Genetic Subtyping using Cluster Analysis

Tom Burr
Los Alamos National Laboratory
Mail Stop E541
Los Alamos, NM 87545
tburr@lanl.gov

James R. Gattiker
Los Alamos National Laboratory
Mail Stop E541
Los Alamos, NM 87545
gatt@lanl.gov

Greggory S. LaBerge
Denver Police Dept. Crime Lab
1331 Cherokee Street Room 648
Denver, CO 80204
gslaberge@qwest.net

ABSTRACT

In this paper we (1) describe state-of-the-art methods to identify clusters in DNA sequence data for taxonomic analysis; (2) describe a new method with better scaling properties based on model-based clustering, and (3) present examples using the nucleoprotein and hemagglutin regions of influenza and the *env* and *gag* regions of human immunodeficiency virus (HIV).

Keywords

Model-based clustering, phylogenetic trees, DNA sequence analysis, HIV, influenza.

1. INTRODUCTION and BACKGROUND

The effort to detect, describe, and explain biological diversity is known as systematics [33]. Taxonomy falls within the systematics discipline and concentrates on identifying groups such as species (which are reproductively isolated) or subspecies. In the cases we consider here, groups are defined via evolutionary relationships. Evolutionary trees display inferred evolutionary relationships and often exhibit clusters of distinct groups. These groups could for example be species, subspecies, or in our applications here, are subtypes of virus. A common way for such groups to arise is geographic and/or temporal isolation in conjunction with evolutionary processes. Phylogenesis is defined as evolution within a species together with speciation processes. Strictly speaking, phylogenetic trees (such as in Fig. 1a) should include speciation events, but we will follow common usage [33] and refer to any tree that depicts evolutionary relationships as a phylogenetic tree. For example, Fig. 1a is a phylogenetic tree of influenza strains that have evolved within three distinct hosts (human, swine, and avian).

Many genetic data types are available for taxonomic analysis. For example, restriction fragment length data is available when deoxyribonucleic acid (DNA) is digested with restriction enzymes and DNA-DNA hybridization data is available as temperatures at which the double-stranded DNA separates. DNA sequence data consisting of A, C, T, or G base pairs (or the associated amino acid sequence data) is increasingly available and is generally believed to provide the most information about evolutionary histories. Here, we consider only DNA sequence data and our examples involve identifying subtypes of viruses (influenza or HIV). However, we emphasize that the clustering techniques we present are more widely applicable.

Identifying and characterizing viral subtypes is useful for theoretical and applied purposes. Theoretical purposes include (1) inferring aspects of the evolutionary history that led to the subtypes; (2) evaluating the molecular basis for differences in subtype behavior, including transmissibility; and (3) predicting future trends. In category (1), [3] evaluates how many subtypes can emerge under several models of HIV evolution and whether the subtypes of HIV could be explained by the known past population dynamics of the virus. In category (2), for HIV, a relationship between subtype and natural resistance against antiretroviral drugs as well as between subtypes and the efficiency of serological and molecular tests for HIV have been observed [17]. The degree to which vaccines based on one subtype will provide protection against other subtypes is not yet well understood. Regarding (3), it is believed that current temporal and geographic patterns in the distribution of genetic subtypes can help to forecast growth rates of the overall epidemic. Applied purposes include identifying subtypes as a simple yet effective summary of the phylogenetic tree and identifying simple signature patterns [5, 22, 27] that distinguish each subtype so that rapid screening studies of the prevalence of each subtype need not build computationally intensive phylogenetic trees each time new HIV samples are available, but instead, the subtype of a sample can be identified by comparing it to the signature of each subtype.

Although cluster analysis is used for many purposes in the analysis of DNA sequence data, our focus is identifying groups for taxonomic analysis. Our goals are to define genetic subtypes, and to associate genetic variation or subtypes with geographic, temporal, or species information. In this paper we (1) describe methods that are in current use to identify clusters in DNA sequence data; (2) introduce a new method based on model-based clustering that has the potential to accommodate larger numbers of sequences, and (3) present examples using simulated and real DNA data. The paper includes a survey of several current phylogenetic tree-based clustering methods so that the new method can be compared to existing methods. Section 2 provides examples using influenza data. Section 3 describes how to define genetic distances based on an evolutionary model, with more detail given in the Appendix. Section 4 describes existing cluster analysis methods with bootstrap resampling to assess confidence in the chosen number of clusters. Section 5 introduces our new application of model-based clustering applied to a multidimensional scaling of the genetic distance data. Section 6 presents an example of model-based clustering using the nucleoprotein (NP) and hemagglutinin (HA) regions of influenza virus and another example using the *env* and *gag* regions of HIV-1, group M (all sequence data is available [36]). Section 7 discusses scalability issues. Section 8 is a summary and indicates directions for future research.

2. EXAMPLES

Cluster analysis is frequently applied to DNA data. For example, Figure 1a is an unrooted phylogenetic tree of influenza (NP region) sequences from 3 host species (human, swine, avian) constructed by a computationally demanding maximum likelihood (ML) algorithm that is commonly used for data sets with small

numbers (less than 100) of sequences [9, 10]. Although it is usually assumed that the sequences evolved from a common ancestor, the location of the ancestor in the tree can be left unspecified, in which case the tree is unrooted and the direction of increasing time is left unspecified. Sometimes a distant taxa sequence can be used to locate the tree's root (the position of the most recent common ancestor, MRCA), in which case time increases as the tree is traversed outward from the root. For example, Figure 1b is a rooted tree of a subset of the sequences used in Figure 1a. These trees are intended to convey information about the genealogy of these influenza viruses. Because the evolutionary process includes random events, there is no guarantee that "closer in genetic distance" implies "closer in

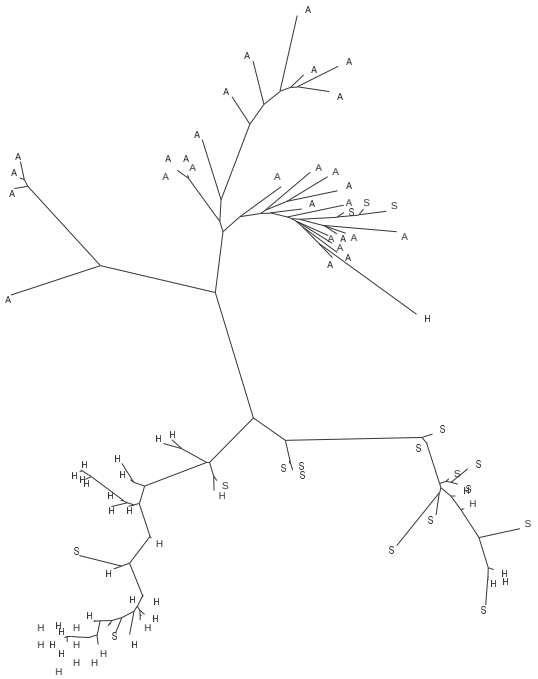

Fig. 1a. Unrooted tree of 85 sequences of the nucleoprotein region of the influenza virus. Several "wrong host" sequences are the result of cross-species transmissions. Generally, the Swine (S) and Human (H) "clusters" are closer to each other than to Avian (A).

