Minimum Contradiction Matrices in Whole Genome Phylogenies

Marc Thuillard
Belimo Automation AG, CH-8340 Hinwil, Switzerland.

Abstract: Minimum contradiction matrices are a useful complement to distance-based phylogenies. A minimum contradiction matrix represents phylogenetic information under the form of an ordered distance matrix $Y_{ij}^n$. A matrix element corresponds to the distance from a reference vertex $n$ to the path $(i,j)$. For an X-tree or a split network, the minimum contradiction matrix is a Robinson matrix. It therefore fulfills all the inequalities defining perfect order: $Y_{ij}^n \geq Y_{ik}^n \geq Y_{kj}^n$, $i \leq k \leq j < n$. In real phylogenetic data, some taxa may contradict the inequalities for perfect order. Contradictions to perfect order correspond to deviations from a tree or from a split network topology. Efficient algorithms that search for the best order are presented and tested on whole genome phylogenies with 184 taxa including many Bacteria, Archaea and Eukaryota. After optimization, taxa are classified in their correct domain and phyla. Several significant deviations from perfect order correspond to well-documented evolutionary events.

Keywords: phylogenetic trees, whole genome phylogeny, minimum contradiction, split network

1. Introduction

The discovery of the importance of lateral transfers, losses and duplications events in the evolution of genetic sequences has motivated the development of new approaches to graphically represent phylogenies. Methods like NeighborNet (Bryant and Moulton, 2004), T-Rex (Makarenkov et al. 2006), SplitTrees (Bandelt and Dress, 1992; Dress and Huson, 2004; Huson, 1998), Qnet (Grünewald et al. 2006), Pyramids (Bertrand and Diday, 1985), Tree of Life (Kunin et al. 2005a) allow visualizing deviations from a tree topology. All these methods have in common that they summarize the information in the form of a planar network. Deviations from an X-tree are often represented by supplementary edges (Makarenkov et al. 2006; Nakhleh et al. 2004) that create cycles in the graph.

Phylogenetic information can be represented by a distance matrix $Y_{ij}^n$. For an X-tree, the elements of the distance matrix $Y_{ij}^n$ correspond to the distance from a reference taxon $n$ to the path $(i,j)$. The taxa can be ordered through permutations, so that the distance matrix is a Robinson matrix (Bertrand and Diday, 1985), with values of both rows and columns decreasing away from the diagonal. The corresponding circular order is defined as a perfect order. We have shown with a probabilistic model that perfect order is quite robust against lateral transfer and crossover (Thuillard, 2007). The search for the order minimizing a measure of the deviation from perfect order can be efficiently done with a multi-resolution algorithm (Thuillard, 2001, 2007). The method has been tested on SSU rRNA data for Archaea. The matrix with the best order corresponds quite well to a Robinson matrix. In this article, the minimum contradiction approach is further developed and applied to whole genome phylogenies.

With the availability of complete genomes, many methods have been proposed to determine the evolution of whole genomes (For reviews see Galperin et al. 2006; Delsuc et al. 2005; Henz et al. 2005). The construction of trees from whole genomes has proved over recent years to be a quite difficult task. This is mainly because of the very limited number of genes shared by Archaea, Eukaryota and Bacteria. Furthermore, gene evolution can sometimes be very different from species evolution. The main difficulty consists in finding a good operator to estimate the distance between genomes. Distances have been estimated with measures based on gene order or arrangement (Wolf et al. 2002; Wang et al. 2006), gene content (Fitz-Gibbon and House, 1999; Snell et al. 1999; Korbel et al. 2002), protein domain organization (Fukami-Kobayashi et al. 2007; Yang et al. 2005), folds (Lin and Gerstein, 2007), combining the information from many genes in a supertree or a superdistance (Dutihl et al. 2007 for a comparative study) or using a local alignment search tool such as Blast (Kunin et al. 2005b; Clarke et al. 2002).
Among genome distances obtained with Blast, the genome conservation (Kunin et al. 2005b) has furnished some of the best trees up to date, if the quality of a whole genome phylogeny is measured by its concordance to broadly accepted classifications. The genome conservation estimates the distance between two taxa using the sum of BlastP reciprocal best hits between two genomes. The method is capable of quite correctly recovering all main phyla. At the phylum level, the evolution of different genes is sufficient similar to form a distinct cluster. The main uncertainties in whole genome phylogenies are on the relationships between phyla. Different evolution rates of the genes, gene losses or duplications, lateral gene transfer may result into large deviations of the distance matrix from a tree topology. In this context, minimum contradiction matrices can furnish information not contained in a single tree or a split network.

