AFD\_Draft1

Russell Dinnage

Wednesday, April 09, 2014

# Introduction

Ecologists have increasingly recognized the need to incorporate genetic information into community analysis methods (ref). The important goals of understanding the different types of diversity, how diversity is generated and maintained, and the consequences of diversity can all be greatly aided when we stop treating individual species as interchangeable entities and instead treat them as non-independent points along an evolutionary continuum. In order to incorporate the genetic relatedness of species, and thus account for their shared evolutionary history, community ecologists have developed new diversity metrics which incorporate the information contained in molecular phylogenies to weight the contribution of species to the diversity of a community based on how much novel or distinctive evolutionary history they contribute (ref). However, this requires the construction of a phylogeny, which is a non-trivial task for ecologists when there is no ready-made phylogeny available for a group of interest. This problem will only get worse as we enter the age of cheap high throughput sequencing technology. With the advent of metagenomics, ecologists will increasingly be faced with the boon of enormous amounts of genetic information, and the curse of having to find a way to efficiently analyse it. Utilizing such large amounts of data to understand diversity will be a major challenge in particular for community phylogenetics, as building a phylogeny becomes more complicated and extremely computationally intensive for large genetic datasets.

One approach to dealing with this explosion of data is to simply filter out most of it and just focus on a few genes, which is largely the approach which has been used to date (refs). The purpose of this study is to investigate ways to utilize all the available data, while greatly reducing the amount of time and computational effort required to achieve results of comparable quality, and perhaps even higher quality, than those of traditional phylogenetic methods.

Phylogenies often contain more information than is required by ecologists to answer their questions of interest. Phylogenies are usually treated as a proxy for what ecologists are really interested in: traits (ref). Species traits – or phenotype – controls how species interact with their environment and other species, and so they drive the ecology of communities (ref). Phylogeny is a proxy for trait divergence because traits are the product of evolutionary history and so – all else being equal – closely related species will share similar traits on average more than distantly related ones, because they have had less time to diverge (ref). Phylogenies contain two main pieces of information: branch lengths – the amount of evolutionary time or genetic divergence between species – and branching order (or topology) – the order in which species have diverged, which determines things like sister species, and monophyly. The development of phylogenetic methods has largely been concerned with achieving high accuracy in the second of these two pieces of information, the branching order or topology of the generated phylogenetic trees, because this is the information most relevant for systematists, who have developed the methods. We would argue, however, that it is the first piece of information contained in phylogenies, the branch lengths, that are of most interest to ecologists. Afterall, the branching order hardly matters if we are mainly interested in the average degree of divergence among species, which is what we expect to be correlated with the amount of trait divergence. Therefore, we propose that ecologists can dispense with time-consuming steps designed to increase topological accuracy in phylogenetic reconstruction, and in many cases, work with raw genetic data to generate diversity metrics of similar utility to phylogenetic diversity metrics.

## Alignment-free Genomic diversity Metrics

One of the most time-consuming and difficult steps in phylogenetic reconstruction is the alignment of multiple sequences together. The purpose of alignment is to estimate homology – that is, which regions of the sequence should be compared with which regions in other sequences (ref). Establishing homology is important to estimate accurate tree topologies, and is also important for estimating branch lengths in traditional phylogenetic methods, because with insertions and deletions in sequences, unaligned sequences may contain evolutionarily unrelated nucleotides at the same position in different sequences. Alignment is computationally expensive because it usually proceeds by aligning all species to all other species, and combining these pair-wise alignments. This means that for N species, N(N-1)/2 alignments need to be made.

Alignment is also difficult for automated algorithms to do, especially in genomic sequences and highly diverged sequences where genome rearrangements make determining homology very challenging (refs). Also, in large genetic datasets, alignments cannot be checked by eye because they are too large. The accuracy of alignments is also affected taxon sampling. Generally, distantly related species are more accurately aligned if the dataset includes other species which are evolutionarily intermediate between them. This can lead to different phylogenetic estimates depending on which species are included in the analysis, which is particularly a problem for community ecologists, because communities are usually a highly incomplete subset of represented taxonomic groups.

In response to the challenge of aligning large genomic datasets, a number of alternative “alignment-free” methods of phylogenetic reconstruction have been proposed (refs). Most of these methods proceed by generating pair-wise genetic distances between species using rapid methods that do not need to establish homology first. Trees are then constructed using hierarchical clustering methods. Despite the very different philosophy of constructing these trees – harkening to the days of ‘phenetics’ (refs) – most work has focused on achieving comparable topological accuracy to traditional alignment-based methods (alignment-free methods so far are less topologically accurate: refs).

For ecologists, alignment-free methods offer an opportunity to streamline the way we measure phylogenetic diversity. Many phylogenetic diversity metrics use pair-wise distances which are derived from a phylogeny. For these metrics, ecologists would no longer even have to do the tree-building step, we can use the alignment-free genetic distances directly. For those methods that do need the tree, ecologists can generate the tree with hierarchical clustering, which is extremely fast.