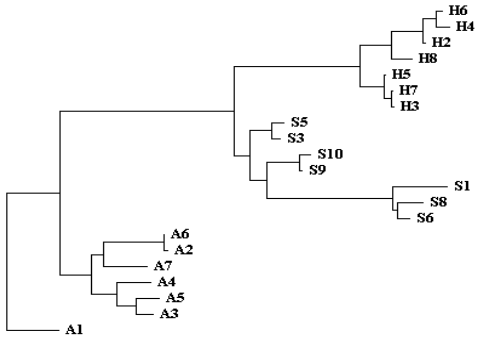

Fig. 1b. Rooted tree of a subset (21) of the 85 sequences from Fig. 1a with 7 sequences from each of A, H, and S hosts.

time," for every pair of influenza sequences. For example, from Figure 2 we note that some 1996 human-host influenza (HA region) sequences are more similar to some 1995 sequences than they are to other 1996 sequences. However, the tree in Figure 2 is rooted using a 1968 sequence and a pattern is evident in this rooted tree: generally, there is a drift away from the 1968 sequence.

3. GENETIC DISTANCES AND EVOLUTIONARY MODELS

It is well known that cluster analysis results can depend strongly on the metric. There are at least three notable metric-related features of DNA data. First, the DNA data is categorical. Second, modern phylogenetic analysis of DNA data includes choosing the evolutionary model using goodness of fit or likelihood ratio tests [18]. 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. Coding regions (genes) are often more constrained due to selective pressure and therefore are expected to be more conserved over time (have a smaller rate of change) than non-coding regions.

Our data is assumed to be n mutually aligned DNA sequences of (typically) hundreds to a few thousand sites. Each site is either an A, C, T, or G representing one of the 4 DNA bases. As an important aside, evolutionary processes sometimes lead to base insertions or deletions (known as indels) so that not all taxa have the same number of base sites. Alignment is a crucial step [33] in which gaps are inserted where DNA indel evolutionary events are inferred. One method for aligning two or more sequences was developed by Needleman and Wunsch [28] that applies a dynamic programming method to a matrix plot in which every nucleotide in one sequence is compared to every nucleotide in all other sequences to find the alignment with the fewest gaps and substitutions. Alignment introduces a possible error source (that is typically ignored) which we do not consider. Distance measures between taxa usually consider only those aligned sites, assumed to have arisen from a common ancestor, for which all taxa have a base pair rather than an alignment character. That is, sites having one or more indels are nearly always removed from the analysis, as we do here.

The most basic model of the data generating process assumes that all mutation events (A to C, A to T, etc.) are equally likely. More realistic models weight the event probabilities by their inferred frequency of occurrence. Because of "convergent evolution" all distance measures eventually saturate at approximately 25% mismatch between any two sequences. An example of convergent evolution is at time t_1 sequence 1 having an A at site i and sequence 2 having a C at site i, but at a later time t_2, both sequences having an A at site i. An evolutionary model should specify the probability per unit time of a substitution from A to C, A to G, etc. in a 4-by-4 substitution probability matrix P. Readers who are interested in the details are referred to the Appendix. Results presented here use either the HKY5 or GTR models, which are described in the Appendix.

Because of the trend toward using the most realistic feasible evolutionary model and the fact that most models have an associated distance measure, distance-based clustering methods

are natural candidates for taxonomic analysis. Another candidate involves using maximum likelihood plus a resampling scheme such as the bootstrap as we describe next.

Fig. 2. Rooted tree of the hemagglutinin region of the influenza virus. The isolation year is the first two digits and with a sequence identification given in the last digit.

4. CLUSTERING METHODS

Any of the typical clustering methods could be applied to DNA data, but some are more suited than others, particularly because of the categorical nature (A, C, T, or G) of the data. We will describe maximum likelihood, hierarchical, kmeans, and model-based clustering in this section. We emphasize throughout that an evolutionary model chosen from goodness of fit or likelihood ratio methods should provide the most defensible basis for selecting the model. Candidate clustering methods should incorporate the chosen model.

4.1 Maximum Likelihood Plus Bootstrap

The substitution probability matrix P leads to a likelihood [9] for each candidate topology (branching order or genealogy of the sample sequences). For example, at site i with an assumed ancestor site of C, the probability (as a function of branch length) of obtaining the observed values for each taxa can be evaluated and therefore can be maximized for a given topology with respect to branch lengths. The topology with the highest maximum (over branch lengths) among all candidate topologies is the maximum likelihood topology for that site. Assuming that sites evolve independently, this procedure can be combined over all sites to find a global maximum. The "molecular clock" hypothesis states that the evolutionary rate μ does not vary across sequences (branches in the tree). But this hypothesis is sometimes rejected in likelihood ratio tests in which case likelihood-based methods produce tree estimates with unequal branch lengths as measured from a root position. Regardless of whether the clock assumption is imposed, software is available (see [33] for a partial listing and PAUP [34] is one example) to find a local maximum likelihood branching order. The number of possible rooted topologies N is approximately factorial in the number of sequences n ($N = \prod_{k=1}^{n-1} 2k-1$), and the number of unrooted topologies is the same but with $n-1$ replacing n. Therefore, for approximately 15 or more sequences, it is not feasible to find the global maximum likelihood topology. Instead, heuristic search methods to find a good tree are used. More recently, stochastic searches [30] and Bayesian posterior probability (via Markov Chain Monte Carlo) [25,31] approaches have been implemented. All of these methods have merit but are computationally intensive and do not lead to a fully automated way to choose the number of clusters. Usually, the human eye is used to identify clusters in the phylogenetic tree and a resampling scheme such as the bootstrap [7] is used to assess confidence in the chosen clusters. The typical implementation of the bootstrap is to resample the sites (columns of DNA data) with replacement to generate bootstrap samples which are treated the same as the original data. Then, for any cluster that is specified by the user, the bootstrap implementation will report how many bootstrap samples maintained that prespecified cluster as being "monophyletic." Monophyletic means that all cluster members clustered before any non-cluster member. We gave an example with 3 clusters (S, H, and A hosts) in Figure 1b. The number (out of 100) of bootstrap samples that supported the stated clusters was 100, and 1000 of 1000 trees sampled from the posterior distribution (estimated via BAMBE [31]) supported the expected clusters. Similarly, when 8 of the 1993 and 8 of the 1996 influenza sequences (hemagglutinin region) were selected from the sequences shown in Figure 2, ML + bootstrap, BAMBE, and a neighbor-joining + bootstrap method all gave very high support (97 percent or higher) to the expected clusters [4]. This indicates that a 3-year gap is sufficient to define a clear "time signature" for this region of the influenza genome. Clearly, maximum likelihood plus bootstrap ("ML + bootstrap") or Bayesian methods are defensible but computationally demanding methods. That is why we applied these methods to relatively small subsets of the data in the examples above.