The paper is organized as follows. After introducing minimum contradiction matrices in section 2 and their connection to Robinson matrices and Kalmanson inequalities, section 3 explains why the identification of deviations from perfect order is a useful complement to phylogenetic studies. Section 4 presents an algorithm to search for the order minimizing a measure of the deviation from perfect order over all taxa. This order can be interpreted as an average best order over all reference taxa \( Y_{i,j}^N \) \((N = 1, \ldots, n)\). The algorithm is applied in section 5 to distance matrices for whole genome phylogenies obtained with the genome conservation method.

### 2. Circular Order and the Minimum Contradiction Approach

#### 2.1. Definitions

Let us start by recalling a number of definitions that are necessary to introduce the notion of circular order. A graph \( G \) is defined by a set of vertices \( V(G) \) and a set of edges \( E(G) \). Let us write \( e(x,y) \), the edge between the two vertices \( x \) and \( y \). In a graph \( G \), a path \( P \) between two vertices \( x \) and \( y \) is a sequence of non-repeating edges \( e(x_1,z_1), e(z_1,z_2), \ldots, e(z_{n-1},y) \) connecting \( x \) to \( y \). The degree of a vertex \( x \) is the number of edges \( e \in E(G) \) to which \( x \) belongs. A leaf \( x \) of a graph is a vertex of degree one. A vertex of degree larger than one is called an internal vertex.

A valued X-tree \( T \) is a graph with \( X \) as its set of leaves and a unique path between any two distinct vertices \( x \) and \( y \), with internal vertices of at most degree 3. The distance \( d \) between leaves satisfies the classical triangular inequality

\[
d(x,z) \leq d(x,y) + d(y,z) \text{ for all } x,y,z \in X
\]

with \( d(x,y) \) representing the sum of the weights on the edges of \( T \) in the path connecting \( x \) and \( y \).

A central problem in phylogeny is to determine if there is an \( X \)-tree \( T \) and a real-valued weighting of the edges of \( T \) that fits a dissimilarity matrix \( \delta \). Typically, a dissimilarity matrix \( \delta \) corresponds to an estimation of the pairwise distance \( d(x_i,x_j) \) between all elements in \( X \). A necessary and satisfactory condition for the existence of a unique tree is that the dissimilarity matrix \( \delta \) satisfies the so-called 4-point condition (Bunemann, 1971). For any four elements in \( X \), the 4-point condition requires that

\[
\delta(x_i,x_k) + \delta(x_j,x_p) \leq \max(\delta(x_i,x_j),
\delta(x_i,x_p) + \delta(x_j,x_k)).
\]

#### 2.2. Circular order and Kalmanson inequalities

Consider a planar representation of a tree \( T \) or a split network \( S \). A circular order corresponds to an indexing of the \( n \) leaves according to a circular (clockwise or anti-clockwise) scanning of the leaves (Barthélémy and Guénoche, 1991; Makarenkov and Leclerc, 1997, 2000; Yushmanov, 1984).

