Theory

For the purpose of predicting phylogenetic utility, Townsend (2007) approximates complex trees as equivalent to a simple quartet with full-length branches. However, as the number of taxa increases, this approximation becomes increasingly inaccurate (Townsend and Leuenberger 2011). Although the approximation holds well if taxa subtending the quartet all are recently divergent lineages, the presence of taxa diverging near the node of interest can impact the power of phylogenetic inference regarding the proper resolution of the quartet (Townsend and Lopez-Giraldez 2010). To achieve a comprehensive approach toward phylogenetic experimental design, calculation of probabilities of resolution would be performed on any set of character rates for any taxon-sampling scheme. Our aim here is to reduce a tree of arbitrarily complex topology with known branch lengths to a simpler quartet in such a way that the expected information concerning interior nodes as inferred from the states at the leaves will be changed as little as possible. In this approach, each step of the algorithm replaces a fork of two neighboring external branches with a single branch (Figure 1A). The branch length in the collapsed tree after each step is selected such that the expected information concerning interior nodes inferred from the character states of the leaves is preserved. This step is iterated on all forks subtending a node of interest (Figure 1A). The resulting quartet tree can be analyzed as equivalent to the original tree (Figure 1B), and can serve as a basis for calculation of the potential phylogenetic utility of nucleotide sequences via existing theory of phylogenetic signal and noise (Townsend et al. 2012, Su et al. 2014).

Following Townsend Townsend (2007), we assume specification of the rates of evolution, , of characters to be used for phylogenetic inference. These rates can be calculated using a guide tree and a corresponding character matrix to infer a rate of character change, , for each site (Lopez-Giraldez and Townsend 2011). While site rates can be quantified using either the alignment of interest and a tree or set trees inferred using this corresponding alignment (Lopez-Giraldez and Townsend 2011, Dornburg et al. 2014), alignments of homologous sequence from sister clades or phylogenetically deeper alignments comprising data from public sources with corresponding topologies can also be used to infer site-specific rates (Townsend 2007, Dornburg et al. in review). In the context of predictive utility of future sequencing, the character states for tip taxa are therefore unknowns represented by character state probability distributions, rather than observed values. Thus, for each character, given the value of , we can collapse two neighboring external branches of node *M* with lengths and that lead to character state probability distributions and , respectively (Fig. 1C). We wish to find a branch of length *T’* such that the amount of information provided by a single character state probability distribution *C’* about *M* is equivalent to the amount of information that *C*1 and *C*2 provide about *M*. More specifically, we wish to find a *T’* that satisfies the following equality



, (1)



where is the information in the context of a single *T*’, given the possible states of *M* as a function of *C*’, and is the information in the context of *T1* and *T2*, given the possible states of M as a function of *C*1 and *C*2.



To solve this equality, we can use the mutual information of two discrete random variables, .  measures the amount of information *X* contains about *Y*, and is defined as



, (2)



which quantifies the reduction in uncertainty (the entropy H=SUM*\_x*(ln *px*); Cover and Thomas 2012) about *Y* due to knowledge of the state of *X*.

Applying Equation 2 to Equation 1 produces

. (3)



To calculate each side of Equation 3, we can use the conditional entropy of one discrete random variable given another, , defined as



, (4)



where Pr{*x*} represents the probability of discrete state *x*. Applying Equation 4 first to the right hand side of Equation 3, and applying the definition of entropy, yields

. (5)



The expected value E on the right side of Equation 5 can then be computed by summing the definition of entropy over *s* character states for *C*1,*C*2,and *M*.

. (6)

Probabilities from a Markov transition matrix representing a model of molecular evolution (*e.g.* the **Q** matrix with transition probabilities *pij* and base frequencies *πA, πG*, *πC*, *πT*; Sullivan and Joyce 2005), can provide the probabilities to parameterize Equation 6. The joint probability of *C*1,*C*2,and *M* can then be expressed as

, (7)



where the branch length *b* traditionally specified in **Q** is equal to , where ** is typically calculated by maximum likelihood as in Lopez-Giraldez and Townsend (2011) and Pond and Muse (2005), and where *T* (representing *T*1 or *T*2) derives from the branch lengths of the specified hypothetical taxon sampling tree or from the branch length of a specified hypothetical tree distribution. Furthermore, the joint probability of *C*1 and *C*2 can be expressed by summing over all possible character states of *M*:



. (8)



Provided Equations 7 and 8, we can then compute the conditional probability of *M* given *C*1 and *C*2 as

. (9)

Equations 7–9 provide the probabilities necessary for calculation of Equation 5, the conditional Shannon entropy of *M* given *C*1 and *C*2:

. (10)

Equation 10 provides the means of calculating .



