Let Them Fall Where They May: Congruence Analysis in Massive Phylogenetically Messy Data Sets

Jessica W. Leigh,*[1] Klaus Schliep,[2] Philippe Lopez,[2] and Eric Bapteste[2]

[1]Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

[2]UMR CNRS 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie, Paris, France

***Corresponding author:** E-mail: jleigh@maths.otago.ac.nz.

Associate editor: Arndt von Haeseler

Research article

Abstract

Interest in congruence in phylogenetic data has largely focused on issues affecting multicellular organisms, and animals in particular, in which the level of incongruence is expected to be relatively low. In addition, assessment methods developed in the past have been designed for reasonably small numbers of loci and scale poorly for larger data sets. However, there are currently over a thousand complete genome sequences available and of interest to evolutionary biologists, and these sequences are predominantly from microbial organisms, whose molecular evolution is much less frequently tree-like than that of multicellular life forms. As such, the level of incongruence in these data is expected to be high. We present a congruence method that accommodates both very large numbers of genes and high degrees of incongruence. Our method uses clustering algorithms to identify subsets of genes based on similarity of phylogenetic signal. It involves only a single phylogenetic analysis per gene, and therefore, computation time scales nearly linearly with the number of genes in the data set. We show that our method performs very well with sets of sequence alignments simulated under a wide variety of conditions. In addition, we present an analysis of core genes of prokaryotes, often assumed to have been largely vertically inherited, in which we identify two highly incongruent classes of genes. This result is consistent with the complexity hypothesis.

Key words: phylogenetic congruence, phylogenetic networks, spectral clustering, lateral gene transfer, prokaryote phylogeny.

Introduction

The notion of phylogenetic incongruence predates molecular phylogeny, though the many biological sources of incongruence in molecular data (e.g., hybridization [McBreen and Lockhart 2006; Koblmüller et al. 2007], incomplete lineage sorting [Hudson 1983; Hobolth et al. 2011], lateral gene transfer [Bapteste et al. 2009]) have certainly raised awareness of the importance of incongruence among evolutionary biologists in recent years. As sequence databases have grown and computational power has increased, numerous congruence assessment methods have been developed. These methods can loosely be divided into two classes. The topology-based methods use as null hypothesis complete lack of correlation between trees and directly compare topologies (Lapointe and Rissler 2005; de Vienne et al. 2007; Nye 2008; Puigbò et al. 2009). These tests have used a variety of distance metrics between tree topologies, including partition metrics (Robinson and Foulds 1981; Penny et al. 1987), maximum agreement subtrees (Bryant et al. 2003), pruning distances (Křivánek 1986; Bordewich and Semple 2004; Wu 2009), quartet distance (Estabrook et al. 1985), and path distance (Steel and Penny 1993). Distances between weighted trees (i.e., taking branch lengths into account) have also been described (e.g., Waddell et al. 2007). Other methods, such as the Congruence Among Distance Matrices test (Campbell et al. 2009, 2011), compare distance matrices rather that trees to assess the null hypothesis of incongruence. Topology-based methods have been very useful in fields such as phylogeography and studies of coevolution, where any correlation in different trees is of interest.

For phylogenomics and multigene phylogeny, however, the interest often lies in determining whether tree topologies for different genes are exactly identical, either to demonstrate that different markers share the same pattern of inheritance or that combined analysis is justified. In this case, a null hypothesis in which genes share the same topology is used and rejected in order to identify incongruence. Normally, these methods are classified as character-based because topologies are not compared directly; rather, the methods evaluate the fit of different topologies to different markers.

However, existing methods have largely been developed with eukaryotes in mind. Yet most available genome sequences are from prokaryotes and viruses (1,454 and 2,567, respectively, compared with 41 from eukaryotes according to NCBI Genome, http://www.ncbi.nlm.nih.gov/sites/genome last accessed: 6 May 2011), which also make up most of the biological diversity on earth. Sequence evolution of these molecules faces very different constraints compared with sequences from eukaryotes, as horizontal (lateral) evolution occurs far more frequently (Bapteste et al. 2009). Whereas statistical tests for congruence often postulate congruence as the null hypothesis, failure to reject congruence with

© The Author 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

phylogenetic data from prokaryotes may indicate a lack of phylogenetic signal, rather than real congruence in the data set. Therefore, congruence methods may underestimate the level of incongruence in these data.

Moreover, most sequence-based congruence tests were not designed for modern phylogenomic data sets. The incongruence length difference (ILD) test (Farris et al. 1994) can assess congruence in a multigene data set only by testing whether the entire data set is congruent and cannot identify which genes are congruent with one another; likelihood-based equivalents (Huelsenbeck and Bull 1996; Waddell et al. 2000) share this property. Both the ILD (Planet and Sarkar 2005) and likelihood ratio tests for congruence (Leigh et al. 2008) have been adapted to identify congruent subsets of multigene data sets, but these hierarchical tests scale poorly with the number of markers in the data set and are plagued with multiple testing problems.

Clustering-based classification methods have also been investigated for the purpose of assessing congruence (Brochier et al. 2002; Gribaldo and Philippe 2002; Matte-Tailliez et al. 2002; Nye 2008). Different clustering methods exist, including hierarchical methods (such as the unweighted pair group method with arithmetic mean [UPGMA]), the classic *k*-means algorithm, and nonlinear dimensionality reduction methods such as multidimensional scaling (Sammon 1969; Edwards and Oman 2003) or spectral clustering methods (e.g., Ng et al. 2001; Zelnik-Manor and Perona 2004; Newman 2006), and unsupervised methods for estimating the number of clusters (e.g., Tibshirani et al. 2001; Von Luxburg 2007) have been developed. Clustering is immune to multiple testing errors associated with repeated or hierarchical statistical tests because there is no confidence level below which some null hypothesis is rejected and because there is no repeated testing involved.

Here, we present a novel algorithm for clustering by phylogenetic distance (CPD), implemented in the application Conclustador for evaluating congruence in phylogenomic data from complete genomes. The distances used by this algorithm are related to distances used in topology-based congruence tests but take into account uncertainty in phylogenetic estimates by representing each marker by a distribution, rather than a single tree. The CPD algorithm uses clustering and scales linearly with the number of markers in the data set with respect to the slow phylogenetic analysis step. We demonstrate the effectiveness of our method with sequence data simulated under a variety of conditions, as well as with a recently published data set of 114 alignments and 100 operational taxonomic units (OTUs) (Puigbò et al. 2009).

Despite the wealth of literature on phylogenetic congruence, definitions of congruence have varied. In this work, we will use "incongruence" to mean specifically topological incongruence, rather than differences in other evolutionary parameters among phylogenetic markers (e.g., relative evolutionary rate in different lineages; Waddell et al. 2007; Leigh et al. 2008). In addition, we do not consider incongruence to be an all-or-nothing property; that is, a pair of tree topologies can be more congruent than another pair if they share a larger number of phylogenetic relationships (local congruence) while remaining globally incongruent.

Materials and Methods

Algorithm Overview

The CPD algorithm involves calculating Euclidean distances between markers based on a distribution of phylogenetic trees for each gene, then clustering these distances. Effectively, this is a topology-based method, but each gene is represented by a distribution of topologies, and no explicit hypothesis of congruence or incongruence is tested. Ultimately, the goal is not to assess whether the set of markers is congruent or incongruent, but rather to "let them fall where they may": that is, to find the other markers with which they share a level of congruence that is distinguishable from the background of the entire data set. This algorithm is implemented in Conclustador, available by request from the authors.

Phylogenetic Analysis

Conclustador takes as input a distribution of phylogenetic trees for each marker. Phylogenetic analysis is not integrated into Conclustador, as numerous available methods might be preferred by different users. We evaluated the performance of Conclustador with tree distributions inferred by the neighbor-joining (NJ) distance method, maximum likelihood (ML), maximum parsimony (MP), and Bayesian inference (BI). All scripts used for phylogenetic analysis were written in Python and are distributed with Conclustador.

