Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis

Damien M. de Vienne,1,2 Sébastien Ollier,2 and Gabriela Aguileta1,2

1Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain
2Laboratoire Ecologie, Systématique et Evolution, UMR 8079, Université Paris-Sud 11, Orsay cedex, France

Abstract

Full genome data sets are currently being explored on a regular basis to infer phylogenetic trees, but there are often discordances among the trees produced by different genes. An important goal in phylogenomics is to identify which individual gene and species produce the same phylogenetic tree and are thus likely to share the same evolutionary history. On the other hand, it is also essential to identify which genes and species produce discordant topologies and therefore evolve in a different way or represent noise in the data. The latter are outlier genes or species and they can provide a wealth of information on potentially interesting biological processes, such as incomplete lineage sorting, hybridization, and horizontal gene transfers. Here, we propose a new method to explore the genomic tree space and detect outlier genes and species based on multiple co-inertia analysis (MCOA), which efficiently captures and compares the similarities in the phylogenetic topologies produced by individual genes. Our method allows the rapid identification of outlier genes and species by extracting the similarities and discrepancies, in terms of the pairwise distances, between all the species in all the trees, simultaneously. This is achieved by using MCOA, which finds successive decomposition axes from individual ordinations (i.e., derived from distance matrices) that maximize a covariance function. The method is freely available as a set of R functions. The source code and tutorial can be found online at http://phylomcoa.cgenomics.org.

Key words: phylogenomics, outlier genes, phylogenetic markers, multivariate analysis, fungal genomes.

Introduction

The analysis of genomic data sets has generated the pressing need to compare a large amount of phylogenetic trees. The trees under comparison may, for example, represent a set of shared orthologous genes from a given organism or taxonomic group, sampled across their full genomes. The goal of such comparisons can be 2-fold: either to identify the set of genes producing a concordant topology that may reflect the species history or to identify the set of outlier genes that generate topological incongruities and are likely to evolve differently or represent noise. While most efforts have been devoted to detecting genes that produce concordant gene trees, less attention has been given to identifying outliers, although they are very important candidates for studying interesting biological processes.

Detecting outlier genes and species is particularly relevant in phylogenomics because genes whose tree topologies are discordant with respect to an overall topological consensus are good candidates for evolving at significantly different rates, to be under positive selection for functional divergence, to have been horizontally transferred, to be paralogous instead of orthologous genes, or to be randomly sampled from incompletely sorted lineages. On the other hand, species that appear as outliers (i.e., that are placed very differently in different gene trees) may represent species that are more likely to experience gene transfers, having sequencing errors, having undergone bottlenecks, or other biological processes at the species level. Even though topological discordances can be detected at the gene or the species level, most methods so far are designed to look at genes only.

In the context of phylogenetic gene tree comparison and phylogenomics, several approaches and methods have been proposed to detect concordant and discordant topologies trying to integrate as much as possible the available information. A solution that attempts to deal with all the available data is the construction of a super matrix from the alignment of all or a large number of the available individual genes, an approach known as “total evidence” (Gadagkar et al. 2005; Aguileta et al. 2008). However, most methods involve some sort of consensus solution either by classical consensus of multiple competing phylogenies (Felsenstein 1985), by building a super tree (Gordon 1986; Sanderson et al. 1998), by Bayesian sampling (Ane et al. 2007), or a summary by agreement subtrees (Cranston and Rannala 2007). Other approaches work by trying to filter out the noise and/or increasing the phylogenetic signal (Rodriguez-Ezpeleta et al. 2007; Roure et al. 2007) or by clustering congruent loci (Leigh et al. 2008, 2011). Another possibility involves identifying the conflicting bipartitions and representing all the possible alternatives in a graphic way, an approach that results in networks.
of genes (Huson and Bryant 2006), or a projection of the conflicting signals in 2Ds (White et al. 2007). Also, in order to obtain a single tree from multiple individual trees of the same organism or species, the coalescent approach can be used (Rosenberg 2002; Suchard et al. 2003; Suchard 2005; Degnan and Rosenberg 2006) or the Bayesian reconstruction of gene trees taking the species tree as a prior (BEST) (Edwards et al. 2007). Finally, approaches for graphically comparing phylogenetic trees include the use of multidimensional scaling to visualize tree-to-tree pairwise distances (Hillis et al. 2005); the use of principal component analysis (PCA) to detect events of horizontal gene transfer (Brochier et al. 2002; Susko et al. 2006); and employing Hadamard spectral analysis represented in a triangular graph called Treeness Triangles to investigate the loss of phylogenetic signal (White et al. 2007). However, by concatenating the multilocus data or by summarizing or obtaining a consensus of their individual gene trees, most of the latter methods lose a wealth of potentially interesting information, especially by removing outlier data or trees. Phylo-MCOA is unique in that it analyzes all the trees simultaneously without discarding data a priori. Furthermore, as far as we know, the possibility to also detect outlier species when comparing phylogenetic trees has not been proposed before.