A recent application of clustering in the context of choosing the number of subtypes of HIV-1, group M required a more automated method (such as model-based clustering) of choosing the number of clusters [3]. We believe that model-based clustering should be useful in other applications.

4.1 Hierarchical Clustering

Hierarchical clustering is one less computationally demanding alternative to the " ML + bootstrap" approach. The input is an n-by-n symmetric matrix of all pairwise distances computed under the chosen metric. The tree begins with each sequence as its own cluster and the sequences are merged successively until all sequences form one cluster. Figure 1b and Figure 2 are examples

of a neighbor joining method [33], which is a specialized type of hierarchical clustering. The choice of how many clusters could be made by appealing to any of several subjective criteria, including those used by the human eye.

4.2 Kmeans Clustering

Kmeans clustering [19] could be applied to the data if the data (A, C, T, and G) were coded so that distances could be computed. The most defensible way to do this that we are aware of is to represent the pairwise distance data via multidimensional scaling. Multidimensional scaling [35] represents the data in new coordinates such that distances computed in the new coordinates very closely approximate the original distances computed using the chosen metric. For n sequences with an n-by-n distance matrix, the result of multidimensional scaling is an n-by-p matrix (p coordinates) that can be used to closely approximate the original distances. Therefore, multidimensional scaling provides a type of data compression with the new coordinates being suitable for input to kmeans or model-based clustering. We use the cmdscale function in Splus5.1 [32] to implement classical multidimensional scaling. The implementation of model-based clustering that we consider includes kmeans as a special case as we describe in the next subsection. Model-based clustering is the new more-scaleable alternative that has several attractive features.

4.3 Model-based Clustering

In model-based clustering, it is assumed that the data are generated by a mixture of probability distributions in which each component of the mixture represents a cluster. Given n p-dimensional observations $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$, assume there are G clusters and let $f_k(\mathbf{x}_i|\theta_k)$ be the probability density for cluster k. The model for the composite of clusters is typically formulated in one of two ways. The classification likelihood approach maximizes

$$L_C(\theta_1,\ldots\theta_G;\gamma_1,\ldots,\gamma_n \mid \mathbf{x}) = \Pi_i \; f_{\gamma i}(\mathbf{x}_i \mid \theta_{\gamma i}),$$

where the γ_i are discrete labels satisfying $\gamma_i = k$ if $\mathbf{x}_i$ belongs to cluster k. The mixture likelihood approach maximizes

$$L_M(\theta_1,\ldots\theta_G;\tau_1,\ldots,\tau_G \mid \mathbf{x}) = \Pi_i \Sigma_k \; \tau_k f_k(\mathbf{x}_i \mid \theta_l),$$

where τ_k is the probability that an observation belongs to cluster k.

Fraley and Raftery describe their latest version of model-based clustering in [11] where the f_k are assumed to be multivariate Gaussian with mean μ_k and covariance matrix Σ_k. Banfield and Raftery [1] developed a model-based framework by parameterizing the covariance matrix in terms of its eigenvalue decomposition in the form $\Sigma_k = \lambda_k D_k A_k D_k^T$, where D_k is the orthonormal matrix of eigenvectors, A_k is a diagonal matrix with elements proportional to the eigenvalues of Σ_k and λ_k is a scalar, and under one convention is the largest eigenvalue of Σ_k. The orientation of cluster k is determined by D_k, A_k determines the shape, while λ_k specifies the volume. Each of the volume, shape, and orientation (VSO) can be variable among groups, or fixed at one value for all groups. One advantage of the mixture-model approach is that it allows the use of approximate Bayes factors to compare models, giving a means of selecting the model parameterization (which of V, S, and O are variable among groups) and the number of clusters. The Bayes factor [1] is the posterior odds for one model against another model assuming that neither model is favored a priori (uniform prior). When the EM algorithm (estimation-maximum likelihood [6]) is used to find the maximum mixture likelihood, the most reliable approximation to twice the log Bayes factor (called the Bayesian Information Criterion, BIC) is BIC = $2l_M(x,\hat{\theta}) - m_M \log(n)$, where $l_M(x,\hat{\theta})$ is the maximized mixture loglikelihood for the model and m_M is the number of independent parameters to be estimated in the model. A convention for calibrating BIC differences is that differences less than 2 correspond to weak evidence, differences between 2 and 6 are positive evidence, differences between 6 and 10 are strong evidence, and differences more than 10 are very strong evidence [20].

We give an example in Figure 3 of the performance of emclust (available from www.stat.washington.edu/fraley for use in Splus) on simulated multivariate Gaussian data with n = 30 observations in each of 3 groups and p = 2 variables. The correct model is model 4 which is denoted VVV (varying volume, shape, and orientation). We are presently doing a more extensive evaluation of emclust for simulated Gaussian data, and evaluating emclust on simulated data from nonGaussian distributions with a catch-all "noise cluster" [1].