With the exception of the BI method, all distributions were obtained from either 100 (ML) or 1,000 (MP and NJ) nonparametric bootstrap replicates for each marker. For the MP method, we used PAUP* (Swofford 2003), with the heuristic search option. For ML, we used RAxML (Stamatakis 2006) with the WAG substitution model (Whelan and Goldman 2001) and the "rapid bootstrapping" method (Stamatakis et al. 2008). We used Tree-Puzzle (Schmidt et al. 2002) to estimate distances quickly under the WAG model with constant rates across sites, along with BioNJ (Gascuel 1997) to infer distance trees.

BI was performed using PhyloBayes (Lartillot and Philippe 2004) with the C20 substitution model (Quang et al. 2008) and four-class discrete Γ distribution. Burnins were determined automatically by a method adapted from that of Beiko et al. (2006). Briefly, the mean log-likelihood was calculated from sliding windows of 500 consecutive samples of the chain, with a slide of 50 samples. The burnin was identified as the center of the first window whose mean log-likelihood was within 1 standard deviation of the mean log-likelihood of the last window of the chain. The burnin was thus constantly adjusted until the estimated value was less than 10% of the remaining postburnin samples, and the remaining samples were used as the distribution of trees.

Distance Calculation and Clustering

For a given distribution of trees, Conclustador begins by calculating a matrix of distances between markers. The distances used are Euclidean distances calculated between observed bipartition frequencies in tree distributions (Bayesian posterior or bootstrap distributions), as described in equation (1), where B is the set of all possible bipartitions of taxa and $\mathrm{P}(b|i)$ is the posterior probability (or bootstrap frequency) of bipartition b, given gene i.

$$d_{i,j} = \sqrt{\sum_{b\epsilon B}(\mathrm{P}(b|i) - \mathrm{P}(b|j))^2}. \qquad (1)$$

If a bipartition is not observed in the posterior or bootstrap distribution for a given gene, its probability is assumed to be 0; thus, in practice, it is unnecessary to consider unobserved bipartitions when calculating $d_{i,j}$ for a given pair of genes. In cases where genes do not share exactly the same set of taxa, taxa missing in either gene are removed prior to calculating the distance between markers. If the number of taxa shared is less than four, the distance is considered infinitely large. Because the distance value described in equation (1) increases as a function of the number of taxa shared between the two genes, distances are divided by $(2x_{i,j} - 6)^{1/2}$, the maximum distance between genes (i.e., if they shared no nontrivial bipartitions with frequency greater than zero; eq. 2), such that they remained constant with respect to the number of taxa shared between the two markers, $x_{i,j}$ (supplementary fig. S1, Supplementary Material online)

$$d'_{i,j} = \frac{d_{i,j}}{(2x_{i,j} - 6)^{1/2}}. \qquad (2)$$

Because elements of the set of bipartitions are not independent (i.e., clans are nested within clans), the "axes" of the space in which distances are estimated are not orthogonal. This could potentially lead to erratic behavior, as a small number rearrangements could produce very large distances. In order to verify that strange behavior is rare, we simulated data under a wide variety of scenarios (described below).

Markers can then be clustered based on these distances. We implemented two different clustering methods in order to compare their performance. We chose not to use hierarchical clustering because errors that occur early on in the clustering process can have drastic effects on the clusters found. Instead, we first used the classic k-means method (MacQueen 1967). In k-means, centroids are first chosen either at random or by some heuristic and then individual observations are assigned to the cluster defined by the nearest clusters. Next, centroids are refined by finding the point that minimizes the average distance to members of each cluster. The process of assigning observations to clusters and then refining centroids continues iteratively until convergence. In this case, because distances between genes were defined but genes themselves were not defined by positions in space, the centroids were necessarily assigned to the gene with the smallest total distance to other genes within a cluster, rather than to the coordinates of the center of the cluster.

With k-means, centroids are normally initially chosen at random from among the data points, which can lead to a local optimum, so clustering is repeated several times from independent starting points to increase the chance of finding the globally optimal clusters. Instead of using a fixed number of iterations, Conclustador implements the Death of Dodos algorithm (Roberts and Solow 2003; Vinh le and Von Haeseler 2004) in order to estimate the number of rounds of k-means required to find the global optimum.

We also implemented a spectral clustering method (Ng et al. 2001; Zelnik-Manor and Perona 2004). Spectral clustering is particularly useful when members of a cluster are sometimes closer to members of other clusters than to some of the members of their own cluster. In our implementation, the distance matrix is used to construct the undirected k-nearest neighbor graph (i.e., connecting each node to its k-nearest neighbors). The choice of k for this graph is important because the graph must be connected or spectral clustering tends to find the trivial clusters formed by connected components (Von Luxburg 2007). The k is initially set to $\log(n)$, where n is the number of nodes in the graph (genes), and then is iteratively incremented until the graph is fully connected. This graph is then used to construct an affinity matrix $\mathbf{A}$ given by equation (3), in which entries corresponding to pairs of nodes are non-zero if and only if these nodes are adjacent in the k-nearest neighbor graph, with affinities calculated according to the method of Zelnik-Manor and Perona (2004), in which local scaling of affinity values is used to accommodate clusters of different densities. Affinity values for a pair of nodes i,j were scaled by $\sigma_i\sigma_j$, the product of the distances of each node to its k^{th} nearest neighbor.

$$A_{i,j} = \begin{cases} \exp\left\{\frac{-d_{i,j}^2}{\sigma_i\sigma_j}\right\} & i,j \text{ adjacent} \\ 0 & \text{otherwise} \end{cases}. \qquad (3)$$

As described by Ng et al. (2001), the normalized Laplacian for the affinity matrix $\mathbf{A}$ was then calculated, and its eigenvectors were determined. For a given number of clusters C, the first C eigenvectors were normalized to norm 1 and used to form a $n \times C$ matrix. The rows of this matrix were then clustered by average linkage. We chose to use hierarchical clustering at this point because we found that it tended to find clusters that were identical to the optimal k-means solution, whereas k-means heuristics found the optimal solution less frequently (results not shown).

Identification of the Number of Clusters

The k-means and spectral clustering methods require that the number of clusters be given as input. Because this is generally not known for multilocus phylogenetic data, we investigated methods of estimating the number of clusters. For spectral methods, Conclustador uses the eigengap heuristic, described in Von Luxburg (2007), in which the number of clusters C is chosen such that the first C eigenvalues $\lambda_1, \ldots, \lambda_C$ are relatively large and λ_{C+1} is much smaller. With k-means clustering, Conclustador estimates the number of clusters using the CH index (Caliński and Harabasz 1974), given in equation (4):

$$\mathrm{CH}(C) = \frac{B(C)/(C-1)}{W(C)/(n-C)} \tag{4}$$

$B(C)$ and $W(C)$ are the between- and within-cluster sums of squares for C clusters, and n is the number of points (i.e., genes).

Simulations

Simulations were used to evaluate the performance of Conclustador under different conditions (supplementary table 1, Supplementary Material online). All genes were simulated using Seq-Gen (Rambaut and Grassly 1997). For a given simulation, 100 genes were produced by randomly generating a single tree topology and then LGT was simulated by rearranging this topology through a series of subtree pruning and regrafting (SPR) operations to produce the underlying topologies of each gene cluster. Tree topologies were rooted so that SPR rearrangements might reflect relationships expected in real LGT events as much as possible; however, to simplify simulations, we did not constrain SPR operations to prevent impossible scenarios, for example, transfer of a gene to an extinct ancestor of a contemporary lineage. For one additional simulation set, underlying topologies for each cluster were chosen independently at random, rather than according to SPR rearrangements. For each gene within a cluster, internal and external edge lengths were drawn from separate Γ distributions. The shape parameter for rates across sites (α) and gene length were also drawn from Γ distributions. For most simulations, a number of pruning operations were performed on the cluster tree. Because our method is aimed at complete prokaryotic genomes, this pruning method was intended to simulate the differences in taxon composition that might be expected in data sets composed of complete genome sequences, rather than those expected in eukaryotic data sets from expressed sequence tag or selective sequencing.