In an X-tree, a circular order has the property that for any integer \( k \) (modulo \( n \)), all the branches on the path \( P(x_k, x_{k+1}) \) between \( x_k \) and \( x_{k+1} \) correspond to the left branch (or right branch if anti-clockwise). A circular order can be obtained by considering the distance matrix \( Y_{i,j}^n \). As illustrated in Figure 1, the matrix element \( Y_{i,j}^n = \frac{1}{2}(d(x_i,x_n) + d(x_j,x_n) - d(x_i,x_j)) \) corresponds to the distance between a reference leaf \( n \) and the path \( P(x_i,x_j) \). A circular order can be computed by ordering the distance matrix \( Y_{i,j}^n \) so that it fulfills the inequalities defining a perfect order

\[
Y_{i,j}^n \geq Y_{i,k}^n, Y_{k,j}^n \geq Y_{k,i}^n \text{ for all } i \leq j \leq k < n.
\]
The above inequalities characterize also a Robinson matrix (Christopher et al. 1996; Thuillard, 2007). Using the definition of $Y_{i,j}^n$ the inequalities become

$$d(x_j, x_n) + d(x_i, x_k) \geq d(x_i, x_n) + d(x_j, x_k)$$

and

$$d(x_j, x_n) + d(x_k, x_j) \geq d(x_i, x_k) + d(x_j, x_i)$$

and

$$d(x_k, x_j) (i \leq j \leq k < n). \quad (3b)$$

These inequalities have a similar form to the 4-point condition (2) and are known as the Kalmanson inequalities.

2.3. Minimum contradiction matrix

In real applications, the distance matrix $Y_{i,j}^n$ does often only partially fulfill the inequalities corresponding to a perfect order. The contradiction on the order of the taxa can be defined as

$$C = \sum_{k > j \geq i, i, j, k \neq n} \left( \max \left( \left( Y_{i,k}^n - Y_{i,j}^n, 0 \right) \right) \right)^\beta$$

$$+ \sum_{k > j > i, i, j, k \neq n} \left( \max \left( \left( Y_{i,k}^n - Y_{j,k}^n, 0 \right) \right) \right)^\beta. \quad (4)$$

For a perfectly ordered X-tree, the contradiction $C$ is zero. A tree with a low contradiction value $C$ is a tree that can be trusted, while a high contradiction value $C$ is the indication of a distance matrix deviating significantly from an X-tree.

3. Why Perfect Order is an Important Property?

Kalmanson inequalities are at the center of a number of important results relating convexity (Kalmanson, 1975), the Traveling Salesman Problem (TSP) (Deineko et al. 1995; Korostensky and Gonnet, 2000), phylogenetic trees and networks (Christopher et al. 1996; Dress and Huson, 2004). Let us explain why perfect order is an important property.

- If the error on the distance in an X-tree is not greater than $x_{min}/2$ with $x_{min}$ the shortest edge on the tree, then the Neighbor-Joining algorithm will recover the correct tree topology and Kalmanson inequalities hold (Atteson, 1999; Korostensky and Gonnet, 2000).

- If a distance matrix fulfills Kalmanson inequalities, then the distance matrix can be exactly represented by a split network (Bandelt and Dress, 1992).

- If Kalmanson inequalities are fulfilled, then the tour (1, 2, ..., $n$) corresponds to a solution of the Traveling Salesman Problem (Christopher et al. 1996).

The last result can be demonstrated starting from the sum $\sum_{i=1, \ldots, n-2} Y_{i,i+1}^n$. When Kalmanson inequalities are fulfilled, the sum $\sum_{i=1, \ldots, n-2} Y_{i,i+1}^n$ is maximized. As $Y_{i,i+1}^n \geq Y_{i,i+m}^n (i + m \leq n, m > 1)$, one gets $\sum_{i=1, \ldots, n-1} Y_{i,i+1}^n = \sum_{i=1, \ldots, n} d_{i,n} - 1/2. (d_{i,n} + \sum_{i=1, \ldots, n} d_{i,i+1})$. The first sum $\sum_{i=1, \ldots, n} d_{i,n}$ is independent of the order and one concludes that a perfect order minimizes $d_{i,n} + \sum_{i=1, \ldots, n-1} d_{i,i+1}$. The tour (1, 2, ..., $n$) is therefore a solution of the TSP.