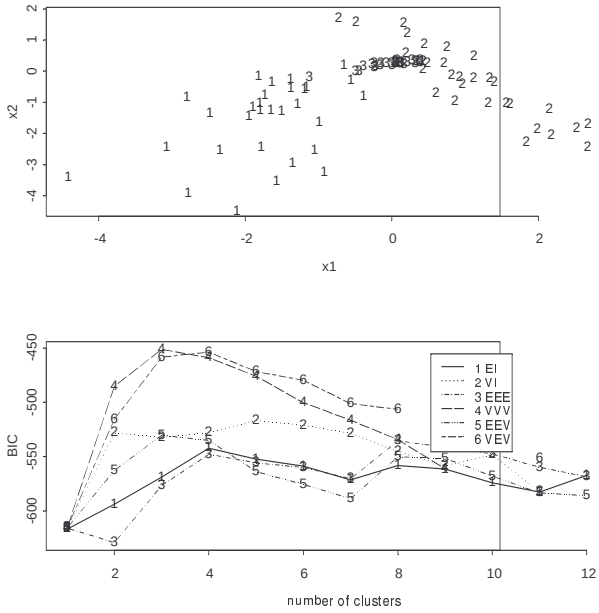

Fig. 3. Evaluation of emclust for a simulated data set of 30 observations from each of 3 clusters (labeled 1, 2, 3 in top plot) with true model VVV denoting that the volume, shape (defined as the ratio R of largest to smallest eigenvalue in the p = 2 case), and orientation all vary among clusters. The BIC correctly chooses 3 clusters and the VVV model 4 with model 6 (VEV) a close second. The VEV model has varying volume, equal shape, and varying orientation among clusters.

5. MODEL-BASED CLUSTERING EXAMPLES

5.1 Influenza Data

Here we return to the examples presented in Section 2. We anticipate host-specific signatures for NP and time-specific signatures for HA Figure 4a is the principle coordinate representation of the pairwise distances (distances computed using the HKY5 model) of 21 of the NP sequences (7 each of H, A, and S). Informally, we observe 3 clusters, or perhaps 4 because of an apparent subcluster among the S. Figure 4b is the BIC calculated from emclust. Figure 5a is the principle coordinate representation of the pairwise distances (distances computed using the HKY5 model) of 8 of the 1993 HA sequences and 8 of the 1996 sequences ("3" denotes the 1993 sequences and "6" denotes the 1996 sequences). Figure 5b is the BIC calculated from emclust. The BIC is not evaluated for any model with insufficient number of observations in a cluster or numerical instability in the estimated covariance matrix. For example, model 4 is not evaluated with 5 clusters in Figure 5. The "correct" number of clusters is arguably 2 or 4 in the collection of 8 1993 and 8 1996 HA sequences. Figure 6 provides a possible explanation for the 2 subclusters in the 1993 sequences. One subcluster went extinct and the other has existing lineages in 1996.

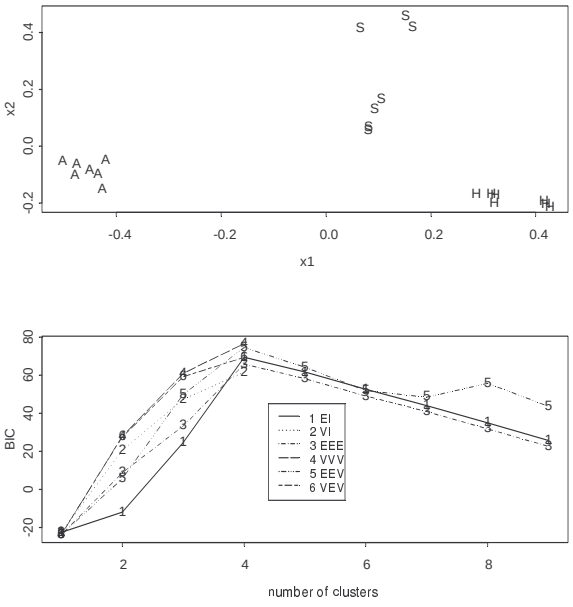

Fig. 4. (Top) Principle coordinate plot of 21 pairwise distances (computed with HKY5 model) of the data from Fig 1b. (Bottom) Result of emclust applied to the principle coordinates in the top plot, suggesting 4 clusters (S has an apparent subcluster).

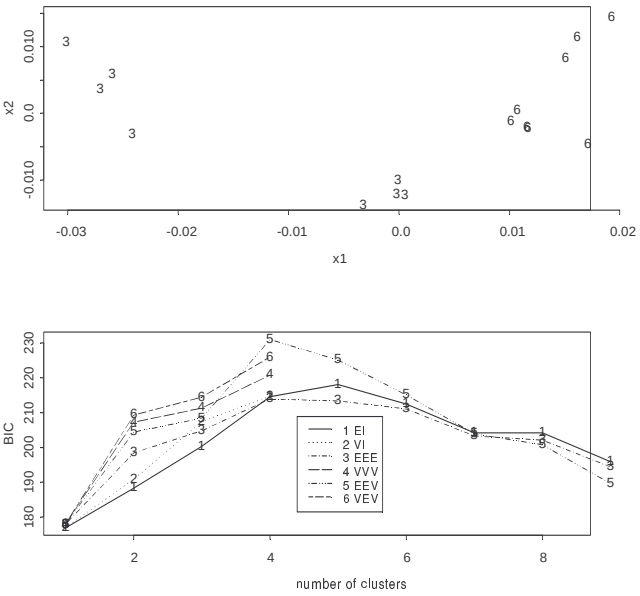

Fig. 5. (Top) Principle coordinate plot of 16 pairwise distances (computed with HKY5 model) of the data from Figure 2. (Bottom) Result of emclust applied to the principle coordinates in the top plot, suggesting 4 clusters (1993 and 1996 each have apparent subclusters).