Here, we present Phylo-MCOA, a fast method for comparing multiple phylogenetic trees and detecting outlier genes and species based on multiple co-inertia analysis (MCOA) (Chessel and Hanafi 1996). The method allows the efficient identification of concordant and outlier genes and species, by enabling the quick comparison of a large number of trees. It builds a consensus topology by extracting the similarities, in terms of the pairwise distances between all the species in all the trees, simultaneously. The consensus topology is a 2D graphic representation of the distance interrelationships between species supported by the majority of the genes. Phylo-MCOA finds successive decomposition axes from individual ordinations (derived from matrices of distances between species) that maximize a covariance function in order to build the consensus topology, and it estimates the contribution of each individual gene tree to this topology (i.e., to what extent each gene deviates or agrees with what the majority of genes support). MCOA is expected to reveal the common features (i.e., pairwise distances and species placements) of single-tree topologies, build a consensus topology and compare individual trees with this reference. Until now, MCOA has been used with ecological (Bady et al. 2004; Hedde et al. 2005) and genetic data (Laloe et al. 2007; Berthouly et al. 2008); however, as far as we know, this is the first time that MCOA is applied to phylogenetic analysis. Phylo-MCOA is unique in tracing the position of each species with respect to all other species across the compared phylogenies. The simultaneous analysis of all the species placements in all trees makes it possible to detect, not only concordant genes but also outlier gene trees, and even more importantly, it can show which species and genes explain those differences. Our method is thus new in allowing the analysis of the data at the species level for comparing the histories of different species across individual gene trees.

Phylo-MCOA can be very useful for phylogenomic analyses, where big genomic data sets are typically analyzed. First, Phylo-MCOA readily identifies the outlier genes and species that cause topological incongruence. This is useful for studying the evolution of specific genes, particularly those that seem to differ from the rest of the genome. The method will also identify candidate outlier species, which may undergo interesting biological processes at the species level. Phylo-MCOA is unique in pointing not only to outlier genes but also to outlier species. Second, when the goal is to construct a species tree or a tree that reflects the evolution of most genes in the genome, Phylo-MCOA can be used to identify which genes produce discordant trees and remove them for subsequent phylogenetic analysis. In this way, only the genes that share the same evolutionary history are kept in order to build a tree that reflects the evolution of the species or that of most of the genome. Third, Phylo-MCOA can be used to investigate specific evolutionary hypotheses including, but not limited to, which genes may have been laterally transferred, which ones may be more readily lost or replaced, or which species produce phylogenetic discordance.

Materials and Methods

MCOA in the Context of Phylogenetic Analysis

In phylogenomics, there is often the need to compare sets of phylogenetic trees, determine how the position of each species changes in every tree, and to find out which trees are more similar between each other in terms of the pairwise topological distance (i.e., either nodal or patristic). It would thus be advantageous to compare the position of each species in all trees at the same time, to know if there is a correlation among the species positions and distances represented in all trees (i.e., create a consensus topology that serves as reference), and to what extent each individual tree deviates from such a consensus. The problem at hand involves measuring the relationship between the position of each species and the distances between them in all the trees of a given set.

In the context of multivariate data analysis, one is interested in ordering objects (e.g., variables, sites, and tables) by their similarity, and the relationships between the objects are measured by different metrics (Manly 2004). Generally, objects are projected onto principal axes and are thus characterized numerically or graphically. This procedure places similar objects near each other and dissimilar ones far apart in geometric space. Traditional ordering techniques (e.g., PCA) typically use a singular decomposition axis value of one table. In the case of our proposed method, MCOA allows the simultaneous ordination of many Euclidean distance matrices derived from every tree in the set by finding successive decomposition axes from each distance matrix that maximize a covariance function. Doing this allows the extraction of the consensus topology contributed by all trees.
Here, we introduce the basic concepts of the method; interested readers are referred to the appendix in the supplementary material (Supplementary Material online) and to Chessel and Hanafi (1996) for its mathematical description.

To better understand how MCOA is applied to phylogenetic data (i.e., a set of K trees), we can decompose the method in two steps. We first explain how one tree can be represented in a 2D space through Principal Coordinate Analysis (Gower 1984; Gower and Legendre 1986) (PCO, fig. 1a), and we then explain how K trees can be simultaneously analyzed through MCOA (fig. 1b).

A single tree can be transformed into a distance matrix by calculating the pairwise distances between species in the trees (fig. 1a, step 1). Distance between species can either be nodal (number of nodes separating two leaves in the tree) or patristic (sum of the branch lengths separating two leaves). We then compute the square root of this distance matrix in order to get a Euclidean distance matrix, which is an optimal condition for the PCO to be performed. Whereas the methods typically used to obtain Euclidean distances impose a geometric distortion on them, taking their square root does not affect the results (de Vienne et al. 2011). The principal coordinates of the leaves (i.e., terminal branches) in a Euclidean space are then derived from the Euclidean distance matrix by a PCO analysis (fig. 1a, step 2). Each axis defined by the PCO analysis maximizes the variance of the coordinates of the species. When original distances between species are Euclidean, all the eigenvalues are positive and correspond to the variance associated to the axis. In that case, the distances between species in the new space defined by the PCO analysis are the same as the original distances defined by the tree. Because most of the variance is generally concentrated in the first axis (depending on the topology of the tree), we can represent the respective position of the leaves in
Detection of complete outliers by Phylo-MCOA. (a) The 2WR matrix is computed by calculating, for every species, the distance separating its position in each gene tree to its reference position (see text). The 2WR matrix provides information about species in rows and genes in columns. (b) The 2WR matrix is transformed into two binary matrices (2WR^{Bgn} and 2WR^{Bsp}) by detecting outlier cells as described in the Methods section. (c) A score is computed for each gene from the 2WR^{Bgn} matrices and for each species from the 2WR^{Bsp} matrix and represented by a bar plot. The dashed gray line on the bar plot represents the threshold chosen by the user over which a gene or a species is considered as a complete outlier.