All simulations were analyzed with Conclustador using tree distributions inferred by the BI method described above. In addition, for the md1 simulation, Conclustador was applied to bootstrap distributions obtained from the distance, ML, and MP methods described above.

Assessing Phylogenetic Structure in Gene Clusters

To visualize the phylogenetic information in groups identified by Conclustador, supernetworks were constructed for each cluster using data available from single-gene phylogenetic analyses performed by BI. For each gene within a cluster, a consensus tree topology was produced from the posterior distribution. In order to reduce the number of insignificant splits in the supernetworks, only splits with posterior probability of 0.5 or greater were retained. SplitsTree4 (Huson and Bryant 2006) was then used to construct a supernetwork from these gene trees by the Z-closure method (Huson et al. 2004), using default options. The reason for this choice (i.e., the use of networks rather than trees) is that we expect the gene trees that fall in a given cluster to share more phylogenetic properties with each other than with any other gene tree in the data set, yet gene trees within a cluster do not necessarily share a single underlying tree. We call these genes with significant evolutionary resemblances "evolutionary doppelgängers," a notion that represents more accurately the evolutionary history of prokaryotic genes, where the traditional notion of global congruence (identical trees) is too strict to identify local phylogenetic congruences (and overlaps) between gene trees.

Analysis of Core Prokaryotic Genes

In order to evaluate the performance of Conclustador with biological sequence data, we analyzed 114 alignments from the data set of Puigbò et al. (2009) in which at least 90% of the 100 prokaryotes in the data set were represented, described by Puigbò et al. as "nearly universal trees" or NUTs. We used Bayesian posterior distributions of gene trees constructed using PhyloBayes (Lartillot and Philippe 2004) with the C20 substitution model (Quang et al. 2008) and with automated burnin estimation. Genes were assigned to clusters using Conclustador with the spectral clustering method, estimating the number of clusters by the eigengap heuristic.

Congruence within individual clusters was then assessed by the ILD test (Farris et al. 1994) and two likelihood ratio tests (Huelsenbeck and Bull 1996; Waddell et al. 2000). Because of restrictions of the likelihood ratio tests, taxa not present in all markers were removed from the data set, leaving a total of 41 taxa. The ILD test was performed on both the complete data set and the 41 taxon data set using PAUP* (Swofford 2003) with 100 repartitioning iterations, saving a single most parsimonious tree at each inference step (i.e., multrees = no). Both likelihood ratio tests were performed using trees and likelihoods estimated by RAxML with the WAG model and the CAT approximation for rates across sites for tree inference and four-class discretized Γ-distributed rates for likelihood estimation. Significance of the Huelsenbeck–Bull test was assessed based on 100 parametric bootstrap replicates as well as 100 repartitioning replicates, as is used in the ILD test (Farris et al. 1994). For parametric bootstraps, alignments of the same length as real alignments were simulated using Seq-Gen (Rambaut and Grassly 1997) under the WAG + Γ model, using the shape parameter and tree inferred from the concatenated cluster data set. Significance of the Waddell test was assessed based on 1000 RELL (Kishino et al. 1990) nonparametric bootstrap replicates. Phylogenetic supernetworks for individual clusters and for the complete NUTs data set were then constructed as described above.

In addition, we analyzed the same data set of 114 genes with CONCATERPILLAR (Leigh et al. 2008) with an alpha level of

0.01 and the WAG substitution model with a four-class discretized Γ model for rates across sites. CONCATERPILLAR uses the likelihood ratio test for congruence developed by Huelsenbeck and Bull (1996) in a hierarchical framework to identify congruent subsets of genes. Finally, we produced cigarette plots (Bapteste et al. 2008) for the combined set of 114 NUTs, as well as for each cluster identified by Conclustador and for subsets identified by CONCATERPILLAR to look for potential issues of tree reconstruction artifacts and lack of phylogenetic signal in each class of congruent genes (see supplementary material 1, Supplementary Material online for a detailed description of this method and its results).

Functional Analysis of Clusters of Core Genes

Markers within each of the clusters identified by Conclustador were assigned to 15 functional COG/NOG categories (Tatusov et al. 1997). We tested whether some functional categories were overrepresented using a Fisher's exact test and a hypergeometric test. The *P* values were adjusted for multiple testing using a Bonferroni correction (Shaffer 1995).

Taxonomic Consistency in Clusters of Core Genes

We defined eight taxonomic categories (Archaebacteria, crenarchaeota, euryarchaeota, proteobacteria, β-proteobacteria, cyanobacteria, firmicutes, and planctomycetales). For each taxonomic category and each tree in each cluster defined by Conclustador, we computed a *P*-score describing the distribution of the taxa belonging to this category on the tree. A *P*-score of 1 indicates that all the members of the taxonomic category grouped together on the tree, whereas *P*-scores $>$ 1 indicate that members of the category are increasingly scattered and not monophyletic (Schliep et al. 2011). For each tree, we summed the *P*-scores and represented the distribution of this sum for trees of cluster 0 and cluster 1 by two histograms. We tested that these two distributions differed by a Wilcoxon rank-sum test.

Results and Discussion

Conclustador Effectively Identifies Congruent Clusters

We used Conclustador to identify congruent clusters with data simulated according to various scenarios. In each case, Conclustador was used with spectral clustering by estimating the number of clusters using the eigengap heuristic, as well as with both *k*-means and spectral clustering with the correct number of congruent sets specified. Figure 1 summarizes the success of Conclustador for all simulations, in terms of the proportion of error due to each falsely identified congruent pairs (analogous to "false negatives" in a statistical testing framework with congruence as the null hypothesis) and falsely identified incongruent pairs ("false positives"). Histograms showing the distribution of the number of estimated clusters for all simulations using the eigengap heuristic are shown in (supplementary fig. S2, (Supplementary Material online).

In simulations with different numbers of missing taxa (fig. 1*a*), the success of Conclustador decreased as the number of missing taxa increased. However, in general, the performance of either spectral or *k*-means clustering was quite good: with spectral clustering and an estimated number of clusters, even in the simulations with up to three branch deletions per gene, on average only 7% of truly congruent pairs were mistakenly assigned to different clusters (i.e., they were assigned to the same cluster with 93% accuracy), whereas a mean of only around 1% of truly incongruent pairs were assigned to the same cluster (i.e., they were assigned to different clusters with 99% accuracy).

Spectral clustering tended to outperform *k*-means more frequently when data were simulated with more branch deletions. This is likely because spectral clustering can recover clusters whose individual members are sometimes quite distant, providing they are close to at least some members of the correct cluster (for an example in 2D, see Ng et al. 2001). This phenomenon is likely to occur when gene alignments share few OTUs in common with some other alignments in the same cluster but more OTUs with alignments in another cluster. As an extreme example, consider a pair of genes that have always been inherited vertically, A and B, that overlap by fewer than four taxa (e.g., due to gene deletion in at the base of different major lineages) and a third gene from a different cluster, C, that shares more than four taxa with A but has been transferred laterally a number of times. With *k*-means, A and C would likely be assigned to the same cluster. With spectral clustering, genes that are close together have a much greater impact on the structure of clusters: The result is that provided there is some gene D (or series of genes) whose bipartition distance to both A and B is small, they will likely be assigned to the same cluster (i.e., A and B are both closer to some gene D than to C).