The goal of this study is to determine the alignment-free methods, and their particular tuning parameters, which maximize the correspondence of phylogenetic diversity metrics calculated using them and the ‘true’ phylogenetic diversity, by finding the best possible correlation between alignment-free genetic distance and the branch lengths of an underlying phylogeny used to simulate genomics sequences. We will also compare the performance of these alignment-free methods with those that use alignments. Lastly, we will use a real genomic dataset to see how well these same alignment-free genetic distances correlate with the ecological trait-based distances between a set of aquatic insect taxa.

# Methods

## Simulated Datasets

## *Simulated genomes from simulated phylogenies*

In order to test the accuracy of different alignment-free methods, we first simulated genomic data using known, underlying phylogenies. We based our analysis on a set of 156 randomly generated phylogenies. Each phylogeny was simulated under a birth-death process using the Artificial Life Framework (ALF) software (ref). All trees used the same birth rate (0.04) and death rate (0.025) parameters, and were run until each phylogeny had 300 terminal species. We used three different divergence rates (52 phylogenies for each) such that the maximum divergence achieved in the phylogeny was high -- 800 PAM units (ref), medium -- 600 PAM units, or low -- 300 PAM units. (explain what a PAM unit is here)(ref)

The ALF software simulates the evolution of genomic sequences along these input phylogenies using various parameters to customize the results. We used a standard set of parameters for all phylogenies for consistently, with the goal of a realistically complex evolutionary model. We simulated 20 independendent genes across each phylogeny to form our 'genome'. The overall sequence substitution model we used was the 'CPAM' model, which simulates sequences evolution by using empirical codon substitution rates. Insertion and deletions happened at a base rate of 0.0001 per substitution with their length determined by a Zipfian (ref) distribution with parameter 1.821 (the default), with a maximum length of 50. Substitution rate variation between different genes was incorporated using 5 randomly assigned substitution classes with a gamma distribution with alpha and beta parameter equal to 1, and a proportion of invariable sites equal to 0.01. Rate heterogeniety among the 20 genes was also allowed and was distributed according to a gamma distribution with alpha and beta parameter of 1. Gene length was distributed according to a gamma distribution (with alternate parameterization) with parameter k set to 2.4019, and theta set to 133.8063 (ref).

## *Simulated communities*

We used the 152 simulated phylogenies and their accompanying genomes to create two different sets of simulated communities, to test different aspects of the alignment-free methods.

* Community Set 1: This set consists of 4000 randomly generated communities, containing from 10-50 species. This set is designed to compare alignment-based phylogenetic distances and topology to alignment-free methods when there is strong taxon sampling. Because alignment-based methods' required computational resources increase rapidly with the number of species, this set is the only one where we could achieve a high replication rate within a reasonable timescale. Each community consisted of 10-50 (uniformly drawn) randomly selected species from a randomly chosen base phylogeny. For each species in each community, a unique genome was assigned by selecting from 2-12 (uniformly drawn) genes out of the 20 available from the sequence simulations described above.
* Community Set 2: This set consists of 30 communities which vary from 30 to all 300 species, in intervals of 30, sampled from only three of the original phylogenies, one from each of the divergence categories. This smaller set was used to test how taxon sampling affects each of the different methods of estimating genetic distance and phylogenetic topology. Additionally, we used this set to get an estimate of the time it takes for each of the methods to achieve their results, and how this scales with the number of species. Each species in each community had the same randomly chosen set of 3 genes in this set. (Note: I want to redo this analysis for two more reps, with 10 genes and 20 genes, to see how time scales with genome length as well, but this will take awhile to run) Each community in these sets will be considered a regional species pool. Later, we also generate sub-communities which represent local communities so that we can test the ability of the different phylogenetic methods to accurately reflect phylogenetic structure based on the underlying tree.

## Estimating genetic distances and phylogenies

For each community we generated a matrix of genetic distances between all species, and created phylogenetic trees, using several different methods.

### \* Alignment-based Method\*

For each community we aligned each of the 2-10 simulated genes seperately using Muscle software (ref) with the default settings. We then concatenated these aligned genes together to create a genetic supermatrix. We chose Muscle because it is used frequently in the literature, and because it is one of the fastest alignment methods, and so is likely to be used in situations where there is a lot of data. We then used this supermatrix to create a maximum likelihood phylogeny using the software RaxML (ref). Again, we chose this software because it is the fastest ML phylogeny program available, and is the only one that can feasibly create phylogenies from massive datasets.

Genetic distances were drawn from this phylogeny by generated all pairwise distance by measuring the branchlengths seperating two species on the tree. We generated two sets of distances, one with the raw phylogeny from the ML analysis, the other from the same phylogeny, only with its branches rescaled to form an ultrametric tree using penalized likelihood (lambda=1) (ref: Sanderson) in the 'chronopl' function in R package 'ape'(ref).

### *Alignment-free Methods*

For each community, we also generated genetic distance estimates using methods which do not require alignment, and then created phylogenies from these genetic distances using heirarchical clustering (UPGMA: ref), using the 'fastcluster' package in R (ref).

## Comparing estimated genetic distances and topology to the truth

## Diversity Metrics

## Comparing ecological distance to genetic distances

## *Dataset*

## *Statistical Analysis*

# Results