Phylo-MCOA: Using MCOA for Phylogenetic Analysis

Following the analysis described above, the reference position of each species (coordinates of the species projected on the \(v^1, v^2, \ldots v^K\) vectors) as well as the position of the species in each individual tree (coordinates of the species projected on the \(u^1, u^2, \ldots, u^N\) auxiliary vectors) can be computed. It is possible to plot the reference positions of each species, as well as their position in each individual tree, using the two first axes of the MCOA analysis (fig. 1b part 4). This plot is usually called the cohesion plot. The set of reference positions of the species in the cohesion plot is called consensus typology. For a given species, the distance between its reference position and its position in each individual tree is represented by lines. Long lines represent cases where the position of the species has a position in a given tree that is not concordant with its position in all other trees (and therefore neither in the consensus typology). On the contrary, short lines represent cases where the position of the species in a given tree is roughly the same as in all the other trees (and therefore in the consensus typology). The cohesion plot thus contains \(N \times T\) dots, where \(N\) represents the number of species in each tree and \(T\) the number of trees.

The cohesion plot is very informative for small data sets, where one can assess the position of each species in each individual tree relative to its reference position. However, this plot becomes difficult to read and to interpret as the number of genes and species increases because species labels superimpose and the cloud of points and lines becomes overcrowded. Moreover, the cohesion plot gives only a graphical representation of the data and does not allow the extraction of statistically significant outlier genes and species. Therefore, we show instead the information contained in cohesion plots but using a different and more readable representation: the two-Way Reference matrix (hereafter referred to as the 2WR matrix). The 2WR matrix is built by computing, for every species, the distance between its reference position and its position in each individual tree in the multidimensional space. The rows in the 2WR matrix represent species and columns represent genes (fig. 2a). Note that the order of species (rows) in the 2WR matrix follows the average relative position of species in the gene trees. In other words, two species that are close
in the 2WR matrix (two contiguous rows) are two species that are closely related in the gene trees.

Detection of Outliers

Phylo-MCOA can detect two types of outliers: “complete” outliers and “cell-by-cell” outliers. Complete outliers are of two types: complete outlier species and complete outlier genes. Complete outlier genes place all the species in a position different from their positions in the other gene trees, whereas complete outlier species are species whose position with respect to other species is different in all the gene trees. Cell-by-cell outliers represent species whose position in some specific gene trees is not concordant with their position in the other gene trees (i.e., single cells in the 2WR matrix whose value seems outlier with respect to the other values in the matrix).

Cell-by-cell outliers cannot be detected by Phylo-MCOA from the 2WR matrix if complete outliers are present because it hides the cell-by-cell signal. The detection of outliers by Phylo-MCOA is thus performed in multiple steps: 1) Phylo-MCOA analysis of the original data set; 2) detection of complete outliers; 3) removal of the complete outliers from the original data set; 4) Phylo-MCOA analysis of the new data set; and 5) detection of cell-by-cell outliers. If no complete outliers are detected, steps 3 and 4 are omitted.

For detecting complete outliers (fig. 2), the 2WR matrix is separated in two matrices, (2WR\textsuperscript{Bsp} and 2WR\textsuperscript{Bgn}, fig. 2b) by detecting outlier cells. For constructing the 2WR\textsuperscript{Bsp} matrix, a cell is considered as an outlier (and assigned the value 1), if its value in the 2WR matrix is higher than Q3 + kIQR, where Q3 represents the third quartile of the distribution of values of its column and IQR represents the interquartile range of these values. Symmetrically, for the 2WR\textsuperscript{Bgn} matrix, a cell is considered as outlier if its value in the 2WR matrix is higher than Q3 + kIQR, where Q3 and IQR are now computed based on the values in the row. We set $k = 1.5$ as the default value as is classically done when detecting outliers with this method. The two obtained binary matrices are represented in figure 2b. From the 2WR\textsuperscript{Bgn} matrix, a score is computed for each gene by computing the proportion of 1s in its column. From the 2WR\textsuperscript{Bsp} matrix, a score is computed for each species by computing the proportion of 1s in its row. This leads to the two barplots presented in figure 2c. In this example, gene 2 and gene 13 are detected as outliers by more than 90% of the species, and 116 and 133 are detected as outlier species by more than 90% of the genes.