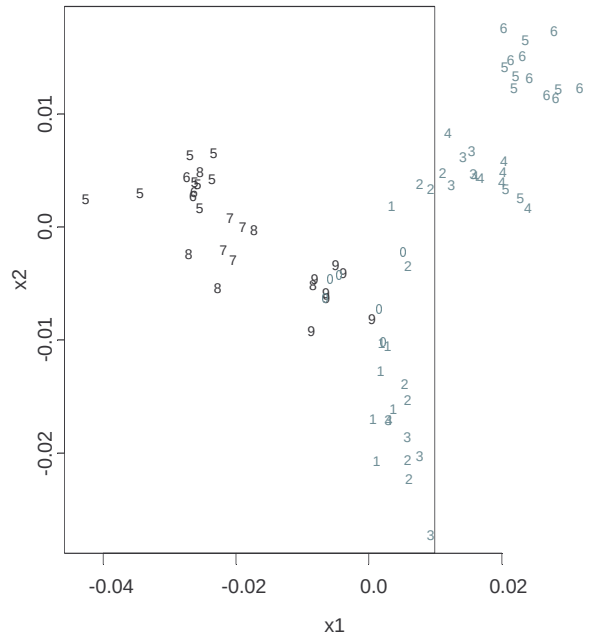

Fig. 6. Principal coordinate plot (distances via HKY-5 model) of 79 HA genes from 1985 to 1996 (1968 case omitted). The digit is the year with the 1980 decade in darker type than 1990 (left-most case is 5 = 1985). The eight 1993 cases separate into a "close to 1996" subgroup (surviving lineage?) and a "far from 1996" (extinct lineage?) subgroup so the "correct" number of clusters is not necessarily two.

5.2 HIV Data

Here we introduce HIV data from the *gag*-p17 and *env*-gp120 regions of the HIV genome. The main HIV epidemic is caused by HIV-1, group M which currently has approximately 10 subtypes [13,16,27]. It is well known that some regions of the HIV genome are more conserved (have smaller genetic distance for the same time separating the sequences) than others, so subtype definitions can depend on the region of the genome. Another important feature of HIV evolution is recombination which could, for example, cause a sequence to have subtype A for the *gag* region and subtype E for the *env* region (as in the "Thailand AE sequences"). For this example, we have excluded all obvious inter-subtype recombinants (as outliers) and we consider both the *gag-p17* (more conserved) and the *env-gp120* (less conserved) regions of the genome.

5.2.1 Data

We selected n = 95 *env* sequences and n = 88 *gag* sequences. Some of the sequences have known isolation times and all have been subtyped using many methods, including ML + bootstrap.

The ML + bootstrap approach suggests 7 clusters (subtypes) in the *env* sequences and 6 clusters (subtypes) in the *gag* sequences. The data is available at hiv-web.lanl.gov [36] and accession numbers are available upon request.

5.2.2 Evolutionary Model

We used PAUP [34] to estimate the substitution rate μ (section 4 and appendix), 5 relative rate parameters, 3 relative frequencies, and the rate heterogeneity parameter in a GTR. The main two summary parameters of the model are μ and γ, which were estimated to be 0.0026 and 0.45 for *gag* and 0.0035 and 0.40 for *env*. Our selected model agrees very closely with published models [23, 24] for slightly different subregions of the *gag* and *env* egions, which provides increasing confidence that our model estimate is close to the best model for the *gag*-p17 and *env*-gp120 regions.

5.2.3 Evolutionary Distances

The evolutionary distances between each pair of sequences were computed using PAUP from the chosen model as described in sections 3, 4, and the appendix.

5.2.4 Model-based Clustering

We applied the cmdscale function in Splus5.1 [32] to implement classical multidimensional scaling [35] of the $\binom{n}{2}$ pairwise distances in a matrix $D_{original}$ for each of the *gag* and *env* regions. There is a choice of the number of principal coordinates (similar to principal components) [19,35] to use to recover the original distance matrix. This choice can be objectively made by evaluating the distance between the original and approximate distance matrices. For each candidate number of principle coordinates, we compute the $\binom{n}{2}$ pairwise distances as D_{approx} and compute $D_{diff} = D_{approx} - D_{original}$ (or a scaled version where each difference is rescaled by $D_{original}$ [35]). As we increase the number of coordinates we find the cluster number above which the decrease in D_{diff} becomes less than some predetermined threshold. We typically retain 2 to 10 coordinates using this strategy. Assume we have retained 3 principal coordinates. These 3 coordinates become the input to emclust to find the suggested number of clusters and the model parameterization. Example results (plotting only the first 2 of 10 principle coordinates) are given in Figure 7 for *env* (BIC peaks at 6 clusters, tending to merge subtypes B and D compared to the ML approach) and Figure 8 for *gag* (BIC peaks at 6 clusters, agreeing with the ML approach).

6. SCALING TO MORE SEQUENCES OR MANY DATA SETS

We envision at least two types of scaling that would be of practical interest. First, we recently evaluated [3] the number of clusters that are produced from many repeated simulations of each of several models of how HIV is evolving (with respect to classical epidemiology in terms of the numbers of new cases and with respect to molecular epidemiology in terms of substitution rates). Using "ML + bootstrap" would have required human interaction with the simulated sequences from each run to choose the number of clusters, so the human time cost would have been prohibitive. Model-based clustering is more amenable to objective automation for this application. Second, we envision applications of a single clustering of n sequences where n is very large (10,000 or more). No distance-based clustering scheme can scale well for very large numbers of sequences n. Recall that we must incorporate the evolutionary model and to do so requires that we work with the joint likelihood in the ML approach or with the matrix of pairwise distances. We have focused here on the pairwise distances approach followed by multidimensional scaling so that model-based clustering can operate on a low dimensional

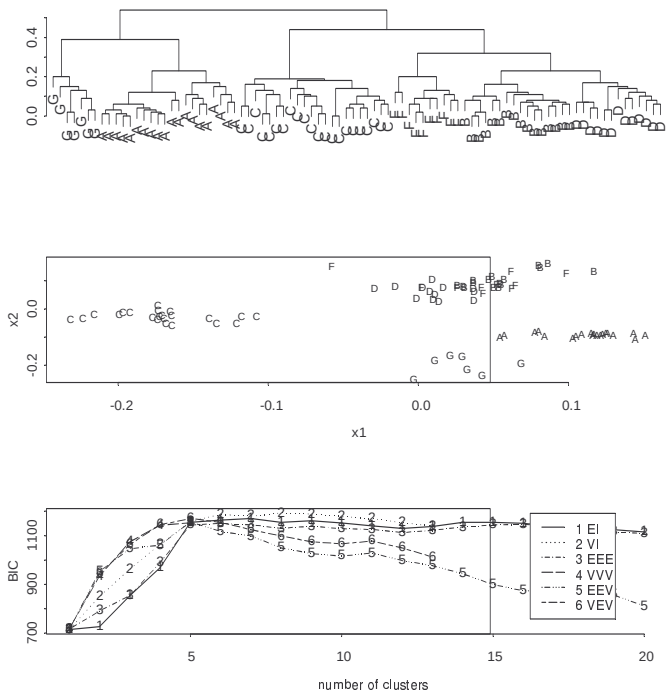

Fig. 7. Env Data. (Top) Hierarchical Clustering; (Middle) Principle Coordinate plot; (Bottom) Results of emclust.