The solution to the TSP has the Master Tour property (Deineko et al. 1995). A Master Tour is a solution of the TSP with the property that the optimal tour restricted to a subset of points is also a solution of the reduced TSP. This result follows directly from the inequalities for perfect order $Y_{i,j}^n \geq Y_{i,k}^n$, $Y_{i,j}^n \geq Y_{k,j}^n$ ($i \leq j \leq k < n$). Any restriction of a perfectly ordered distance matrix...
to a subset of taxa is perfectly ordered and consequently is a solution to the reduced TSP. In contrast to this result, one finds with numerical experiments that, if the minimum contradiction matrix does not fulfill the inequalities for perfect order, the best order is not always preserved when a number of taxa are removed. The order minimizing the contradiction over n taxa does not always minimize the contradiction when restricted to a subset of taxa. It follows that one cannot exclude that the topology of a tree or a split network may change when taxa contradicting perfect order are removed. Deviations from perfect order correspond to problematic regions that have to be interpreted very carefully. For that reason we suggest that minimum contradiction matrices are a useful complement to any distance-based phylogeny.

4. Searching for the Best Order in Whole Genome Phylogenies

4.1. Fast algorithm to search for the best order

The choice of the reference taxon \( n \) in \( Y^{n}_{i,j} \) can significantly influence the best order, when the distance matrix cannot be perfectly ordered. For that reason, an average best order is determined by minimizing the contradiction over all reference taxa.

The contradiction over all \( n \) reference taxa is given by

\[
C = \sum_{m=1}^{n} \left( \sum_{k>j>i, k \neq n} \left( \max \left( \left( Y^{n(m)}_{i(m),k(m)} - Y^{n(m)}_{j(m),k(m)} \right) , 0 \right) \right) \right) + \sum_{k>j>i, k \neq n} \left( \max \left( \left( Y^{n(m)}_{i(m),k(m)} - Y^{n(m)}_{j(m),k(m)} \right) , 0 \right) \right) \right)^{\beta} \tag{5}
\]

with \( i(m) = \text{mod}(m + i_{0} - 2, n) + 1; j(m) = \text{mod}(m + j_{0} - 2, n) + 1; n(m) = n_{0} - m + 1 \) and \( \beta = 2 \).

The best order is the order \( (1, \ldots, i_{0}, \ldots, j_{0}, \ldots, n_{0}) \) minimizing the contradiction. The computation of the contradiction requires \( O(n^4) \) operations. For a large ensemble of taxa, the computational cost may become quite high. We will therefore introduce below an algorithm requiring only \( O(n^3) \) operations to compute a (slightly different) measure of the contradiction.

Let us start by considering an X-tree and the 3 vertices \( i,j,k \) as in Figure 2. The distance matrix fulfills the inequalities for perfect order. The order between the vertices \( i,j,k \) is preserved for any reference vertex not in the interval \((i,k)\) and the inequalities \( Y^{n}_{i,j} \geq Y^{n}_{i,k} \) and \( Y^{n}_{k,j} \geq Y^{n}_{k,i} \) \( n = 1, \ldots, i, k, \ldots, N \) hold. The inequalities can be summed up over all \( n \) and one obtains two new inequalities:

\[
Sa(i,j) = \sum_{n=1}^{i} Y^{n}_{i,j}; \quad Sb(i,j) = \sum_{n=k}^{N} Y^{n}_{i,j} \tag{6c}
\]

With

\[
Sa(i,j) = \sum_{n=1}^{i} Y^{n}_{i,j} \quad (i \leq j \leq k) \tag{6a}
\]

\[
Sb(i,j) = \sum_{n=k}^{N} Y^{n}_{i,j} \quad (i \leq j \leq k) \tag{6b}
\]

If the contradiction \( c_{i,j} \) between the vertices \( i,j \) is defined as the sum of two terms

\[
c_{i,j} = c_{a_{i,j}} + c_{b_{i,j}} \text{ with} \tag{7a}
\]

\[
c_{a_{i,j}} = \sum_{k>j>i} \max(0,(Sa(i,k) - Sa(i,j))^2) \tag{7b}
\]

\[
c_{b_{i,j}} = \sum_{k>j>i} \max(0,(Sb(i,k) - Sb(i,j))^2) \tag{7c}
\]
then the best order is the order minimizing 
\[ c = \sum_{i=1}^{n} c_{a_{i},j} + c_{b_{i},j}. \]
Computing the contradiction requires \( O(n^3) \) operations (As the computation of the contradiction is the most computer-intensive, the algorithm requires approximately \( n \) times less computing time than the \( O(n^4) \) algorithm).