The user can fix a threshold for considering that a species or a gene is a complete outlier. By default, we set this threshold to 50% (i.e., species detected as outlier for more than 50% of the genes are seen as complete outliers and genes detected as outliers for more than 50% of the species are considered as complete outliers).

For the detection of cell-by-cell outliers (fig. 3), the 2WR matrix (fig. 3a) is transformed into the 2WR\textsuperscript{Bgn} and 2WR\textsuperscript{Bsp} matrices by normalizing the 2WR matrix by rows and columns, respectively; the product of these two normalized matrices gives the 2WR\textsuperscript{Bmsp} matrix that is used for the cell-by-cell outlier detection.

The 2WR\textsuperscript{Bmsp} matrix is first transformed into a binary matrix by detecting outlier cells, according to the values in the rows, and to a second binary matrix by detecting outlier cells according to the values in the columns. These two binary matrices are finally multiplied in order to conserve only cells detected as outliers for both rows and columns. This last binary matrix is called 2WR\textsuperscript{Bgmsp} (fig. 3b).

This procedure for detecting cell-by-cell outliers is, however, too permissive, and therefore, a second filtering step is necessary. As explained before, Phylo-MCOA starts by transforming trees into distance matrices. Species in the trees are not independent from each other. Therefore, even a single modification in a tree will change all the values in the associated distance matrix. When we detect a cell-by-cell outlier, we have to distinguish between “true” outliers and species that are affected by this outlier just because...
they are close in the distance matrix. To solve this problem, we take advantage of a Phylo-MCOA property that orders species in the 2WR matrix by their mean proximity in the cohesion plot. When two or more species that are contiguous in the 2WR\_gnsp matrix for a given gene are detected as outliers, we define this group as an “island” (dashed circles in fig. 3b). In each island, we need to distinguish between the true outlier species and the contiguous species that are erroneously detected as outliers. To do so, we consider that the outlier cell with the highest value in the 2WR\_gnsp is the true outlier, and the contiguous cells are artifacts. This filtering step allows obtaining a binary matrix free of islands that contains only “true” cell-by-cell outliers (fig. 3c).

Handling Tree and Node Support
When comparing phylogenetic trees, it is very important to consider the statistical support of both trees and nodes. To take into account measures of tree support such as likelihood scores, posterior probabilities, or any other suitable scores, Phylo-MCOA can assign different weights, specified by the user, to the different trees according to their support. To do so, during the MCOA analysis, each table (representing Euclidean spaces, see fig. 1) is multiplied by the square root of the weight that is associated to it (in absolute values). As a result, the values in the tables corresponding to a specific gene with a high weight are higher and the importance of the corresponding auxiliary vectors for retrieving the \( \nu \) vectors also increases (fig. 1b, step 3). In this sense, those genes will weigh more in the MCOA. It is also possible to take into account node support (e.g., bootstrap values) when building the consensus typology, by choosing a cutoff level of node confidence below which nodes will be collapsed. In this way, poorly supported nodes (i.e., unlikely species bipartitions) will have less or no influence in the consensus typology.

Handling Missing Data
Phylo-MCOA is capable of analyzing data sets that are incompletely sampled (i.e., with missing data), in the cases where a given gene is not available for one or more species but is found in all other species. To handle incomplete samples, the distances that correspond to missing species in a given pairwise distance matrix are assigned the value of the average distance obtained from all the other distance matrices where the species is present.

Results
Phylo-MCOA Evaluation: Simulated Data Sets
Simulated data sets were generated in order to evaluate the efficiency and robustness of Phylo-MCOA. All the data sets were created in the same way: we generated a random tree with \( N \) species by random splitting of edges as implemented in function \( \text{rtree} \) of the “ape” package in R language (Paradis et al. 2004) and duplicated this gene tree in order to get \( T \) identical gene trees. A new gene tree was randomly generated for each test performed. To create a complete outlier species, we changed its position randomly in all \( T \) trees. To generate a complete outlier gene, we replaced one of the gene trees by a completely random tree.

Subsequently, we evaluated the efficiency of Phylo-MCOA by its ability to retrieve the correct complete outlier gene and species, using the corresponding 2WR matrix. All outliers were detected based on nodal distances, using the option \( \text{distance} = \text{"nodal"} \) in Phylo-MCOA.

To simulate cell-by-cell outliers, we randomly sampled one species and one tree and changed randomly the position of the selected species in the selected tree. By repeating this operation multiple times, we obtained data sets with a variable proportion of cell-by-cell outliers.

Effect of Sample Size on the Ability of Phylo-MCOA to Retrieve Complete Outliers
Simulated data sets were obtained containing between 10 and 100 genes (10, 20, 30, 50, and 100, respectively) and between 10 and 100 species (10, 20, 30, 50, and 100, respectively); all possible combinations were explored. In all these data sets, we artificially introduced three complete outlier genes and three complete outlier species.