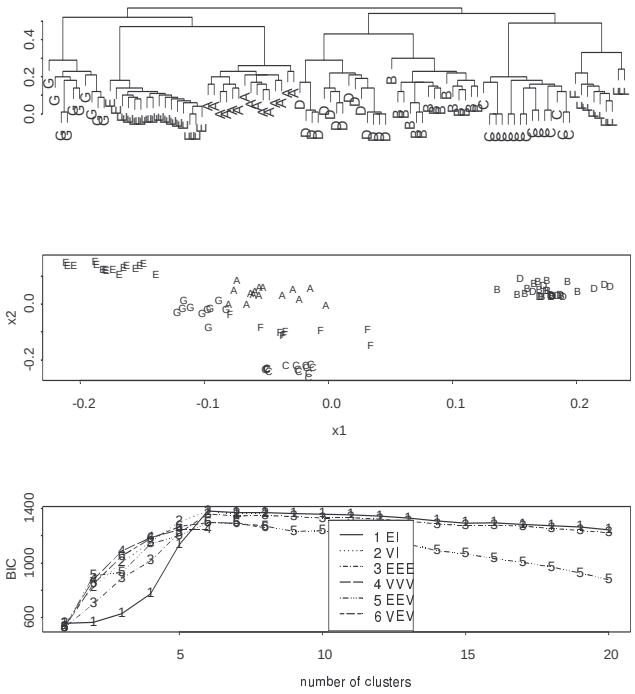

Fig. 8. Gag data. (Top) Hierarchical Clustering; (Middle) Principle Coordinate plot; (Bottom) Results of emclust.

representation that can approximately recover the original distance matrix. We used cmdscale in Splus, but a faster alternative is FASTMAP [8]. Given some type of multidimensional scaling in low dimensions (two to ten usually) we have illustrated the application of emclust. The emclust implementation of model-based clustering in Splus [32] calls Fortran routines that implement the EM algorithm which can be slow to converge especially for marginally well separated clusters. Perhaps the approach by Moore [26] would lead to a faster implementation. Also, we recognize that the output of some type of multidimensional scaling could also be input to any of the relatively fast algorithms that allow for variable-shaped clusters, such as BIRCH [37], CURE [14], or CLARANS [29], provided they use a criterion such as the BIC to suggest the number of clusters. Reference [2] provides a good review of some general issues to consider regarding scalability of clustering methods.

Fortunately, for the application of choosing the number of clusters in DNA data, we normally can appeal to statistical sampling theory in conjunction with ideas from coalescent theory [21] to argue that the estimated number of clusters $\hat{G}$ increases as n increases up to some limit, and then $\hat{G}$ saturates for large n. Therefore, there is no need to perform cluster analysis using extremely large numbers of sequences n. Whether this saturation effect actually occurs can easily be objectively measured by tracking the behavior of the estimated number of clusters $\hat{G}$ as n increases. We show an example in Figure 9. Figure 9 shows the estimated number of clusters $\hat{G}$ in data simulated from a simple coalescent model for n =100 (left column) and n = 500 (right column) sequences. Note that $\hat{G}$ is approximately 5 for both n =100 and n = 500. The coalescent model specifies how each new generation of HIV cases is produced from the previous generation, and we use an implementation of coalescent theory (Treevolve in [13], available at website [36]) to infer typical sample genealogies and the commensurate DNA data. For real data, we take the same approach by first clustering with n = 50, n = 60, n = 70, etc. until the estimated number of clusters stabilizes at some constant value.

In cases where geographic and/or temporal separation of sequences leads to "subtypes," there could be substantial genetic theory required to design the appropriate statistical sampling plan. If the estimated number of clusters fails to stabilize, we would accept that as empirical evidence that the clustering did not scale well, and perhaps a suitable "subtype" definition is not available.

Finally, in situations where we have demonstrated that there is a stable number of estimated clusters as n increases, we note that in many cases there will be relatively simple methods to: (1) assign a "cluster signature" to each cluster; and (2) assign new sequences to the appropriate cluster via a "supervised learning" effort. Several "signature patterns" have been compared, each with good success in certain applications [5]. Usually a straightforward empirical comparison of the performance of candidate signature pattern methods on held-out test sequences can be used to choose a signature pattern method. One successful method (VESPA, or viral epidemiology signature pattern analysis) searches for sites that are strongly conserved within a cluster, but not outside that cluster [5, 22]. To find signature sites for group 1, VESPA searches for sites for which group 1 has high relative frequency (p_{min} = .8 is the default value) of a given nucleotide and the non-group 1 sequences have low relative frequency (p_{max} = .2 is the default value). The same process is repeated to find signatures for all groups. There is no guarantee of finding any signature sites, but clearly, the more we find with strict p_{min} and p_{max} values, the better the group separation. Test sequences are assigned to the group having the highest percent agreement between the group's signature site values and the test sequence values at the signature sites. The parameters p_{min} and p_{max} can be selected on the basis of performance on training data at separating the groups. Less formal procedures are also sometimes useful. For example, [12] manually reviewed aligned influenza sequences to discover host-specific clusters.

