to deal with variation in genome size. Separate BIONJ is BIONJ on the distance matrix from each conditioning genome separately, scored as correct if all the subtrees were correct. This corresponds to the original conditioned genome reconstruction method.



We simulated sets of 5000 genes, with 1000 genes expected present at the root. Lighter colours mean more frequent recovery of the true topology. The average order of performance was modified BIONJ > separate BIONJ > SHOT > summing over subtrees. All methods do worse when $q_{01}^{(2)}$ is large and t_w and t_z are long, so that w and z tend to have large genomes and all other taxa have small genomes. When this happens, w and z are often placed together. SHOT was most affected, and adding more data made the problem worse. Modified BIONJ was least affected.

Conclusions

- The choice of conditioning genome can matter in practice.
- Using a supertree method, we can avoid having to choose a conditioning genome
- Modified BIONJ did best, outperforming conditioned genome reconstruction and SHOT.

References

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695.

Korbel, J. O., Snel, B., Huynen, M. A., and Bork, P. (2002). SHOT: a web server for the construction of genome phylogenies. *Trends in Genetics*, 18(3):158–162.

Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences*, 91:1455–1459.

Lake, J. A. and Rivera, M. C. (2004). Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Molecular Biology and Evolution*, 21(4):681–690.

Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, 11(4):605–612.

Rivera, M. C. and Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431:152–155.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003). The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.