We analyzed these data sets with Phylo-MCOA and evaluated the proportion of complete outlier genes and species that the method was able to retrieve (table 1). We observed that as long as the number of both species and genes is large enough (\( \geq 20 \)), Phylo-MCOA always correctly retrieved the outliers (black cells in table 1). On the other hand, when the number of species is small (\( \leq 10 \)), Phylo-MCOA is unable to retrieve the outlier genes correctly, even for a large number of genes (first column, light gray cells in table 1). However, outlier species are easier to retrieve as long as the sample size is large enough (\( \geq 20 \)), even if there is a small number of genes (\( \leq 10 \), first row, dark gray cells in table 1). Finally, for moderately small number of outlier species (\( = 20 \)), outlier genes are easier to retrieve than species (second column, dark gray cells in table 1).

Effect of the Size of the Data on the Ability of Phylo-MCOA to Correctly Retrieve Complete Outlier Genes and Species

<table>
<thead>
<tr>
<th>Number of Species</th>
<th>Retrieved Genes/Species</th>
<th>Number of Outlier Genes Retrieved</th>
<th>Number of Outlier Species Retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0/0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0/0</td>
<td>1/2</td>
<td>3/3</td>
</tr>
<tr>
<td>30</td>
<td>0/0</td>
<td>0/3</td>
<td>3/3</td>
</tr>
<tr>
<td>50</td>
<td>0/0</td>
<td>0/3</td>
<td>3/3</td>
</tr>
<tr>
<td>80</td>
<td>0/0</td>
<td>0/3</td>
<td>3/3</td>
</tr>
<tr>
<td>100</td>
<td>0/0</td>
<td>0/3</td>
<td>3/3</td>
</tr>
</tbody>
</table>

*Note.—Three outlier genes and three outlier species were simulated. Each cell contains two values separated by a slash (/): number of outlier species retrieved/number of outlier genes retrieved. Nodal distances were used in the Phylo-MCOA analysis. The darker the shading, the better the retrieval of outliers by Phylo-MCOA.*
60% of the species were outliers (with values 2%, 10%, 20%, 40%, and 60%). All possible combinations of these proportions were tested. We then evaluated the number of complete outlier genes and species that Phylo-MCOA was able to retrieve correctly in each case (table 2). We observed that even when 20% of genes and species are outliers, Phylo-MCOA was still able to retrieve almost all of them correctly (black cells in table 2). A large proportion of outlier genes (40%) does not prevent retrieving outlier species. Similarly, a large proportion of outlier species is not a problem for retrieving outlier genes (dark gray cells in table 2). However, if a large proportion of both genes and species are outliers, this will make it difficult to retrieve outliers (light gray cells in table 2).

**Effect of the Proportion of Missing Data on the Ability of Phylo-MCOA to Retrieve Complete Outliers**

In typical biological data sets, there are plenty of missing data, therefore, methods for real data set analysis should be robust to incompletely sampled data sets. In order to evaluate the ability of Phylo-MCOA to retrieve outlier genes and species despite missing data, we created a data set with 100 genes and 100 species containing three complete outlier species and five complete outlier genes. We increased the proportion of missing data progressively by randomly removing species from the original set of genes. We removed 1%, 2%, 5%, 10%, 15%, 20%, 30%, and 50% of the data, respectively (table 3). Note that a proportion of 20% of missing data means that, on average, each tree has lost 20% of its leaves. We observed that Phylo-MCOA is very robust to missing data. Even with 20% of the data missing, the three complete outlier species and one of the five complete outlier genes were correctly retrieved (table 2).

**Table 2. Effect of the Proportion of Outlier Genes and Species in the Ability of Phylo-MCOA to Correctly Retrieve Them.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2 [1]</td>
<td>1/1</td>
<td>5/1</td>
<td>10/1</td>
<td>0/1</td>
<td>0/0</td>
<td></td>
</tr>
<tr>
<td>10 [5]</td>
<td>1/5</td>
<td>5/5</td>
<td>8/5</td>
<td>0/5</td>
<td>0/0</td>
<td></td>
</tr>
<tr>
<td>20 [10]</td>
<td>1/10</td>
<td>5/10</td>
<td>10/10</td>
<td>0/0</td>
<td>0/0</td>
<td></td>
</tr>
<tr>
<td>40 [20]</td>
<td>1/0</td>
<td>5/0</td>
<td>8/0</td>
<td>0/0</td>
<td>0/0</td>
<td></td>
</tr>
<tr>
<td>60 [30]</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td></td>
</tr>
</tbody>
</table>

*NOTE.—The data set was composed of 50 genes and 50 species. The proportions of outliers are given in percentage and the absolute number of outliers is given under bracket ([ ]). Each cell contains two values separated by a slash (/): number of complete outlier species/retrieved/number of complete outlier genes retrieved. Nodal distances were used in the Phylo-MCOA analysis. The darker the shading, the better the retrieval of outliers by Phylo-MCOA.*

**Table 3. Effect of the Proportion of Missing Data on the Number of Complete Outliers Retrieved (genes and species).**

<table>
<thead>
<tr>
<th>Proportion of Missing Data</th>
<th>Species retrieved</th>
<th>Genes retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2%</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>5%</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>10%</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>20%</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>30%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50%</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*NOTE.—The data set was comprised of 100 genes and 100 species with five complete outlier genes and three complete outlier species. Nodal distances were used in the Phylo-MCOA analysis. The darker the shading, the better the retrieval of outliers by Phylo-MCOA.*