To summarize the scalability issue, we do not envision a way for "ML + bootstrap" to scale well to repeated application for the same small or moderate number of sequences n. Model-based clustering appears to have a role in this type of problem regardless of whether the bootstrap is included (emclust + bootstrap scales better than ML + bootstrap). In a single application with large n, no method will work directly. It will be necessary to start with small to moderate n and apply model-based clustering (or a more manual ML + bootstrap) for each of several values of n as n increases. Presumably, ML + bootstrap will continue to be the "gold standard" as the most defensible method of semi-manually choosing the number of clusters. Even in the cases where a large careful effort is appropriate for each case considered (so that ML + bootstrap is appropriate), we believe that model-based

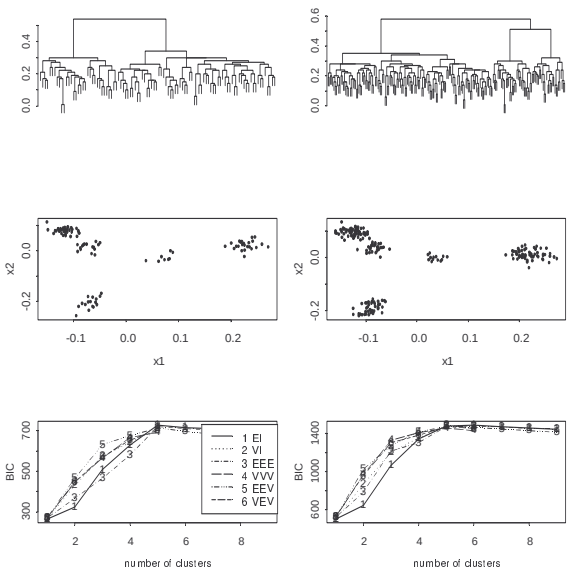

Fig. 9. Example of the impact of increasing n. (Left column) n = 100, (Right column) n = 500. The estimated number of clusters $\hat{G}$ is approximately 5 in both cases.

clustering would add valuable complementary support for the chosen number of clusters. In cases where many simulated or real data sets are evaluated, model-based clustering (perhaps with the bootstrap) could defensibly be chosen as a convenient way to automate high-throughput screening of the apparent number of clusters in each data set.

7. SUMMARY

There will always be a subjective element in any cluster analysis. Nevertheless, current implementations of model-based clustering have demonstrated good results with simulated and real data and we believe that model-based clustering has great potential for choosing the number of subtypes in genetic data. A truly novel ability of emclust is its ability to apply the BIC to choose the number of clusters and the model parameterization.

Because we rely on first computing the distance matrix of all pairwise distances, an exception is any case where the best evolutionary model is too complex to allow a defensible distance measure to be defined. To date, we are unaware of any real data set for which there is a compelling reason to work with a model that does not allow a commensurate distance measure. Future work should address the robustness of current implementations of model-based clustering to non-Gaussian data, include more evaluations of the clustering results in simulated data with various degrees of cluster separation and sample sizes, and compare model-based clustering results to results using ML + bootstrap. The BIC is established as a good criterion for choosing the number of clusters, but there are others, such as the Akaike's information criterion (AIC). The AIC tends to overestimate the number of clusters. Therefore, in cases where the BIC tends to underestimate the number of clusters, it might be possible to claim that with high probability, the AIC-based result is an upper bound while the BIC-based result is a lower bound. Finally, methods to combine clustering results over different regions of the genome in cases where no recombination has occurred are of interest. Distance-based methods could compute an average distance matrix over all regions of the genome that could be input to model-based clustering. The alternative is to somehow combine clustering results from each genome region, which is less straightforward.

8. ACKNOWLEDGMENTS

We thank Gerry Myers and Mac Hyman for informative discussions about clustering genetic data and are grateful to the referees whose comments greatly improved this paper.

9. REFERENCES

[1] Banfield, J. and Raftery, A. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49:803-821, 1993.

[2] Bradley, P., Fayyad, U., and Reina, C. Scaling Clustering Algorithms to Large Databases. *Proceedings of the 4th International Conf. on Knowledge Discovery and Data Mining* (KDD-98). AAAI Press, Aug. 1998.

[3] Burr, T., Myers, G., and Hyman, J. The origin of AIDS – Darwinian or Lamarkian? *Phil. Trans. R. Soc. Lond. B.356:877-*887, 2001

[4] Burr, T., Skourikhine, A.N., Macken, C., and Bruno, W. Confidence measures for evolutionary trees: applications to molecular epidemiology. P*roc. of the 1999 IEEE Inter. Conference on Information, Intelligence and Systems*, 107-114, 1999.

[5] Burr, T., Charlton, W., and Stanbro, W. Comparison of signature pattern analysis methods in molecular epidemiology. *Mathematical and Engineering Methods in Medicine and Biological Sciences*, 473-479, 2000.

[6] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1-38, 1977.

[7] Efron, B., Halloran, E., and Holmes, S. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* 93: 13429, 1996.

[8] Faloutsos, C. and Lin, K. FastMap: A fast algorithm for indexing, data-mining, and visualization of traditional and multimedia datasets. *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, San Jose, California, pages 163-174, May 22-25, 1995

[9] Felsenstein, J. Evolutionary trees from DNA sequences*:* a maximum likelihood approach. J. Mol. Evol. 17:368-376, 1981.

[10] Felsenstein, J. Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* 22:521-565, 1997.

[11] Fraley, C. and Raftery, A. MCLUST: Software for model-based cluster analysis. *Journal of Classification* 16:297-306, 1999.

[12] Gammelin, M., Mandler, J., and Schholtissek, C. Two subtypes of nucleoproteins (NP) of the influenza viruses. *Virology* 170:71-80, 1989.

[13] Grassley, N.C., Harvey, P.H., and Holmes, E.C. Population dynamics of HIV-1 inferred from gene sequences. *Genetics* 151: 427-438, 1999.

[14] Guha, S., Rastogi, R., and Shim, K. CURE: An efficient clustering algorithm for large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 73-84, New York, 1998. ACM.

[15] Hasegawa, M., Kishino, H., and Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 21: 160-174, 1985.

[16] Holmes, E.C., Pybus, O.G., and Harvey, P.H. The molecular population dynamics of HIV-1. In Crandell, K. *The Evolution of HIV,* Baltimore: Johns Hopkins University Press, 1999.

[17] Hu, D.J., Buve, A., Baggs, J., van der Groen, G., and Dondero, T.J. What role does HIV-1 subtype play in transmission and pathogenesis? An epidemiological perspective. AIDS 13:873-881, 1999.

[18] Huelsenbeck, J. and Rannala, B. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*, 276: 227-232, 1997.

[19] Johnson, R. and Wichern, D. *Applied Multivariate Statistical Analysis*, 2nd edition. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[20] Kass, R. and Raftery, A. Bayes Factors. *J. American Statistical Association*. 90:773-795, 1995.

[21] Kingman, J.F.C. On the genealogy of large populations. *J. Appl. Prob*. 19: 27-43. 1982.