When the number of clusters was estimated using the eigengap heuristic, the error increased slightly (both in terms of the number of congruent pairs of markers falsely identified as incongruent and the incongruent pairs falsely identified as congruent). In the simulation with no branch deletions, the number of cases in which congruence was falsely identified appears to be substantially higher than when the number of clusters was known. This drop in performance can largely be attributed to a very small number of simulations in which the number of clusters was drastically underestimated (supplementary fig. S2a, Supplementary Material online) combined with near perfect performance for this set of simulations when the number of clusters was known. Similarly, when data were simulated along trees with up to three branch deletions, the eigengap heuristic drastically overestimated the number of clusters in one occasion.

Figure 1*b* shows the results of simulations with different mean SPR distances between underlying tree topologies. Alignments for these simulations were simulated along trees with up to two branch deletions. Similarly to results shown in figure 1*a*, the error associated with spectral clustering was less than for *k*-means for all simulations

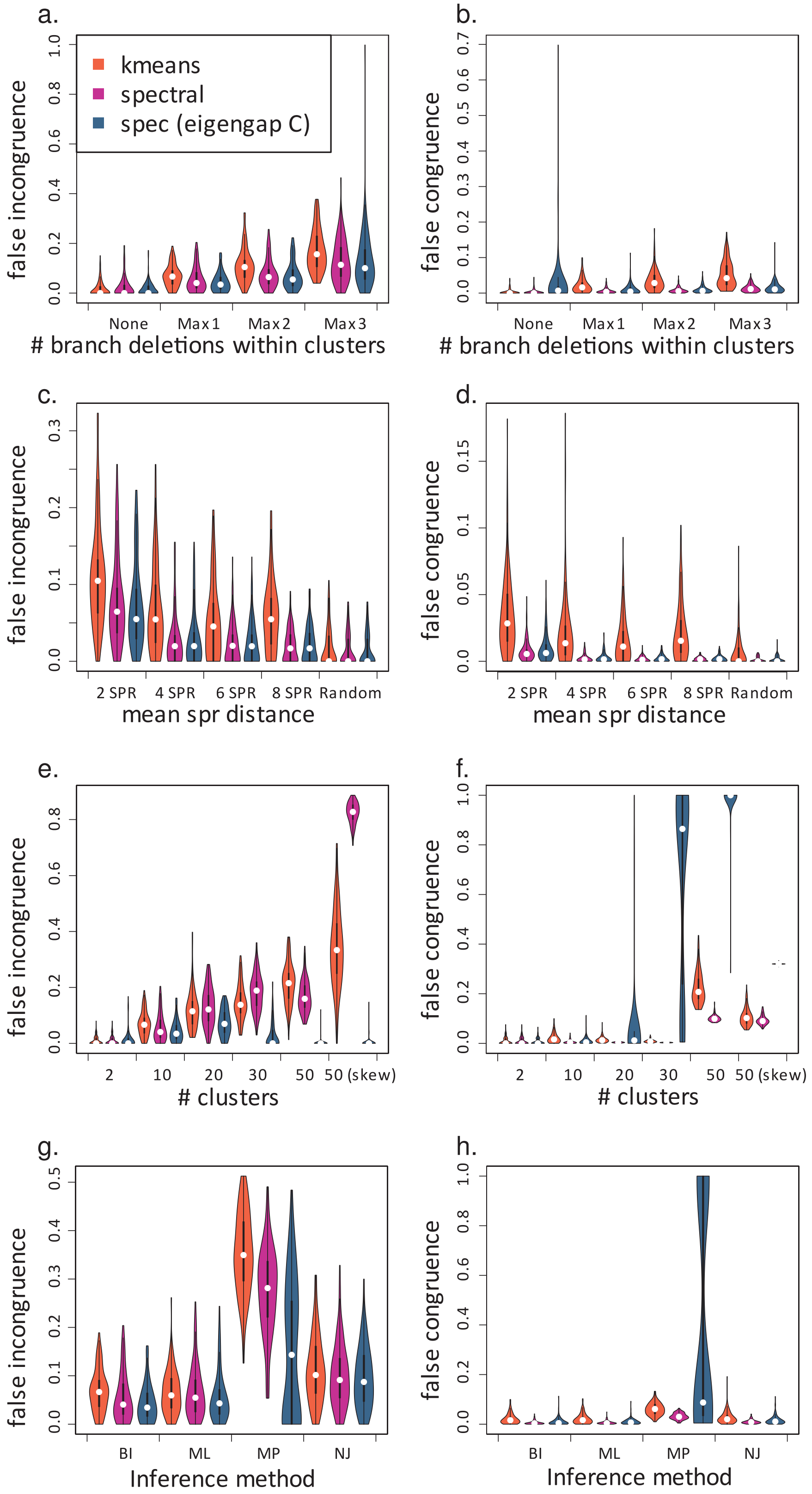

Fig. 1. Error in cluster assignment with simulated data. Data were simulated under a variety of conditions, subjected to phylogenetic analysis, and clustered using Conclustador. The error in cluster assignment was evaluated as the proportion of pairs of genes mistakenly assigned to the same or different clusters. Violin plots represent the result of 100 simulations, with the white circle indicating the mean value and the width of the violin indicating the density of values. *a*, *c*, *e*, and *g*, the vertical axis indicates proportion of congruent pairs of genes assigned to different clusters; *b*, *d*, *f*, and *h*, the vertical axis indicates proportion of incongruent pairs of genes assigned to the same cluster. *a* and *b*, different numbers of within-cluster branch deletions are indicated along the horizontal axis. *c* and *d*, mean SPR distance between underlying topologies of different clusters is indicated on the horizontal axis. *e* and *f*, number of underlying topologies ("true clusters") is indicated on the horizontal axis. *g* and *h*, inference method used to produce the tree distribution inferred from simulated data sets is indicated on the horizontal axis.

(in terms of the frequency with which both congruence and incongruence were falsely identified), whether the number of clusters was known or estimated. As the similarity between topologies underlying different sets of genes decreased (i.e., through an increased number of SPR operations), the performance of Conclustador improved, though even with a mean number of only two SPR events, spectral clustering misidentified incongruence in just over 5% of pairs (i.e., correctly identified congruent pairs with almost 95% accuracy), and almost never falsely identified congruence. Clearly, even when the clusters themselves are close in terms of underlying tree topology, Conclustador is able to distinguish clusters.

Clusters Do Not Represent Shared Tree-Like History When Number of True Trees Is Large

We also looked at the performance of Conclustador when alignments were simulated under different numbers of topologies (i.e., different numbers of clusters). Figure 1*c* shows the results of these simulations. For either two or ten true topologies, Conclustador performed very well, regardless of the clustering algorithm. Estimation of the number of clusters was also accurate (supplementary fig. S2c, Supplementary Material online).

When the data were simulated along 20 true topologies, the number of clusters tended to be somewhat underestimated by the eigengap heuristic, with modes around 15 and 18 clusters, although the mean of both errors remained less than 10% (congruent pairs were assigned to the same cluster and incongruent pairs to different clusters each with 93% accuracy). Resolving 20 clusters from only 100 data points is a difficult problem in principle; that Conclustador only slightly underestimates the number of clusters at this point is in itself impressive. As the number of clusters increased to 30 and 50, the underestimation of the number of clusters became more severe: With 50 true clusters, the number of clusters was nearly always estimated as one. For data sets with a high degree of incongruence, then, it is important to recognize that the identification of a single cluster does not indicate a single tree-like evolutionary history. We refer to genes included in such a cluster as "evolutionary doppelgängers": Genes that come to resemble one another by Conclustador's criteria but do not strictly share the same pattern of inheritance.