The quantities \( Sa \) and \( Sb \) in Eq. (6) can be related to the NJ algorithm. For 3 consecutive vertices \((i, j = i + 1, k = i + 2)\), Eq. (6a) can be written, assuming perfect order, as

\[ \sum_{n \neq i, i + 1, i + 2} Y_{i,j}^n \approx \sum_{n \neq i, i + 1, i + 2} Y_{i,j+2}^n \]

Writing \( r_i = \sum_{n=1}^{N} d(x_i, x_n) \) and \( S_{i,j} = r_i + r_j - (N-2) \cdot d(x_i, x_j) \) one obtains

\[ S_{i,j+1} - S_{i,j+2} \geq 0. \]

The value \( S_{i,j} \) is central to the NJ algorithm (Saitou and Nei, 1987; Gascuel and Steel, 2006). Two vertices \((i, j)\) are joined by the NJ algorithm, if they maximize \( S \) (i.e. \( \text{max} (S) = S_{i,j} \)). From the above discussion, it seems natural to initialize the search for the best order on the NJ tree. The search for the best order of \( Y_{i,j}^n \) is initialized with the NJ algorithm and a small supplementary procedure that we describe below. Given two vertices \((a, b)\) that are joined by the NJ algorithm and the leaves \( a_1, a_2, \ldots, a_i \) (resp. \( b_1, b_2, \ldots, b_j \)) that have the vertex \( a \) (resp. \( b \)) as first ancestor. The best order of the leaves is chosen so as to minimize the contradiction among 4 possibilities: \((ab, \bar{a}b, a\bar{b}, \bar{a}\bar{b})\) with \( ab \) the order \( a_1, a_2, \ldots, a_i, b_1, b_2, \ldots, b_j \) and \( \bar{a} \) the inverted order \( a_i, a_{i-1}, \ldots, a_1 \). Once the order is optimized over the NJ tree, the best order is refined with a multiresolution search algorithm (Thuillard, 2001, 2007).

4.2. Similarity matrix for whole genomes phylogenies

For whole genome phylogenies, the search for appropriate measures to estimate the evolutionary distance between taxa is still the subject of significant research efforts (Korbel et al. 2002; Kunin et al. 2005b; Yang et al. 2005; Fukami-Kobayashi, 2007). Distance matrices obtained from BlastP scores have been quite successful to generate good trees. The similarity score obtained with BlastP programs can be given a probabilistic interpretation. The statistics of high scoring segments in the absence of gaps tends to an extreme value distribution (Karlin and Altschul, 1990). The probability \( P \) of finding at least a high scoring segment is well approximated, for small values of \( P \), by the formula \( P = m_1 \cdot m_2 \cdot 2^{-\text{Score}} \) with \( m_1, m_2 \) the length of the 2 sequences. It follows that \( \text{Score} = -\log_2 P + \log_2 (m_1 \cdot m_2) \). Defining the distance \( d \) between two sequences as \( d = -\text{Score} \) and assuming equal lengths one has \( d = \log_2 (P/m^2) \). Using that definition, the distance matrix \( Y_{i,j}^n \) becomes for 3 sequences

\[ Y_{i,j}^n = 1/2 \cdot \log_2 \left( \frac{P(i|n) \cdot P(j|n)}{P(i|j) \cdot m^2} \right) \]

The log term has the form of a mutual information and furnishes a measure of the similarity of the genomes \( i \) and \( j \) in reference to the genome \( n \).

