Guide to getting good results using empirical data with STEM.

Dec 3rd, 2009

Bryan C. Carstens

Welcome to the STEM Google Group. Many of you are here because you would like to use STEM to analyze some empirical data that you have collected; typically, the data consist of several unlinked loci from multiple individuals from some number of independent groups. I have some experience with empirical data and STEM, and the following is based on this experience.

First some assumptions of STEM:

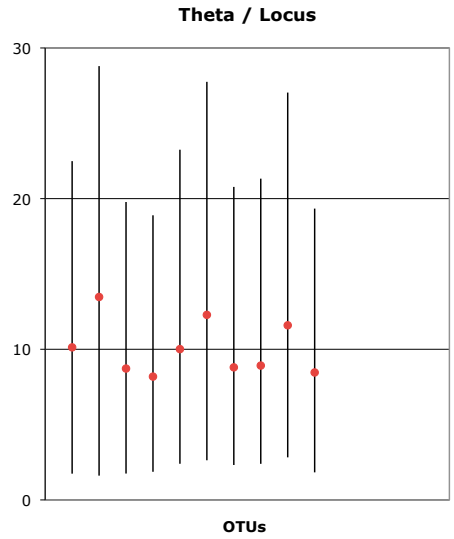
1. The OTUs of at STEM analysis are evolutionarily-independent lineages.
2. The model implemented by STEM assumes that these lineages are not exchanging migrants.
3. The model implemented by STEM assumes that $\theta = 4N\mu$ does not change during the period of time represented by the species tree.
4. The model assumes that the genealogies from each locus are evolving in a clock-like manner so that coalescent events can be compared across loci.
5. The model assumes that the genealogies from each locus are at linkage equilibrium (e.g., separated by recombination).

These assumptions have practical consequences related both the manner by which the gene trees in the '*genetree.tre*' file are estimated and some entries in the '*settings*' file.

Estimating gene trees. A single point-estimate of the genealogy is needed; this genealogy should have branch lengths estimated under an appropriate model of sequence evolution. After DT-Modsel[1] or similar program is used to select a model, estimate the gene tree using a program such as PAUP*[2] or Garli[3]. Once the gene tree is estimated, conduct a likelihood ratio test of the molecular clock[4], and if the data are clocklike enforce the molecular clock and re-estimate the branch lengths of the ML gene trees. After this has been done for each locus, the gene trees need to be rooted (either via an outgroup or at the midpoint) and fully resolved. I find that TreEdit[5] to be a useful software package for the later steps.

Geneflow. STEM, like other species-tree estimation programs, assumes that the OTUs are not exchanging migrants. How is this best tested? One solution is to use a program such as Migrate-n[6] to estimate migration rates, but the nature of the empirical data set may make this problematic. The model implemented in STEM assumes that shared polymorphism results from incompletely sorted ancestral polymorphism, this same shared polymorphism may be interpreted by Migrate-n (or other coalescent-based approaches to estimating migration) as resulting from migration. One solution to this dilemma would be to conduct demographic model selection using information theory[7]. In any case, the type of gene flow is likely to make a difference: if speciation proceeds via a process of genetic isolation with migration between sister taxa, STEM will probably produce relatively good estimates of the

species tree, but if gene flow occurs among non-sister taxa (e.g., an n-island model), then the species tree estimates will not be accurate[8].



Theta. For a set of closely-related species or subspecies, is it reasonable to assume that θ is constant? One approach to answering this question is to compare the point estimates of this parameter across OTUs. I have done this for one empirical data set consisting of ten OTUs and 4-8 individuals per OTU, and the results are presented to the left. My interpretation is that the per locus θ (estimated using Migrate-n) does not differ greatly across OTUs. The estimates are similar, particularly given that the sampling was not equal across OTUs, but more importantly the point estimates of each OTU are contained within the 95% confidence interval of the other estimates.

STEM requires that users provide a per-site estimate of θ, so the per-locus values output from a program such as Migrate-n can be converted by dividing this value by the average length of each locus. This value is primarily used to convert the tree length to coalescent units and to calculate the –ln L of the species tree.

---

[1] Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Systematic Biology, 52:674-683.

[2] Swofford, D. L. 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

[3] Zwickl, D. J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.

[4] Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

[5] Rambaut, A. http://tree.bio.ed.ac.uk/software/treeedit/

[6] Beerli, P. and J. Felsenstein. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations using a coalescent approach. PNAS USA 98(8):4563-4568.

[7] Carstens BC, Stoute HN, Reid NM (2009) An information theoretical approach to phylogeography. Molecular Ecology, 18, 4270-4282

[8] Eckert AJ, Carstens BC. (2008) Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Molecular Phylogenetics & Evolution 49, 832-842.