For a highly skewed distribution of genes among true tree topologies (data set 50tS, in which 51 genes were assigned to a single topology and the remaining 49 each evolved along a different topology), the eigengap heuristic nearly always identified two clusters. From the violin plot (fig. 1*c*), it is clear that these clusters were generally organized such that the 51 alignments simulated along the same topology fell within one cluster and the 49 simulated along different topologies fell within the other. Again, the 49 genes are evolutionary doppelgängers: clustering represents increased phylogenetic similarity relative to the entire data set but not a single tree-like history.

Obviously, there is no way to distinguish between a single true topology and genes that all evolved along different topologies (i.e., *n* true clusters for *n* genes) whatsoever when estimating the number of clusters. Furthermore, clustering methods tend to perform poorly when the number of true clusters is large relative to the number of data points. However, although the clustering methods used here failed to separate genes that evolved along different topologies as the number of clusters increased, there is an important difference between, for example, the simulations of the 50tS data set and the 2t data set, though in both cases, two clusters were usually identified. Figure 2 shows phylogenetic supernetworks estimated from typical simulations of the 50tS and 2t data set. In the case of the 50tS data set, one cluster's network (fig. 2*a*) is clearly more tree-like than the other's (fig. 2*b*); moreover, because the underlying topologies for the genes assigned to this cluster were related through only two SPR operations on average, there are many regions of the cluster phylogeny that remain tree-like, and network-like regions are evident in the splits graphs. In contrast, when there were only two true tree topologies underlying the data set, both cluster phylogenies looked generally tree-like (fig. 2*c* and 2*d*).

When a data set might contain a high level of incongruence, investigating whether clusters represent a shared tree-like history or significant phylogenetic similarity is then very important if tree-based phylogenetic inference is to be performed on clusters. Although we have examined the presence or absence of tree-like structure using splits graphs, other data set exploration methods have been developed for this purpose (e.g., Lento et al. 1995; White et al. 2007; Bapteste et al. 2008). In supplementary material 2 (Supplementary Material online), we show the application of the method of Bapteste et al. (2008), which uses heatmaps to determine whether apparent phylogenetic resolution is attributable to the presence of sites to which the model fits poorly.

Sets Identified in Prokaryotic Core Genes Reflect Different Frequencies of Lateral Gene Transfer

In addition to our studies with simulated sequence alignments, we used Conclustador to analyze a set of prokaryotic data published by Puigbò et al. (2009). One hundred prokaryote OTUs were represented in this data set of 114 gene alignments; in each alignment, at least 90% of OTUs were present (i.e., 90 sequences). Although Puigbò et al. concluded that this data set was congruent, such a result is inherently suspicious, given that frequent LGT is expected for prokaryotic data. Most interestingly, Conclustador recovered not one but two distinct clusters. We assessed phylogenetic congruence of the genes within each cluster using the ILD test (Farris et al. 1994) and likelihood ratio tests described Waddell et al. (2000) and Huelsenbeck and Bull (1996); congruence was rejected by these tests ($P < 0.001$ for the Waddell test and $P < 0.01$ for both the ILD and Huelsenbeck–Bull test). Based on this result, we chose to infer networks, rather than trees, to visualize phylogenetic relationships in each of the clusters. We used the individual gene majority rule

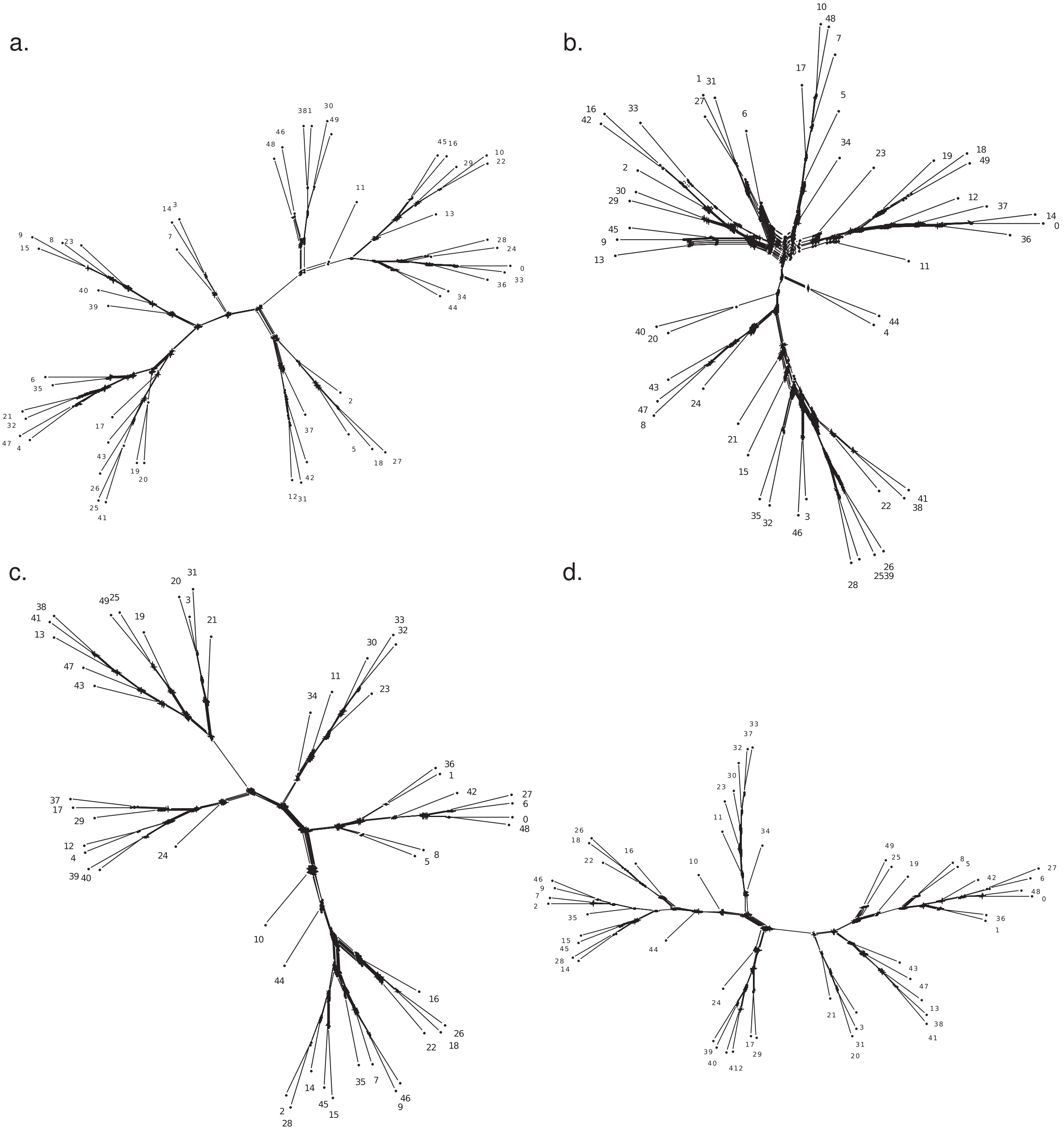

FIG. 2. Splits graphs from examples of simulations 50tS and 2t. Supernetworks were constructed for clusters inferred from simulations under either 50 topologies to which gene assignment was skewed (*a* and *b*) or 2 topologies (*c* and *d*). For both of these simulations, two clusters were inferred. However, the splits graph produced from one of the clusters from the 50tS simulation (*b*) is clearly less tree-like than the other supernetworks, indicating that there are multiple distinct trees underlying the genes of this cluster.

consensus trees (including only bipartitions with posterior probabilities ≥0.5) to construct a supernetwork for each cluster, shown in figure 3. The source of the incongruence is clear from these networks. In figure 3*a* (cluster 0, 47 genes), the network is far more reticulated than the network shown in figure 3*b* (cluster 1, 67 genes), with a number of short edges joining the eubacterial and archaebacterial domains. These networks indicate that Conclustador was able to effectively distinguish between genes in cluster 1, which have undergone substantial LGT, but only within domains, and those in cluster 0, which appear to have undergone a great deal more LGT events, including transfers between Eubacteria and Archaebacteria. Interestingly, the network inferred from the complete NUTs data set (supplementary fig. S3, Supplementary Material online) shows none of the interdomain reticulation found in cluster 0; the level of LGT in this data set is thus hidden until the two clusters are analyzed separately. Our analysis has thus uncovered additional aspects of congruence in this particular data set.