Different approaches have been proposed to normalize the distance matrix using the marginal entropy (Kraskov et al. 2005), the self-score (Kunin et al. 2005b), Korbel normalization (Korbel et al. 2002) or the average score. The normalization by the self-score in the genome conservation gives some of the best results. It is based on a nonlinear weighted sum of the BlastP scores. The gene conservation method computes the distance between two taxa by normalizing the sum of reciprocal best hits between genome \( j \) and \( j \) by the self-score. The effect of duplication is limited by using only reciprocal best hits. The normalization by the self-score is important to correct, at least partially, the effect of different genome sizes. The genome conservation similarity matrix is given by

\[ S_{i,j} = \min \left( \frac{\sum (i,j), \sum (j,i)}{\min (\sum (i,i), \sum (j,j))} \right) \]

with \( \sum (i,j) \) the sum of reciprocal best hits between the genomes of the two taxa.

5. Whole Genome Phylogenies

5.1. Search for the best average order

The algorithms described in section 4 have been used to search for the best order. The distance matrix
was computed using the data furnished by the genome phylogeny server (Kunin et al. 2005b) obtained with an e-value cut-off set to $10^{-10}$. The contradiction is significantly lower with the score $(1 - S_{i,j})$ than with the logarithm of the score. Figure 3 shows the best order after optimization with the algorithms described in section 4 followed by 5000 steps of the multiresolution search algorithm using Eq. (7) to compute the contradiction.

Table 1 gives the order of the different taxa corresponding to the best order. Archaea and Eukaryota are grouped into two adjacent clusters of taxa. One observes, for Bacteria, that all the members of a class or a phylum are neighbors. All proteobacteria (together with Aquifex?) are grouped together. The best order obtained with the minimum contraction approach differs from the NJ tree on the following aspect: all spirochetes and δ-proteobacteria form a cluster. This is not the case of the NJ tree.

5.2. Interpreting minimum contradiction matrices

This article focus on the mathematical aspects of Minimum Contradiction Matrices. We will limit the discussion to 3 examples showing how to interpret Minimum Contradiction Matrices. The matrix $Y_{i,j}^n$ can be imaged for different reference taxa using the best order of Figure 3 given in the annex. Figure 4 shows the matrix $Y_{i,j}^n$ using *Pirellula* (taxa 177) as reference taxa. The scale on the right of the figure gives the color code used to represent $Y_{i,j}^n$ after rescaling. The minimum value of $Y_{i,j}^n$ corresponds to dark blue, while the largest values are coded red. Low values of $Y_{i,j}^n$ are associated to two vertices $(i, j)$ having a first common ancestor vertex close to the reference taxa. A cluster of adjacent taxa with large values (red cluster) can be interpreted as a group of close taxa. One observes that Archaea and Eukaryota are not only adjacent but form also a cluster.

![Figure 3. Minimum contradiction matrices corresponding to the best order found after optimization with Eq. 6,7. The contradiction is minimized over the lines of the matrix $Sa(i, j) = \sum_{s=k...l} Y_{i,j}^s$ (left) and the columns $Sb(i, j) = \sum_{s=k...N} Y_{i,j}^s$ (right).](image-url)
The best order in Figure 3 is obtained by minimizing the contradiction using all taxa as reference vertex at least once. The best order is therefore a kind of “average” best order. The matrix $Y_{ij}^n$ (resp. $\sum_{n=n_1,\ldots,n_k} Y_{ij}^n$) with $n$ corresponding to a unique taxon (resp. a group of taxa belonging to some phylum) allows the identification of large contradictions from the best order. These contradictions can often be specifically related to the reference taxon. A loss of a gene, a lateral gene transfer or a crossover in the reference taxon modifies all elements of the distance matrix $Y_{ij}^n$. A similar perturbation on a taxon that is not a reference taxon affects at most the row and the column corresponding to that taxon.

Many contradictions in Figure 5 can be associated to well accepted endosymbiotic events (Chloroplasts in plants or mitochondria in Eukaryota). Figure 5a shows $Y_{ij}^n$ for Archaea, Eukaryota and some Bacteria (Taxa 72–116) using Rickettsiales (Taxa 1–4 in annex) as reference taxa. The average best order is used to order the taxa. Contradictions on the order of the taxa are identified by looking for regions with $Y_{ij}^n$, increasing away from the diagonal (i.e. $Y_{ij}^n < Y_{ik}^n$, $i < j < k < n$). Contradictions are observed for $i = $ Bacteria (without Mycoploasma) $j = $ Eukaryota. One observes that $Y_{ij}^n$ decreases away from the diagonal except between Eukaryota and Archaea (dark blue compared to light blue for Archaea). This result is, at first glance, somewhat surprising. Similar values of $Y_{ij}^n$ for Archaea and Eukaryota are expected.

