Bioinformatics Advance Access published October 11, 2012

TIBA: A Tool for Phylogeny Inference from Rearrangement Data with Bootstrap Analysis

Yu Lin[1], Vaibhav Rajan[1]*, Bernard Moret

Laboratory for Computational Biology and Bioinformatics, EPFL,
EPFL-IC-LCBB INJ 230, Station 14, CH-1015 Lausanne, Switzerland
[1]who contributed equally to this work

Associate Editor: Prof. David Posada

ABSTRACT

Summary: TIBA is a tool to reconstruct phylogenetic trees from rearragement data which consists of ordered lists of synteny blocks (or genes) where each synteny block is shared with all of its homologs in the input genomes. The evolution of these synteny blocks, through rearrangement operations, is modeled by the uniform Double-Cut-and-Join model. Using a true distance estimate under this model and simple distance based methods TIBA reconstructs a phylogeny of the input genomes. Unlike any previous tool for inferring phylogenies from rearrangement data, TIBA uses novel methods of robustness estimation to provide support values for the edges in the inferred tree.
Availability: http://lcbb.epfl.ch/softwares/tiba.html
Contact: vaibhav.rajan@epfl.ch

1 INTRODUCTION

"Rare genomic changes" such as rearrangements (Rokas and Holland, 2000) cause large-scale structural changes in the genome, clarify distant or problematic relationships among organisms and have been used in many phylogenetic studies.

The first algorithm for phylogeny inference from rearrangement data was BPAnalysis (Blanchette *et al.*, 1997). The algorithm seeks to reconstruct the tree and ancestral genomes with the minimum *breakpoint distance* along each edge of the tree. This approach was extended in GRAPPA (Moret *et al.*, 2001) by using *inversion distances*. These methods were restricted to unichromosomal genomes; the tool MGR (Bourque and Pevzner, 2002) was the first to handle multichromosomal genomes. All these parsimony-based approaches must produce good approximations to the NP-hard problem of computing the rearragement median of three genomes (summarized in (Tannier *et al.*, 2008)) which limits their scalability. Despite using clever heuristics, MGR does not scale well particularly for high resolution data. Yet to date MGR (and its more recent derivative MGRA (Alekseyev and Pevzner, 2009)) had remained the only tool available for the analysis of multichromosomal genomic rearrangements.

Distance-based methods like Neighbor Joining (NJ) (Saitou and Nei, 1987), in contrast with parsimony-based methods, run in time polynomial in the number and size of genomes—and fast and accurate heuristics exist for those where the scoring function cannot be computed in polynomial time, such as FastME (Desper and Gascuel, 2004). Pairwise distances, given as input to a distance-based method, are usually the edit distances, that is, the minimum-cost distances under the assumed model of evolution. However, an edit distance typically underestimates the true distance causing poor accuracy of trees inferred from distance-based methods. When given true evolutionary distances, NJ provably returns the true tree. The true evolutionary distance—the actual number of evolutionary events between the two genomes—is impossible to measure but can be estimated using statistical techniques. Such *distance corrections* have long been used for sequence data and, more recently, for rearrangement data. For multichromosomal genomes, we have designed such an estimator (Lin and Moret, 2008) and have demonstrated the accuracy and scalability of a reconstruction method that uses NJ with this distance estimator (Lin *et al.*, 2010).

A major shortcoming of phylogeny reconstruction from rearrangement data has been the lack of any way to assess the robustness of the inferred edges. In phylogenetic analysis from sequence data, such an assessment is *de rigueur* and is carried out by an adaptation of the standard non-parametric tests—the bootstrap and the jackknife, first proposed by Felsenstein (Felsenstein, 1985). Recently, we have designed several new methods for statistically assessing the robustness of trees reconstructed from rearrangement data (Lin *et al.*, 2010, 2011). Through careful and extensive experiments we have shown that our bootstrapping approach for rearrangement data is on par with the classic phylogenetic bootstrap used in sequence-based reconstruction. Combining these methods with our distance based reconstruction method, we provide the first tool for phylogeny inference from rearrangement data that is accurate, scalable and provides bootstrap support values for the edges of the tree.

2 METHODS

Rearrangement data for a genome consists of lists of syntenic blocks (genes are an example) in the order in which they are placed along one or more chromosomes. Each syntenic block is identified by a marker, which is shared with all (or most) of its homologs in the genomes under study, and a sign which represents the strandedness of the syntenic block. The markers typically used for syntenic blocks are integers. Any two adjacent syntenic blocks can be represented by a set of two integers—we call this an adjacency. A telomere in a linear chromosome is represented by a singleton set containing just the end marker. A genome is thus represented by a set of

*to whom correspondence should be addressed

adjacencies and telomeres. Any rearrangement operation changes up to three adjacencies or telomeres in the genome. For multichromosomal genomes, all the rearrangement operations can be modeled by a single operation called "Double-Cut-and-Join" (DCJ) (Yancopoulos *et al.*, 2005).

For reconstruction, we use either NJ or FastME. The pairwise distances used are estimates of the true evolutionary distances under a model of evolution that assumes uniform distribution of DCJ operations. The evolutionary model is used to infer an estimate of the true distance by deriving the effect of a given number of DCJ operations on the number of shared adjacencies and telomeres and numerically inverting the derivation to produce a maximum-likelihood estimate of the true distance under the model. See (Lin and Moret, 2008) for details. Our extensive experiments on simulated and real datasets, described in (Lin *et al.*, 2010), show that the error rates of trees reconstructed by NJ using this distance estimator is below 10% in all but the oversaturated cases. With FastME, the error rates are even lower. Further, error rates are significantly reduced by an increase in the size of the genome—because the larger number of syntenic blocks reduces the relative error in the estimated distances. Trees can be reconstructed on up to 500 genomes each containing up to 10,000 markers within a few seconds on a PC.