**Capacity of Phylo-MCOA to Detect Complete Outliers in the Presence of Cell-by-Cell Outliers**

The detection of cell-by-cell outliers by Phylo-MCOA is regularly performed on a data set from which complete outliers have been removed. In order to verify whether Phylo-MCOA is able to correctly detect complete outliers in the presence of cell-by-cell outliers, we simulated a data set with 100 genes and 100 species containing three complete outlier species and three complete outlier genes, and evaluated the ability of Phylo-MCOA to correctly detect the complete outliers given different numbers of cell-by-cell outliers, from 1 to 1,500. Results are shown in table 4. If <1,200 cell-by-cell outliers are present, Phylo-MCOA is always able to retrieve correctly the complete outliers. For 1,500 cell-by-cell outliers, all complete outlier genes are still retrieved, but the complete outlier species are not retrieved.

**Table 4. Number of Complete Outliers Retrieved with Increasing Number of Cell-by-Cell Outliers.**

<table>
<thead>
<tr>
<th>Number of Cell-by-Cell Outliers</th>
<th>Species retrieved</th>
<th>Genes retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>300</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1,000</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1,200</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1,500</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

*NOTE.—The data set was composed of 100 genes and 100 species with three complete outlier genes and three complete outlier species. Nodal distances were used in the Phylo-MCOA analysis. The darker the shading, the better the retrieval of outliers by Phylo-MCOA.*
Phylo-MCOA Evaluation: Real Data Set
We used Phylo-MCOA on a real data set comprising 246 gene trees for 21 fungal species. This data set was used previously by Aguileta et al. (2008) for retrieving the best genes to reconstruct the phylogeny of fungi, and the data are available in FUNYBASE (Marthey et al. 2008). Aguileta et al. (2008) had compared each individual gene tree to the species tree obtained from the concatenation of the 246 genes. From this comparison, they had obtained a score for each gene tree representing how suitable it was for finding the reference (concatenated) tree. Using Phylo-MCOA, we analyzed the data set of Aguileta et al. (2008) using nodal distance between species and obtained a score for each species and gene as explained in the Materials and Methods section. These scores are plotted in figure 4 for the genes (top) and for the species (bottom). Phylo-MCOA detected five complete outlier genes (black bars in fig. 4) and no complete outlier species. The fact that no species appeared as outliers in this data set is not surprising because the goal of their work was to find optimal gene markers for reconstructing the fungal species tree, therefore, species producing incongruent trees were removed from the data set in Aguileta et al. (2008). In order to check for consistency, we compared the score that Aguileta et al. obtained for each gene (a topological similarity score referred to as topscore) to the score we obtained with Phylo-MCOA. Figure 5 (white dots) represents the scatter plot of these two scores. Phylo-MCOA is able to retrieve the correct ranking of trees and to identify genes that can be considered as outliers. We found that trees considered as outliers in Aguileta et al. (2008) were indeed also outliers according to the analysis using Phylo-MCOA (gray arrows in fig. 5). When using patristic distances instead of nodal distances for the Phylo-MCOA analysis (black dots in fig. 5), the same trend is still present but the correlation with the topscore values of Aguileta et al. (2008) seems to slightly decrease. This is expected because the topscore used by Aguileta et al. (2008) is a topological distance measure, thus more similar to Phylo-MCOA with nodal distances than to Phylo-MCOA with patristic distances.

Overall, this illustrates the general utility of Phylo-MCOA for analyzing real data sets. Moreover, the method is very fast: using the set of trees in Aguileta et al. (2008), it took about 1 min on a pc laptop.

Using Phylo-MCOA for Filtering Phylogenomic Data Sets
Deep relationships in the animal phylogeny are still debated. In 2009, Schierwater et al. (2009) proposed a phylogeny of the Metazoa that was questioned because of the low quality of the data set used to construct it. More precisely, Philippe et al. (2011) detected important problems in the data set of Schierwater et al. (2009), such as contaminations, deep paralogies, or horizontal gene transfers. They
reanalyzed the data set of Schierwater after having replaced problematic sequences by correct ones and showed that the obtained tree, even if it had less supported nodes, was more realistic given previous studies.

In order to test whether Phylo-MCOA is able to filter problematic sequences from a real data set, we analyzed the set of phylogenetic trees used by Schierwater et al. (2009). We removed the outliers detected by Phylo-MCOA, built a phylogeny using only the filtered data set (i.e., without problematic sequences) and compared our tree with the one obtained without filtering and the one produced by Philippe et al. (2011) after filtering the data.