\begin{table}[h!]
\centering
\caption{Best Order (Fig.3, 4).}
\begin{tabular}{ll}
1. & \textbf{\textit{\alpha-}}\textit{-Proteobacteria} & 1–14 \\
2. & \textbf{\textit{\gamma-}}\textit{-Proteobacteria} & 15–18 \\
3. & \textbf{\textit{\beta-}}\textit{-Proteobacteria} & 19–29 \\
4. & \textbf{\textit{\gamma-}}\textit{-Proteobacteria} & 30–54 \\
5. & \textbf{\textit{\varepsilon-}}\textit{-Proteobacteria} & 55–59 \\
6. & Aquificae & 60 \\
7. & \textbf{\textit{\delta-}}\textit{-Proteobacteria} & 61–63 \\
8. & Chlorobi & 64 \\
9. & Bacteroidetes & 65–66 \\
10. & Spirochetes & 67–71 \\
11. & Thermotogae & 72 \\
12. & Fusobacteria & 73 \\
13. & Firmicutes & 74–116 \\
14. & Eukaryota & 117–135 \\
15. & Archaea & 136–152 \\
16. & Actinobacteria & 153–166 \\
17. & Deinococcus-Thermus & 167–168 \\
18. & Cyanobacteria & 169–176 \\
19. & Planctomycetes & 177 \\
20. & \textbf{\textit{\chi-}}\textit{-Proteobacteria} & 178–184 \\
\end{tabular}
\end{table}
when \( i, n \) correspond to Bacteria. The low values for Eukaryota can be explained by a lateral transfer between the Rickettsiales and Eukaryota. We have shown with a probabilistic model (Thuillard, 2007) that a lateral transfer between the reference taxa and some taxa reduces the expected values of \( Y_{i,j}^n \) for those taxa. In this model, the expected value \( \hat{Y}_{i,j}^n \) after an \( \alpha \)-lateral transfer is given by

\[
\hat{Y}_{i,j}^n = (1 - \alpha) Y_{E_1,E_2}^R + \alpha Y_{R_1,R_2}^R \leq Y_{E_1,E_2}^R
\]

with \( \alpha \) the proportion of the genome laterally transferred (\( \alpha \leq 1 \)) from the reference taxa \( R \), and \( R_1, R_2 \) the laterally transferred sequence after further evolution into the Eukaryota genomes \( E_1, E_2 \). The observed contradiction and the small values of \( Y_{i,j}^n \) for Eukaryota are consistent with a lateral transfer between the reference taxa (Rickettsiales) and Eukaryota. Let us recall here that mitochondria are believed to be the result of an endosymbiotic event involving Rickettsia (Timmis et al. 2004), an event that resulted also into the transfer of some Rickettsia genes into the nucleus of the host.

Conclusions

For an X-tree or a split network the minimum contradiction matrix \( Y_{i,j}^n = \frac{1}{2} (d(x_i, x_n) + d(x_j, x_n) - d(x_i, x_j)) \) fulfills all the inequalities defining perfect order (i.e. \( Y_{i,j}^n \geq Y_{i,k}^n, Y_{j,k}^n \), \( i \leq j \leq k \leq n \)). In real applications a number of taxa may typically be in contradiction to the inequalities for perfect order. In that case, the Master Tour property does not hold. It follows that the removal or the addition of taxa in contradiction to the inequalities may change the topology of the associated NJ tree or split network.

An average best order can be obtained by searching for the best circular order over \( Y_{i,j}^n \) \( (N=1, \ldots, n) \). The matrix \( Y_{i,j}^n \) can be used to localize a problematic taxon, as large deviations from the average best order are often related to the reference taxon \( n \). This approach was applied to whole genome phylogenies using distances computed with the genome conservation method.
Several large deviations from the average best order were found to correspond to well-documented evolutionary events.

