Methods for Choosing Clusters in Phylogenetic Trees

Tom Burr
Los Alamos National Laboratory, USA

INTRODUCTION

One data mining activity is cluster analysis, of which there are several types. One type deserving special attention is clustering that arises due to evolutionary relationships among organisms. Genetic data is often used to infer evolutionary relations among a collection of species, viruses, bacterial, or other taxonomic units (taxa). A phylogenetic tree (Figure 1, top) is a visual representation of either the true or the estimated branching order of the taxa, depending on the context. Because the taxa often cluster in agreement with auxiliary information, such as geographic or temporal isolation, a common activity associated with tree estimation is to infer the number of clusters and cluster memberships, which is also a common goal in most applications of cluster analysis. However, tree estimation is unique because of the types of data used and the use of probabilistic evolutionary models which lead to computationally demanding optimization problems. Furthermore, novel methods to choose the number of clusters and cluster memberships have been developed and will be described here. The methods include a unique application of model-based clustering, a maximum likelihood plus bootstrap method, and a Bayesian method based on obtaining samples from the posterior probability distribution on the space of possible branching orders.

BACKGROUND

Tree estimation is frequently applied to genetic data of various types; we focus here on applications involving DNA data, such as that from HIV. Trees are intended to convey information about the genealogy of such viruses and the most genetically similar viruses are most likely to be most related. However, because the evolutionary process includes random effects, there is no guarantee that "closer in genetic distance" implies "closer in time," for every pair of sequences.

Sometimes the cluster analysis must be applied to large numbers of taxa, or applied repeatedly to the same number of taxa. For example, Burr, Myers, and Hyman (2001) recently investigated how many subtypes (clusters) arise under a simple model of how the *env* (gp120) region of HIV-1, group M sequences (Figure 1, top) are evolving. One question was whether the subtypes of group M could be explained by the past population dynamics of the virus. For each of many simulated data sets each having approximately 100 taxa, model-based clustering was applied to automate the process of choosing the number of clusters. The novel application of model-based clustering and its potential for scaling to large numbers of taxa will be described, along with the two methods mentioned in the introduction section.

It is well-known that cluster analysis results can depend strongly on the metric. There are at least three unique metric-related features of DNA data. First, the DNA data is categorical. Second, a favorable trend in phylogenetic analysis of DNA data is to choose the evolutionary model using goodness of fit or likelihood ratio tests (Huelsenbeck & Rannala, 1997). For nearly all of the currently used evolutionary models, there is an associated distance measure. Therefore, there is the potential to make an objective metric choice. Third, the evolutionary model is likely to depend on the region of the genome. DNA regions that code for amino acids are more constrained over time due to selective pressure and therefore are expected to have a smaller rate of change than non-coding sequences.

A common evolutionary model is as follows (readers who are uninterested in the mathematical detail should skip this paragraph). Consider a pair of taxa denoted x and y. Define F_{xy} as

$$NF_{xy} = \begin{pmatrix} n_{AA}n_{AC}n_{AG}n_{AT} \\ n_{CA}n_{CC}n_{CG}n_{CT} \\ n_{GA}n_{GC}n_{GG}n_{GT} \\ n_{TA}n_{TC}n_{TG}n_{TT} \end{pmatrix},$$

where N is the number of base pairs (sites) in set of aligned sequences, n_{AA} is the number of sites with taxa x and y both having an A, n_{AC} is the number of sites with taxa x having an A and taxa y having a C, etc. The most general time-reversible model (GTR) for which a distance measure has been defined (Swofford, Olsen, Waddell, & Hillis, 1996) defines the distance between taxa x and y as d_{xy} = -trace$\{\Pi \log(\Pi^{-1}F_{xy})\}$ where Π is a diagonal matrix of the average base frequencies in taxa x and y and the trace is

the sum of diagonal elements. The GTR is fully specified by 5 relative rate parameters (*a, b, c, d, e*) and 3 relative frequency parameters (π_A, π_C, and π_G with π_T determined via $\pi_A + \pi_C + \pi_G + \pi_T = 1$) in the rate matrix Q defined as

$$Q/\mu = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ a\pi_A & d\pi_C & - & f\pi_G \\ a\pi_A & e\pi_C & f\pi_G & - \end{pmatrix},$$

where μ is the overall substitution rate. The rate matrix Q is related to the substitution probability matrix P via $P_{ij}(t) = e^{Qt}$, where $P_{ij}(t)$ is the probability of a change from nucleotide *i* to *j* in time t and $P_{ij}(t)$ satisfies the time reversibility and stationarity criteria: $\pi_i P_{ij} = \pi_j P_{ji}$. Commonly used models such as Jukes-Cantor (Swofford et al. 1996) assumes that $a = b = c = d = e = 1$ and $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$. For the Jukes-Cantor model, it follows that $P_{ij}(t) = 0.25 + 0.75e^{-\mu t}$ and that the distance between taxa *x* and *y* is $-3/4 \log(1 - 4/3D)$ where *D* is the percentage of sites where *x* and *y* differ (regardless of what kind of difference because all relative substitution rates and base frequencies are assumed to be equal). Important generalizations include allowing unequal relative frequencies and/or rate parameters), and to allow the rate μ to vary across DNA sites. Allowing μ to vary across sites via a gamma-distributed rate parameter is one way to model the fact that sites often have different observed rates. If the rate μ is assumed to follow a gamma distribution with shape parameter γ then these "gamma distances" can be obtained from the original distances by replacing the function $\log(x)$ with $\gamma(1-x^{-1/\gamma})$ in the d_{xy} = -trace{P $\log(P^{-1}F_{xy})$} formula (Swofford et al. 1996). Generally, this rate heterogeneity and the fact that multiple substitutions at the same site tend to saturate any distance measure make it a practical challenge to find a metric such that the distance between any two taxa increases linearly with time.

Figure 1. HIV Data (env region). (Top) Hierarchical Clustering; (Middle) Principle Coordinate plot; (Bottom) Results of model-based clustering under six different assumptions regarding volume (V), shape (S), and orientation (O). E denotes "equal" among clusters and "V" denotes "varying" among clusters, for V, S, and O respectively. For example, case 6 has varying V, equal S, and varying O among clusters. Models 1 and 2 each assume a spherical shape (I denotes the identify matrix, so S and O are equal among clusters, while V is equal for case 1 and varying for case 2). Note that the B and D subtypes tend to be merged.