Now working on the left hand side of Equation 3, application of Equation 4 yields the conditional Shannon entropy of *M* given *C’*

. (11)

Just as we summed over all possible character states using Equation 6 to compute the value of the right hand side of Equation 5, we can sum over all possible character states for *C’* and *M* to solve Equation 11:

. (12)

We can rewrite the right hand side of equation 13, deriving in terms of the Markov transition probabilities, following the same approach as in equations 7 and 8, so that



. (13)

Equations 10 and 13 thus provide the left and right sides of Equation 3 derived in terms of known parameters and a single free parameter *T’*. Thus, we can perform the information collapse of the specified bifurcation by solving

 (14)

numerically for the free parameter *T’*.

Equation 14 can be applied serially, collapsing dichotomous terminal branches into a single branch with a length that corresponds to the equivalent information content regarding the ancestral node *M* (Figure 1C). This serial collapse can be performed in any order to result in a quartet around the node of interest. Parameters describing this quartet are the appropriate parameters for calculations of signal and noise (Su et al. 2014) pertaining to the node of interest in the complete tree.

Although we derive equations for trees with only dichotomous branching patterns, these equations can be extended to incorporate nodes with *n* subtending terminal branches as

. (15)



Note that the **Q** matrix supplies the transition probabilities *pij* and base frequencies *πA, πG*, *πC*, *πT*; for equation 15, and therefore enables application of the commonly used models of nucleotide substitution (*e.g.* JC, HKY, *etc*.; Su et al. (2014)) to the information collapse. Furthermore there is no necessary reason why the model needs to be specified identically for more than the collapse of one set of subtending branches. Model settings can easily be altered to reflect any prior knowledge of clade specific patterns of molecular evolution (Soltis et al. 2002, Dornburg et al. 2012, Romiguier et al. 2013, Cox et al. 2014, Li et al. 2014).

From Cover and Thomas Equations (2.60) and (2.92), it follows that

, (16)



Since monotonically increases in starting at , we conclude that there will always be a solution for on the interval.



References

Cover, T. M., and J. A. Thomas. 2012. Elements of information theory. John Wiley & Sons.

Cox, C. J., B. Li, P. G. Foster, T. M. Embley, and P. Civáň. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. Systematic Biology **63**:272-279.

Dornburg, A., M. C. Brandley, M. R. McGowen, and T. J. Near. 2012. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Molecular Biology and Evolution **29**:721-736.

Dornburg, A., A. H. Moeller, D. Greaves, F. Lopez-Giraldez, T. J. Near, E. J. Sargis, and J. P. Townsend. in review. PhyNOIR: An algorithmic approach to phylogenetic noise isolation and reduction based on phylogenetic informativeness. BMC Evolutionary Biology.

Dornburg, A., J. P. Townsend, M. Friedman, and T. J. Near. 2014. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. BMC Evolutionary Biology **14**:169.

Li, B., J. S. Lopes, P. G. Foster, T. M. Embley, and C. J. Cox. 2014. Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. Molecular Biology and Evolution **31**:1697-1709.

Lopez-Giraldez, F., and J. P. Townsend. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evolutionary Biology **11**.

Pond, S. L. K., and S. V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. Pages 125-181 Statistical methods in molecular evolution. Springer.

Romiguier, J., V. Ranwez, F. Delsuc, N. Galtier, and E. J. Douzery. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Molecular Biology and Evolution:mst116.

Soltis, P. S., D. E. Soltis, V. Savolainen, P. R. Crane, and T. G. Barraclough. 2002. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. Proceedings of the National Academy of Sciences of the United States of America **99**:4430-4435.

Su, Z., Z. Wang, F. Lopez-Giraldez, and J. P. Townsend. 2014. The impact of incorporating molecular evolutionary model into predictions of phylogenetic signal and noise. Phylogenetics, Phylogenomics, and Systematics **2**:11.

Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. Annual Review of Ecology Evolution and Systematics **36**:445-466.

Townsend, J. P. 2007. Profiling phylogenetic informativeness. Systematic Biology **56**:222-231.

Townsend, J. P., and C. Leuenberger. 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Systematic Biology **60**:358-365.

Townsend, J. P., and F. Lopez-Giraldez. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Systematic Biology **59**:446-457.

Townsend, J. P., Z. Su, and Y. I. Tekle. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Systematic Biology **61**:835-849.

**Figure 1**: A visual overview of the Information Collapse method. (a) the subtree τ1 is collapsed into a single branch by recursively collapsing neighboring external branches. (b) A tree of arbitrarily complex topology, viewed as four subtrees τ1 – τ4 centered about an internode of interest of length *t0*, is collapsed into a quartet with four unequal branch lengths *T*1 *– T*4. (c) Neighboring external branches with leaves *C*1 and *C*2 of lengths *T*1 and *T*2, respectively, are reduced to a single branch with leaf *C’* of length *T’*, where *T’* is chosen such that the mutual information between *C’* and *M* is as close as possible to the mutual information between *M* and *C*1, *C*2.

Graveyard of text snippets

Although the quartet is the fundamental unit of the tree graph (Bandelt and Dress 1986),

In this paper, we have extended the Townsend et al. (2012) phylogenetic signal and noise analysis by extending the formal domain of the model from only quartets to trees of arbitrarily complex topology.