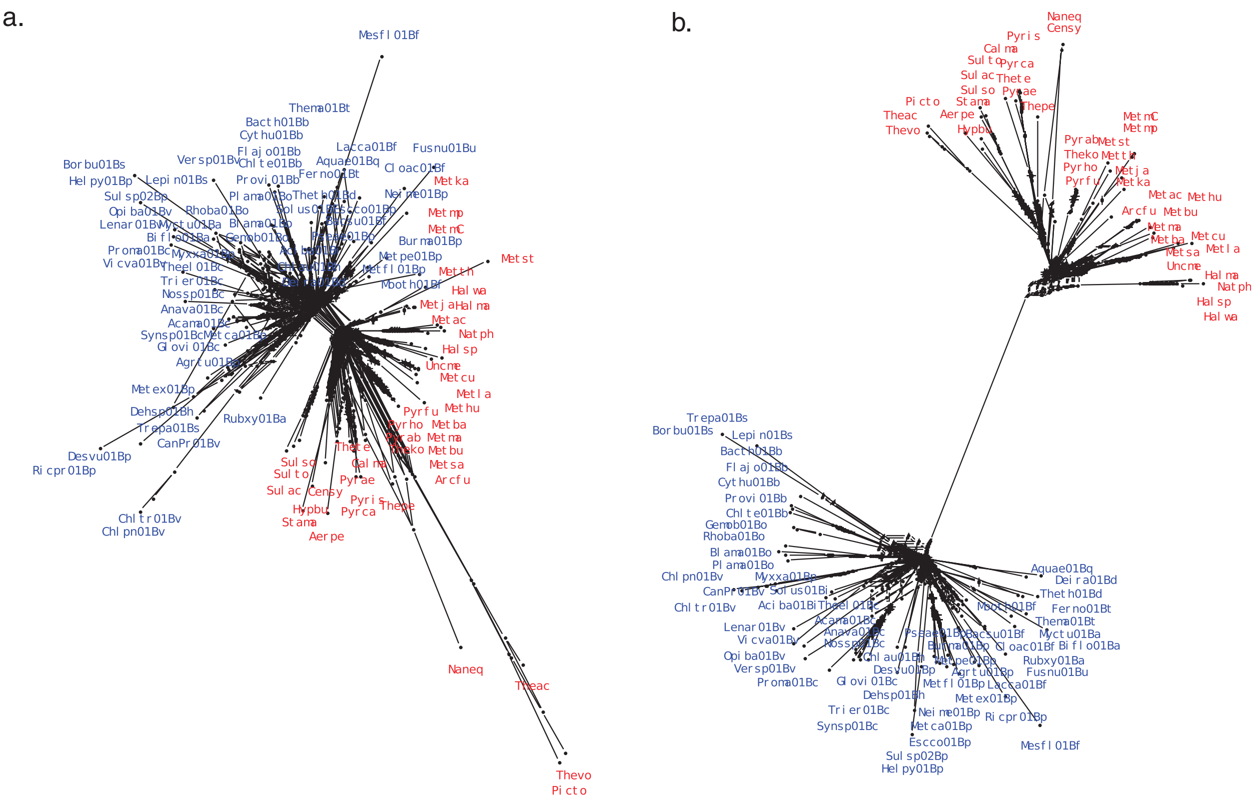

Fig. 3. Splits graphs inferred from clusters identified among NUTs. For the two clusters identified by Conclustador in the 114 NUTs of the data set of Puigbò et al. (2009), inferred supernetworks are shown here. Eubacterial taxa are shown in blue, whereas Archaebacteria are in red. For cluster 0 (*a*), the splits graph is much more highly reticulated than that of cluster 1 (*b*), though both clusters display a high level of non-tree-like evolution.

The two sets identified by Conclustador can be explained in two mutually exclusive ways: either they are mostly composed of genes with a common phylogenetic history (as was observed in the 2t simulation) or one or both of these clusters corresponds to a heterogeneous association of genes with distinct evolutionary stories (as in the 50tS simulation).

For the same data set, CONCATERPILLAR proposed 73 distinct clusters, most of which contained a single gene. Most of these singletons had been assigned to cluster 0 by Conclustador. The five largest sets of congruent genes identified by CONCATERPILLAR contained at least six genes, most of which corresponded to Conclustador's cluster 1 (supplementary table 2, Supplementary Material online). Because CONCATERPILLAR uses congruence as its null hypothesis, it is possible that the genes in this data set evolved along more than 73 distinct tree topologies but that there was insufficient phylogenetic signal to reject congruence in some cases (see Bapteste et al. 2008 for an example). With 73 or more true topologies underlying this data set, Conclustador would likely assign genes to the same cluster if the patterns of phylogenetic relationships they supported were similar (i.e., if they exhibit some local congruence), even if they did not evolve along the same tree. If CONCATERPILLAR groupings are meaningful, the two clusters identified by Conclustador may result from combining the most heterogeneous genes in cluster 0 while grouping the genes with the greatest local congruence into cluster 1.

Our empirical results indeed more closely resemble the 50tS situation than the 2t situation met in our simulated analyses: The two clusters recovered reflect different levels of congruence. Using splits graphs, we verified whether CONCATERPILLAR's five largest groups were truly congruent subsets rather than data sets with little phylogenetic signal. Splits graphs for all five sets (supplementary fig. S4, Supplementary Material online) displayed either very little resolution of relationships or displayed a pattern evolution that was not at all tree-like. Therefore, it appears that Concaterpillar did not identify shared vertical histories in its clusters but rather insufficient phylogenetic signal to reject the null hypothesis of congruence. Cigarette plots (see supplementary Material 2, Supplementary Material online) were consistent with this result.

Consequently, the recovery of two clusters by Conclustador (combined with the understanding that these clusters represent evolutionary doppelgängers) is a more honest result than the apparently meaningless "congruent" subsets recovered by CONCATERPILLAR, as there is evidence for multiple underlying trees for each of these clusters. The two Conclustador clusters are thus an important step in unraveling phylogenetic diversity within the NUTs.

Support for the Complexity Hypothesis: Informational Genes Are Less Frequently Transferred

In addition to revealing the diversity of evolutionary processes underlying the NUTs data set, the clusters identified by Conclustador are biologically meaningful. They are