To obtain the animal phylogeny, we used the original data set of Schierwater et al. (2009). It was first aligned and trimmed with the same tools and parameters as in Philippe et al. (2011), and the individual gene trees were reconstructed using the same model as Philippe et al. (2011). As in Philippe et al. (2011), four genes (AT8, DCR, N4L, and ND6) were removed from the analysis because they contained too few positions after trimming. All the trees were analyzed with Phylo-MCOA using patristic distances and all default parameters for the outlier detection. Four complete outliers genes were detected by Phylo-MCOA and removed before performing the cell-by-cell detection: rna16S, rna18S, rna28S, and rna58S. Of the 984 cells in the 2WR matrix (41 genes × 24 species), 21 cell-by-cell outliers were detected (see supplementary table S1, Supplementary Material online) and were removed from the original data set (by removing the sequences in the original sequence files). The remaining sequences (available as supplementary data set S1, Supplementary Material online) were then concatenated and used to reconstruct a phylogenetic tree with 100 bootstrap replicates using RAxML (Stamatakis 2006) and the same model as in Schierwater et al. (2009). The tree we obtained without filtering (fig. 6a) was similar to the one proposed by Schierwater et al. (2009), which is known to contain noisy data and thus produce an inconsistent topology. After filtering, there was a general decrease of bootstrap support for nodes in the “diploblasts” group (fig. 6b), while the support

![Fig. 5. Scatter plot of the topological congruence (top score) between 246 fungal gene trees and the consensus species tree obtained from the concatenation of individual trees (x axis) against the score of the genes according to Phylo-MCOA (y axis). Black dots represent the scores when using patristic distances. White dots represent the scores when using nodal distances. Outlier genes according to Phylo-MCOA and identified by black bars in figure 4 are shown with small gray arrows. They appear to be poorly congruent with the species tree (small values on the x axis).](image)

![Fig. 6. Comparison of the phylogeny of animals from the Schierwater et al. (2009) data set. (a) Tree obtained from the original sequences. (b) Tree obtained after removal of cell-by-cell outliers using Phylo-MCOA. Species with a star (*) represent Porifera. The gray arrows point to the node support of the “diploblasts” group in the two trees. Nodes with black dots represent bootstrap supports higher than 95. A hundred bootstrap replicates were performed for each tree.](image)
of the nodes outside the “diploblasts” were almost not affected. Overall, the bootstrap support for the monophyly of “diploblasts” went from 100 in the original data set to 63 in the filtered data set (gray arrows in fig. 6), which is in accordance with the accepted vision that “diploblasts” are certainly not monophyletic. The monophyly of Porifera (species with an * in fig. 6) was lost, which is also in accordance with previous studies and coherent with the result of Philippe et al. (2011) on this data set, who obtained a node for the Porifera with a support value of 36 (Philippe et al. 2011).

Phylo-MCOA was thus able to identify true outliers that need to be filtered out prior to further phylogenetic analysis in order to estimate a more consistent topology. However, we could not retrieve exactly the same topology as the one obtained by Philippe et al. This can be due to the fact that we do not replace the sequences identified as problematic by correct sequences as done in Philippe et al. (2011). Moreover, Phylo-MCOA considers that the correct signal is the one that is present in the majority of the genes, but this is not true in this data set for the Hexactinellida, where the majority of the sequences are slow evolving mitochondrial sequences being contaminations from the Demospongiae, while true nuclear sequences of Hexactinellida are fast evolving (Philippe et al. 2011). As a consequence, Hexactinellida and Demospongiae still form a very solid group after filtering, and the branch length leading to Hexactinellida is reduced in the new tree (fig. 6b) compared with the original tree (fig. 6a).

The Phylo-MCOA Program: Availability and Requirements
Phylo-MCOA is provided as a set of R functions which make use of the ade4 (Chessel et al. 2004; Dray et al. 2007), ape (Paradis et al. 2004), and lattice (for graphical functions, Sarkar 2010) packages and newly developed functions. It requires a recent version of R (>=2.8.0) installed on the machine. Phylo-MCOA is freely distributed under the GNU General Public License. It can be downloaded at http://phylomcoa.cgenomics.org.

Discussion
It has become a standard practice to use multiple genes in order to obtain a reliable and concordant phylogeny. When sampling many different genes for the same set of species, the individual gene trees will often be partly incongruent. In this case, one is interested in detecting which subset of genes and species is responsible for the discrepancies, in other words, in finding the outliers. Phylo-MCOA, the method proposed here, provides a quick and powerful way to detect outliers by integrating the phylogenetic information contained in all the individual gene trees. Phylo-MCOA determines how different the compared topologies are but more importantly, it identifies which genes and species explain those differences. This may prove very useful when determining possible causes for incongruities, such as horizontal gene transfers, different selective pressures, recombination events, among others. Outlier genes are typically affected by these factors, so identifying which subset of species explain the differences may help to focus the efforts for studying the affected species. On the other hand, by building a consensus topology Phylo-MCOA also finds the genes that produce concordant phylogenies. These genes are excellent candidates for recovering the species tree, which can be subsequently verified by other methods. Phylo-MCOA thus provides good candidates, not only for outliers, but for phylogenetic markers producing concordant topologies.