As described in (Lin *et al.*, 2011), we design and test several bootstrapping methods that can be used with distance based reconstruction from rearrangement data. The fact that our distance estimator computes the estimated true distance between two genomes based only on the number of shared adjacencies in each genome allows us to design sampling methods for bootstrapping that can handle replicate genomes which may be invalid (e.g., because of additional copies of adjacencies), and yet be sufficient for computing the pairwise distance (by tallying the number of shared adjacencies). Two of these methods, BC and PJ are equivalent in their performance and are better than all the other methods designed. Their names come from their equivalent counterparts in the sequence world: the classic bootstrap (BC) of Felsenstein and parsimony jackknifing (PJ).

The key idea behind these bootstrapping methods is to create replicates by sampling adjacencies: from the list of all possible adjacencies, BC samples with replacement to form a collection of adjacencies; only adjacencies in this collection are then used to count the number of shared adjacencies and then estimate the pairwise distances. PJ is (asymptotically) equivalent to sampling with replacement (as in BC), but without overcounting, that is, when sampling gives an adjacency that has been previously selected, it is not added to the replicate. In other words, selected adjacencies are not counted more than once for computing the number of shared adjacencies between leaf genomes. From each replicate a tree is reconstructed using our true distance estimator and NJ or FastME. The bootstrap support of an edge (viewed as a bipartition of leaves) in the inferred tree is the proportion of the trees from replicates that contain the edge (the same bipartition of leaves). Both BC and PJ show very high sensitivity even at high levels of specificity making them excellent bootstrap methods. We have also demonstrated, in (Lin *et al.*, 2011), that they outperform jackknifing methods based on sampling markers such as (Huang *et al.*, 2010; Shi *et al.*, 2010).

3 SOFTWARE

TIBA is implemented in C++ and can be compiled and executed on the command line in any UNIX-based platform and in the Cygwin environment on Windows. To run a phylogenetic analysis, the program must be run with the following input parameters: the input filename, the output filename, the reconstruction method, and the bootstrap method. The input file format is the same as that used by GRAPPA and MGR: FASTA like headers for the names of the genomes ('>' followed by an alphanumeric sequence followed by a newline), each chromosome represented by a signed permutation of integers ending with a '$' symbol and a newline character. The output filename provided by the user is suffixed with '_1' and '_2' to create two output files, both in Newick format: the first contains the inferred tree with branch lengths and the second contains the same inferred tree with support values replacing the branch lengths. The reconstruction method can be either NJ or FastME. The two bootstrap methods dicussed above are implemented. They can be specified by 'BC' or 'PJ'. The default number of replicates is 100 but the user can change this with an additional input. Installation and usage details, with examples are provided in the package and on our website.

4 CONCLUSION

TIBA is very fast, scalable, accurate and provides support values for the edges in the inferred tree. Fast and scalable distance-based methods, precise estimates of true pairwise distances and finally, for the first time in any rearrangement-based phylogeny inference method, the use of bootstrap scores – together make TIBA unique.

ACKNOWLEDGEMENT

Funding: NA

REFERENCES

Alekseyev, M. and Pevzner, P. (2009). Breakpoint graphs and ancestral genome reconstructions. *Genome Res.*, **19**(5), 943–957.

Blanchette, M., Bourque, G., and Sankoff, D. (1997). Breakpoint phylogenies. In S. Miyano and T. Takagi, editors, *Genome Informatics*, pages 25–34. Univ. Academy Press, Tokyo.

Bourque, G. and Pevzner, P. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, **12**, 26–36.

Desper, R. and Gascuel, O. (2004). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.*, **9**(5), 687–705.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evol.*, **39**, 783–791.

Huang, Y.-L., Huang, C.-C., Tang, C. Y., and Lu, C. L. (2010). Sort2: a tool for sorting genomes and reconstructing phylogenetic trees by reversals, generalized transpositions and translocations. *Nucleic Acids Research*.

Lin, Y. and Moret, B. (2008). Estimating true evolutionary distances under the DCJ model. In *Proc. 16th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'08)*, volume 24(13) of *Bioinformatics*, pages i114–i122.

Lin, Y., Rajan, V., and Moret, B. (2010). Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator. In *Proc. 8th RECOMB Workshop Comp. Genomics (RECOMB-CG'10)*, volume 6398 of *Lecture Notes in Comp. Sci.*, pages 137–148. Springer Verlag.

Lin, Y., Rajan, V., and Moret, B. (2011). Bootstrapping phylogenies inferred from rearrangement data. In *Proc. 9th Workshop Algs. in Bioinf. (WABI'11)*, volume 6833 of *Lecture Notes in Comp. Sci.*, pages 175–187. Springer Verlag.

Moret, B., Wyman, S., Bader, D., Warnow, T., and Yan, M. (2001). A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. on Biocomputing (PSB'01)*, pages 583–594. World Scientific Pub.

Rokas, A. and Holland, P. (2000). Rare genomic changes as a tool for phylogenetics. *Trends in Ecol. and Evol.*, **15**, 454–459.

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

Shi, J., Zhang, Y., Luo, H., and Tang, J. (2010). Using jackknife to assess the quality of gene order phylogenies. *BMC Bioinformatics*, **11**(1), 168.

Tannier, E., Zheng, C., and Sankoff, D. (2008). Multichromosomal genome median and halving problems. In *Proc. 8th Workshop Algs. in Bioinf. (WABI'08)*, volume 5251 of *Lecture Notes in Comp. Sci.*, pages 1–13. Springer Verlag.

Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, **21**(16), 3340–3346.