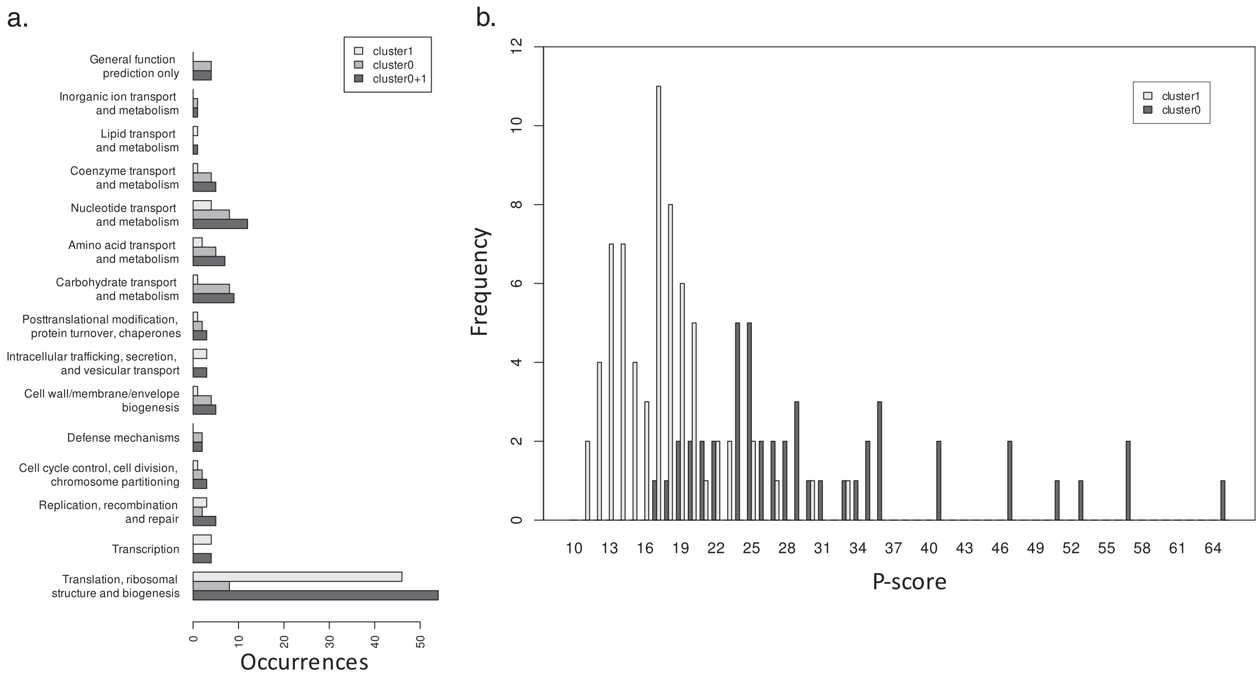

Fig. 4. Differences in functional classification and taxonomic consistency between clusters. (*a*) Distribution of functional classes of genes assigned to each cluster using Conclustador, as well as the combined data set of 114 NUTs. (*b*) Histogram of *P*-scores summed over all eight taxonomic groups for each cluster. Larger *P*-score sums indicate that members of described taxonomic groups tend not to branch together in inferred trees.

not random with respect to the functional classes of proteins encoded by member genes (fig. 4). "Informational" genes involved in translation and ribosomal structure and biogenesis are significantly overrepresented in cluster 1 with respect to cluster 0 (Fisher exact test $P = 4.8 \times 10^{-8}$; hypergeometric $P = 3.4 \times 10^{-8}$). This overrepresentation of informational genes among the less-frequently transferred genes offers a remarkable independent confirmation of the complexity hypothesis (Jain et al. 1999; Wellner et al. 2007), which suggested that informational genes, typically members of large complex systems, are less prone to LGT than other ("operational") genes.

The complexity hypothesis is often invoked to argue that the phylogeny of informational genes provides a valuable backbone to evaluate the vertical component of microbial evolution. In that regard, Conclustador's cluster 1 defines a set of gene trees that more closely reflects a vertical phylogenetic signal within the NUTs. These gene families have certainly been subjected to less frequent LGT, and as a consequence informational genes have remained in given genomes longer than others, but many other gene families have moved around at a significantly higher rate and have followed a different phylogenetic history from those in cluster 1.

Around a tiny core of translation, ribosomal structure and biogenesis genes, less affected by LGT, virtually all other genes in cells have been more frequently transferred over evolutionary time. Conclustador showed not only that there is no large set of mostly vertical (i.e., congruent in the classical sense) genes but also that the history of prokaryotic core genes appears increasingly fluid. Therefore, the somewhat coherent phylogenetic signal of cluster 1, even if it is useful for inferences about the most stable features of prokaryotic phylogenesis, should still not be conflated with the rich and more complex phylogenetic history of the organisms and that of species. Typically, genes from many additional functional categories beyond those found in cluster 1 (and even beyond the NUTs data set) are required to make an organism. Genes encoding defense mechanisms, inorganic ion transport, and metabolism and "general" functions are absent from cluster 1, whereas those encoding transcription, intracellular trafficking, secretion and vesicular transport, and lipid transport and metabolism functions are completely absent from cluster 0. Energy production and conversion, cell motility, secondary metabolites biosynthesis, transport and catabolism, cytoskeleton, and signal transduction mechanisms are absent in the entire NUTs data set.

Computation Time and Effectiveness of Phylogenetic Shortcuts

Results presented above were all produced using Conclustador with Bayesian posterior gene tree distributions estimated with PhyloBayes (Lartillot and Philippe 2004). However, Conclustador can also be used to assess phylogenetic congruence using bootstrap distributions. We compared the performance of Conclustador using BI with the md1 data set with bootstrap distributions from ML, MP, and NJ (fig. 1*d*). Performance was similar with ML and BI and also reasonably good with NJ, although the number of clusters was somewhat less accurately estimated (supplementary fig. S2d, Supplementary Material online). With MP-based bootstrap distributions, however, performance of Conclustador was substantially worse, particularly when the number of clusters was not known. In fact, the number of clusters was most frequently estimated as one, suggesting that clusters are less distinct than with other methods. Clearly, if the tree distributions analyzed with Conclustador are sufficiently bad, the result will suffer.

One advantage of using bootstrap distributions rather than posterior distributions is that bootstrap analysis can trivially be parallelized. In addition, NJ is considerably faster than BI, and its performance was still relatively good. This result is in agreement with the result of Deusch et al. (2008): for large-scale analyses in which the phylogenetic trees themselves are not the desired result, NJ is a sufficiently good reconstruction method and is considerably faster than ML or BI. This method might be a good choice for phylogenomic analyses with more limited computational resources. However, even phylogenetic analysis by BI can trivially be parallelized insofar as each gene analysis can be performed independently.

Conclustador itself does not run in parallel, and its computational complexity is theoretically $O(n^3)$ for the eigenvector decomposition step, assuming spectral clustering is used, whereas k-means is an NP-hard problem (though in practice run time scales reasonably well with the Death of Dodos algorithm and other heuristics). However, in practice, the time required to analyze even a fairly large number of genes is relatively short: Analysis of a data set of 1,100 posterior distributions from simulated alignments with 50 taxa took only an hour with an Intel Xeon 2.66 GHz CPU. The analysis of the 114 alignments of the NUTs data set (Puigbò et al. 2009) took 33 min on the same computer. In contrast, analysis of this same data set with CONCATERPILLAR (Leigh et al. 2008) took 13 days running in parallel over 20 Intel Xeon 3.0 GHz CPUs. The time initially required to perform the phylogenetic analysis of the 114 alignments should be added to Conclustador's analysis time: With PhyloBayes, this took around 9 h on 20 CPUs.

Conclusions

The results of our analyses of simulated data have shown that Conclustador is able to discriminate between genes evolving according to different tree topologies, even when those topologies share large areas of local congruence. Identification of sets of genes that share only local regions of congruence is useful in phylogenomics because phylogenetic analysis of these combined markers allows resolution of shared relationships that individual genes might not contain sufficient data to reveal, even if network-based methods must be used. This is an advantage that Conclustador holds over CONCATERPILLAR, which is unable to identify genes that share only local congruence. In the case where all or nearly all genes evolved along different trees, "congruent" subsets identified by CONCATERPILLAR might reflect only insufficient signal to reject the null hypothesis. Differences in OTU composition of the individual alignments affects the success with which Conclustador recovers phylogenetically congruent clusters, but performance was good even with substantially different taxon representation. However, as performance declined when missing taxa increased, Conclustador might perform poorly for investigating eukaryotic data sets from incomplete genomes.

Yet, when applied to prokaryotic core genes from complete genomes, Conclustador recovered clusters that did not reflect distinct, strictly vertical signals, but instead separated the data into clusters that contained genes with different rates of conflicting signal. Genes belonging to cluster 0 show much more frequent LGT than genes in cluster 1, and distribution of functions between clusters was also contrasted. Although care must be taken not to immediately interpret clusters as congruent subsets of genes, these clusters are clearly biologically meaningful. Future uses of Conclustador should unravel even more of the evolutionary processes at play in complete prokaryotic genomes.