[22] Korber, B. and Myers, G. Signature pattern analysis: a method for assessing viral sequence relatedness. *AIDS Research and Human Retroviruses* 8: 1549-1560, 1992.

[23] Leitner, T., Kumar., S., and Albert, J. Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in HIV Type 1 populations with a known transmission history. *Virology* 71: 4761-4770, 1997.

[24] Leitner, T., et al, Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci., USA* 93: 10864 –10869, 1996.

[25] Mau, B., Newton, M., and Larget, B. Bayesian phylogenetic inference via Markov Chain Montre Carlo Methods. *Biometrics* 55:1-12, 1999.

[26] Moore, A. very fast EM-based mixture model clustering using multiresolution kd-trees. *Neural Information Processing Systems*, December 1998 Issue. The paper is available online at http://www.cs.cmu.edu/~awm/papers.html#fastem

[27] Myers, G. HIV: between past and future. *AIDS Res Human Retro* 10: 1317-1324, 1994.

[28] Needleman, S. and Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol Biol*. 48:443-453, 1970.

[29] Ng, R. and Han, J. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the VLDB Conference*, Santiago, Chile, September 1994.

[30] Salter, L. Algorithms for phylogenetic tree reconstruction. *Mathematical and Engineering Methods in Medicine and Biological Sciences*, 459-465, 2000.

[31] Simon, D. and Larget, B. Bayesian Analysis in Molecular Biology and Evolution (BAMBE) version 1.01 beta, Dept. of Mathematics and Computer Science, Duquesne University, 1998.

[32] S-Plus 5.1 MathSoft, Seattle Washington, 1999.

[33] Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. Phylogenetic inference In *Molecular Systematics, 2nd edition*, pp. 407-514 (Hillis et al., eds.) Sunderland, Massachusetts: Sinauer Associates, 1996.

[34] Swofford, D.L. PAUP* Phylogenetic analysis using parsimony; Version 4; Sunderland, Massachusetts: Sinauer Associates, 1999.

[35] Venables, W. and Ripley, B. *Modern applied statistics with S-PLUS*, 2nd ed., Springer-Verlag: NY, 1997.

[36] Web sites: *hiv-web.lanl.gov* for the HIV sequences; *linker.lanl.gov/flu* for the influenza sequences; *www.stat.washington.edu/fraley* for emclust code for use in Splus; *http://evolve.zoo.ox.ac.uk* for Treevolve code to simulate DNA data under various coalescent models.

[37] Zhang, T., Ramakrishnan, R., and Livny, M. BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 103--114, Montreal, Canada, 1996.

About the authors:

Tom Burr received his BS in math (Clemson, 1981), and in physics (UNC-Asheville, 1985) and a Ph.D. in statistics from Florida State University in 1992. He is a technical staff member at Los Alamos National Laboratory in New Mexico, specializing in the application of statistical and data mining methods to a range of problems in the physical sciences, including evolutionary biology and genetics.

James R. Gattiker received his BS and MS degrees from Penn State University, and Ph.D. in 1996 from Binghamton University in Computer Engineering. His studies and research included data mining models and algorithms, knowledge representation, artificial intelligence and cybernetics, and models of biological and evolutionary systems.

Greggory S. LaBerge received his B.Sc. Hons. in Molecular Biology and Genetics from the University of Guelph, his M.Sc. degree in Biostatistics from the University of Colorado Health Sciences Center and is currently studying for his Ph.D. in Human Medical Genetics at the University of Colorado Health Sciences Center. His studies and research include mathematical models of Human population genetics, forensic genetics with casework applications, and Human disease genetics.

Appendix. EVOLUTIONARY MODELS AND ASSOCIATED GENETIC DISTANCES

Consider a pair of aligned sequences denoted *x* and *y*. Define F_{xy} as

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

N is the number of base pairs (sites) and n_{ij} is the number of sites where sequence *x* has base *i* and sequence *y* has base *j*. The most general time-reversible model (GTR) for which a distance measure has been defined [33] defines the distance between sequences *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 sequences *x* and *y*. 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 (each row of Q sums to zero and f=1 by convention)

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

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 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 sequences *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). Another common model is the HKY5 model (Hasegawa et al [15, 31]) with unequal base frequencies and different rates for transitions from purine to purine (A and G are purines) or from pyrimidines to pyrimidines (C and T are pyrimidines) than for transversions between purine and pyrimidine or vice versa. Another important generalization is to allow the rate μ to vary across DNA sites. Allowing for hot (higher μ) and cold (lower μ) sites via a gamma-distributed rate parameter μ is one way to model the fact that "silent" and "replacement" sites often have different observed rates. A silent mutation does not alter the amino acid for which the DNA is coding, and usually occurs at the third site because of the nature of the redundancy in the amino acid code. 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 formula

d_{xy} = -trace{$\Pi \log(\Pi^{-1}F_{xy})$} [33].

For sequences having small to moderate percent difference *D*, if the correct model is chosen, the expected distance will increase linearly with time. However, 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 the correct metric such that the distance between any two sequences increases linearly with time.

It is possible to generalize further by relaxing the time reversibility assumption and use "log-determinant" transformations that remain "additive" under a wider set of models [33]. An additive distance satisfies the rule that the expected distance between each pair of sequences is equal to the sum of the lengths of each branch on the path connecting the sequences (and hence is an ultrametric [33]). Nearly all published analyses assume the GTR or GTR + rate heterogeneity models or a special case of these. The most general model we consider here is GTR + rate heterogeneity. A known weakness of all such models is that they assume each site evolves independently. There have been a few attempts to relax this site-by-site independence assumption, but computational and data demands have not permitted significant progress toward relaxing the assumption.

Another important fact regarding cluster analysis of DNA data is that different regions of the genome might give different cluster results. We consider this to be an important topic that deserves more attention than space permits here. For our purposes, we will consider only one region at a time, or assume that an average distance matrix has been derived by combining distances based on different regions of the genome. An important issue is whether recombination could have led to a situation in which the true genealogy of the sampled sequences depends on the region of the genome. We assume here that there is no recombination so that there is only one true genealogy for all regions of the genome.

Because of the trend toward using the most realistic feasible evolutionary model and the fact that most models have an associated distance measure, distance-based clustering methods are appropriate for our intended applications. Another candidate involves using maximum likelihood plus a resampling scheme such as the bootstrap as we describe in Section 4.