To our knowledge, this is the first time that MCOA is applied directly to analyze phylogenetic distances. Based on a covariation optimization criterion, MCOA is a method that captures the similarities (and dissimilarities) obtained from the ordination of several tables (types) of data, as it has been previously demonstrated for the analysis of ecological data (Bady et al. 2004; Hedde et al. 2005; Laloe et al. 2007). Furthermore, it is a statistically solid technique that lends itself well to the analysis of phylogenetic data (i.e., in terms of topological distances), where the covariation of the histories of genes and species can be integrated in a single ordination analysis. Even though a consensus topology is proposed by the method, MCOA takes into account all the gene trees in order to build it. We believe this makes a better use of the individual tree information than other forms of phylogenetic consensus that do not take into account outliers or use only a part of the available data. Our method is unique in simultaneously detecting outlier species and genes from the comparison of multiple gene trees.

One important aspect of tree comparisons involves the estimation of pairwise distance matrices. The Eigen decomposition and further ordination represented in the consensus topology take the distances as absolute distances. Therefore, the consensus topology will change depending on the type of distance that is actually being measured. There are two ways to calculate the distance between two species in a phylogenetic tree: the sum of the branch lengths or the number of nodes separating the species in the trees. Both distance metrics have advantages and disadvantages: using the sum of the branch lengths (patristic distances) allows incorporating the relative rates of evolution but can artificially bring closer species that have small branch lengths and separate species with longer branches. Conversely, if the internodal distance separating species is used, then the consensus topology informs us about a degree of phylogenetic proximity, but it tells us nothing about the relative rates of evolution. It is important to be explicit about the pairwise distance method employed in order to derive conclusions, as genes identified as outliers when using one method may seem concordant with other genes when using another one. Comparing the results given by both methods can be very informative. A gene that is identified as an outlier when using patristic distances but not when using nodal distances will represent a gene evolving at higher or slower rates in some species compared with other genes.
Phylo-MCOA allows weighing trees according to different scores (e.g., likelihood scores), which ensures that more credible trees are given more weight in the analysis. It is equally important to assign weights to nodes to distinguish supported relationships from spurious or dubious ones. In principle, one could incorporate a measure of node support into MCOA metrics; however, this is not straightforward. At present, we have arrived at a solution that may alleviate the problem of node support assignment to some extent. We have incorporated the possibility of collapsing poorly supported nodes, according to a cutoff value given by the user. In this way, the nodes that are poorly supported will contribute little or nothing to the construction of the consensus typology. Assessing node support across phylogenies is difficult, especially with cases of missing or duplicated data, as not all nodes will be present in all trees. For this reason, we suggest the use of Treeko (Marcet-Houben and Gabaldón 2011) for trees with duplicated genes, prior to analysis with Phylo-MCOA. Interestingly, our method can be used even in the absence of completely sampled data sets. The program can deal with gene losses, as demonstrated by randomly deleting species from the simulated data analyzed.

We have shown that Phylo-MCOA is able to correctly detect outlier genes and species from the comparison of multiple phylogenetic trees. We have shown this by analyzing simulated data sets but, more interestingly, from the analysis of real data sets. We have shown, with two different examples, how Phylo-MCOA can detect true outliers. First, with the fungal data set by Aguileta et al. (2008), where we found the same concordant and discordant genes as in the original study; and second, where we use Phylo-MCOA to predict which cell-by-cell outliers need to be removed from phylogenetic reconstruction in order to recover the animal phylogeny obtained after removing noisy data, as done by Philippe et al. (2011).

Phylo-MCOA can be used to detect candidate genes for further analysis, or for filtering out problematic genes in phylogenetic reconstruction. Phylo-MCOA can detect two types of outliers. Complete outliers can identify which genes and species deviate from the evolutionary history of most genes analyzed. This information can be very useful for filtering out noisy data in phylogenetic reconstruction. On the other hand, cell-by-cell outliers point to particular cases of topological discordance, where one specific species is an outlier in a given gene. This information can be used to postulate specific hypotheses concerning particular genes or species. Although cell-by-cell outlier detection is implemented in the current Phylo-MCOA version, there is room for improvement, specifically in the way outliers are scored to define significance. Phylo-MCOA can be used to detect candidate genes for further analysis, or for filtering out problematic genes in phylogenetic reconstruction. Nevertheless, caution should always be exerted because the method will give more weight to the information given by the majority of genes under study, and if the data set analyzed is plagued with dubious sequences, the method will favor an incorrect signal. This is what we observed when we analyzed the data set of Schierwater et al. (2009) that is known to be ridden with multiple faulty sequences. External information on the quality of the sequences should be employed whenever possible prior to analysis with our method.

### Supplementary Material

Supplementary table S1 and data set S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

### Acknowledgments

We would like to thank Hervé Philippe for the time he spent on this manuscript and the helpful comments and suggestions he made. We thank David Moreira for constructive discussions, Avner Bar-Hen for interesting comments on the use of MCOA for phylogenomics analysis, Vincent Berry for suggestions on future developments and comments on distances between species, and Fran Supek for his suggestions on statistical significance testing. Finally, we thank Tatiana Giraud and Toni Gabaldón for insightful comments on the manuscript as well as three anonymous reviewers. D.M.V. acknowledges postdoctoral grants from the Centre National de la Recherche Scientifique (CNRS) and Université Paris-Sud 11. G.A. acknowledges postdoctoral grants from the CNRS, Université Paris-Sud 11, and Agence National de la Recherche (ANR-06-BLAN-0201).

### References