Supplementary Material

Supplementary figures S1–S7, tables S1 and S2, and material S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank François-Joseph Lapointe for helpful discussions regarding the direction of this work. We would also like to thank Andrew Roger and the Centre for Evolutionary Genomics and Bioinformatics (CGEB) for generously giving us access to computational resources and Thom Leigh for suggesting the excellent name "Conclustador." We also thank two anonymous referees for thoughtful comments that greatly improved the quality of this work. J.W.L. was supported by a postdoctoral fellowship from Université Pierre et Marie Curie and is currently supported by a postdoctoral fellowship from the Natural Science and Engineering Research Council of Canada.

References

Bapteste E, O'Malley MA, Beiko RG, et al. (11 co-authors). 2009. Prokaryotic evolution and the tree of life are two different things. *Biol Direct*. 4:34.

Bapteste E, Susko E, Leigh J, Ruiz-Trillo I, Bucknam J, Doolittle WF. 2008. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol Biol Evol*. 25:83–91.

Beiko RG, Keith JM, Harlow TJ, Ragan MA. 2006. Searching for convergence in phylogenetic Markov chain Monte Carlo. *Syst Biol*. 55:553–565.

Bordewich M, Semple C. 2004. On the computational complexity of the rooted subtree prune and regraft distance. *Ann Combin*. 8:409–423.

Brochier C, Bapteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet*. 18:1–5.

Bryant D, MacKenzie A, Steel M. 2003. The size of maximum agreement subtree for random binary trees. In: Janowitz M, Lapointe F-J, McMorris FR, Mirkin B, Roberts FS, editors.

Bioconsensus. Vol. 61. DIMACS series in discrete mathematics and theoretical computer science. Providence (RI): American Mathematical Society. p. 55–66.

Caliński T, Harabasz J. 1974. A dendrite method for cluster analysis. *Commun Stat Simul Comp*. 3:1–27.

Campbell V, Legendre P, Lapointe FJ. 2009. Assessing congruence among ultrametric distance matrices. *J Classif*. 26:103–117.

Campbell V, Legendre P, Lapointe FJ. 2011. The performance of the Congruence Among Distance Matrices (CADM) test in phylogenetic analysis. *BMC Evol Biol*. 11:64.

de Vienne DM, Giraud T, Martin OC. 2007. A congruence index for testing topological similarity between trees. *Bioinformatics* 23:3119–3124.

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol*. 25:748–761.

Edwards J, Oman P. 2003. Dimensional reduction for data mapping. *R News*. 3:2–7.

Estabrook GF, McMorris FR, Meacham CA. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst Zool*. 34:193–200.

Farris JS, Källersjö M, Kluge AG, Bult C. 1994. Testing significance of incongruence. *Cladistics* 10:315–319.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 14:685–695.

Gribaldo S, Philippe H. 2002. Ancient phylogenetic relationships. *Theor Popul Biol*. 61:391–408.

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res*. 21:349–356.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–217.

Huelsenbeck JP, Bull JJ. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Syst Biol*. 45:92–98.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 23:254–267.

Huson DH, Dezulian T, Klopper T, Steel MA. 2004. Phylogenetic super-networks from partial trees. *IEEE Trans Comput Biol Bioinform*. 1:151–158.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*. 96:3801–3806.

Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol*. 30:151–160.

Koblmüller S, Duftner N, Sefc KM, Aibara M, Stipacek M, Blanc M, Egger B, Sturmbauer C. 2007. Reticulate phylogeny of gastropod-shell-breeding cichlids from Lake Tanganyika—the result of repeated introgressive hybridization. *BMC Evol Biol*. 7:7.

Křivánek M. 1986. Computing the nearest neighbor interchange metric for unlabeled binary trees is NP-complete. *J Classif*. 3:55–60.

Lapointe FJ, Rissler LJ. 2005. Consensus, congruence, and the comparative phylogeography of codistributed species in California. *Am Nat*. 166:290–299.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21:1095–1109.

Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing congruence in phylogenomic analysis. *Syst Biol*. 57:104–115.

Lento GM, Hickson RE, Chambers GK, Penny D. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Mol Biol Evol*. 12:28–52.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. p. 281–297.

Matte-Tailliez O, Brochier C, Forterre P, Philippe H. 2002. Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol*. 19:631–639.

McBreen K, Lockhart PJ. 2006. Reconstructing reticulate evolutionary histories of plants. *Trends Plant Sci*. 11:398–404.

Newman ME. 2006. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev*. 74:36104.

Ng AY, Jordan MI, Weiss Y. 2001. On spectral clustering: analysis and an algorithm. In: Dietterich TG, Becker S, Ghahramani Z, editors. Advances in neural information processing systems. Vol. 14. Cambridge (MA): MIT Press p. 849–856.

Nye TMW. 2008. Trees of trees: an approach to comparing multiple alternative phylogenies. *Syst Biol*. 57:785–794.

Penny D, Hendy MD, Henderson IM. 1987. Reliability of evolutionary trees. Cold Spring Harbor symposia on quantitative biology. Vol. 52. p. 857–862.

Planet PJ, Sarkar IN. 2005. mILD: a tool for constructing and analyzing matrices of pairwise phylogenetic character incongruence tests. *Bioinformatics* 21:4423–4424.

Puigbò P, Wolf YI, Koonin EV. 2009. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol*. 8:59.

Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 2317–2323.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci*. 13:235–238.

Roberts DL, Solow AR. 2003. Flightless birds: when did the dodo become extinct? *Nature* 426:245.

Robinson D, Foulds L. 1981. Comparison of phylogenetic trees. *Math Biosci*. 53:131–147.

Sammon JW. 1969. A nonlinear mapping for data structure analysis. *IEEE Trans Comput*. 18:401–409.

Schliep K, Lopez P, Lapointe F, Bapteste E. 2011. Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Mol Biol Evol*. 28:1393–1405.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.

Shaffer JP. 1995. Multiple hypothesis testing. *Annu Rev Psychol*. 46:561–584.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol*. 57:758–771.

Steel MA, Penny D. 1993. Distributions of tree comparison metrics: some new results. *Syst Biol*. 42:126–141.

Swofford DL. 2003. PAUP* 4.0 b10. Phylogenetic analysis using parsimony (* and other methods). Sunderland (MA): Sinauer Associates.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–637.

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Ser B Biol Sci*. 63:411–423.

Vinh le S, Von Haeseler A. 2004. IQPNNI: moving fast through tree space and stopping in time. *Mol Biol Evol*. 21:1565–1571.

Von Luxburg U. 2007. A tutorial on spectral clustering. *Stat Comput*. 17:395–416.

Waddell PJ, Kishino H, Ota R. 2000. Rapid evaluation of the phylogenetic congruence of sequence data using likelihood ratio tests. *Mol Biol Evol*. 17:1988–1992.

Waddell PJ, Kishino H, Ota R. 2007. Phylogenetic methodology for detecting protein interactions. *Mol Biol Evol*. 24:650–659.

Wellner A, Lurie MN, Gophna U. 2007. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biol*. 8:R156.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*. 18:691–699.

White WT, Hills SF, Gaddam R, Holland BR, Penny D. 2007. Treeness triangles: visualizing the loss of phylogenetic signal. *Mol Biol Evol*. 24:2029–2039.

Wu Y. 2009. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics* 25:190–196.

Zelnik-Manor L, Perona P. 2004. Self-tuning spectral clustering. In: Saul LK, Weiss Y, Bottou L, editors. Advances in neural information processing systems. Vol. 17. Cambridge (MA): MIT Press. p. 1601–1608