Disclosure
The authors report no conflicts of interest.

References
Atteson, K. 1999. The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction. *Algorithmica*, 25:251–78.

Barthélémy, J.P. and Guénoche, A. 1991. Trees and proximity representations. New York: Wiley.

Bandelt, H.J. and Dress, A. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetic Evolution*, 1:242–52.

Bertrand, P. and Diday, E. 1985. A visual representation of the compatibility between an order and a dissimilarity index: the pyramids. *Computational Statistics Quarterly*, 2:31–44.

Bryant, D. and Moulton, V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21:255–65.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. In: *Buneman, P.*, ed. *Proceedings of the 2nd International Colloquium on Graphs, Mathematics and Theoretical Computer Science*. Providence: Amer. Math. Soc., 183–208.

Christopher, G.E., Farach, M. and Trick, M.A. 1996. The structure of circular decomposable metrics. In European Symposium on Algorithms (ESA’96), Lectures Notes in Computer Science, 1136:455–500.

Clarke, G.D.P., Beiko, R., Ragan, M.A. and Charlebois, R.L. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BlastP scores. *Journal of Bacteriology*, 184:2072–80.

Delsuc, F., Brinkmann, H. and Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6:361–76.

Deineko, V., Rudolf, R. and Woeginger, G. 1995. Sometimes traveling is easy: the master tour problem, Institute of Mathematics, SIAM.

Dress, A. and Huson, D. 2004. Constructing split graphs. IEEE Transactions on Computational Biology and Bioinformatics, 1:109–15.

Dutill, B.E., Noort, V., Heijden, R.T.J.M., Boekhout, T., Snel, B. and Huynen, M.A. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics*, 23:815–24.

Fitz-Gibbon, S.T. and House, C.H. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, 27:4718–222.

Fukami-Kobayashi, K., Minezaki, Y., Tateno, Y. and Nishikawa, K. 2007. A tree of life based on protein domain organizations. *Molecular Biology and Evolution*, 24:1181–9.

Galperin, M.Y. and Kolker, E. 2006. New metrics for comparative genomics. *Current Opinion in Biotechnology*, 17:440–7.

Gascoel, O. and Steel, M. 2006. Neighbor-joining revealed. *Molecular Biology and Evolution*, 23:1997–2000.

Grünewald, S., Forstlund, K., Dress, A. and Moulton, V. 2006. QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. *Molecular Biology and Evolution*, 24:532–8.

Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K. and Schuster, S.C. 2005. Whole-genome prokaryotic phylogeny. *Bioinformatics*, 21:2329–33.

Huson, D. 1998. Splitstree- a program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14:68–73.

Kalmanson, K. 1975. Edgeconvex circuits and the traveling salesmen problem. *Canadian Journal of Mathematics*, 27:1000–10.

Karlín, S. and Alpschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings National Academy of Sciences U.S.A.*, 87:2264–8.

Korbel, J.O., Snél, B., Huynen, M.A. and Bork, P. 2002. SHOT: a web server for the construction of genome phylogenies. *Trends Genetics*, 18:518–62.

Korostensky, C. and Goumet, G.H. 2000. Using traveling salesman algorithms for evolutionary tree construction. *Bioinformatics*, 16:619–27.

Kraskov, A., Stögbauer, H., Andrzezej, R.G. and Grassberger, P. 2005. Hierarchical clustering using mutual information. *Europhysics Letter*, 70:278–84.

Kunin, V., Goldovsky, L., Darzentas, N. and Ouzounis, C.A. 2005a. The net of life: reconstructing the microbial phylogenetic network. *Genome Research*, 15:954–9.

Kunin, V., Ahren, D., Goldovsky, L., Janssen, P. and Ouzounis, C.A. 2005b. Measuring genome conservation across taxa: divided strains and united kingdoms. *Nucleic Acids Research*, 33(2):616–21. http://cgg. ebi.ac.uk/cgi–bin/gps/GPS.pl

Lin, J. and Gerstein, M. 2007. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different recent levels. *Genome Research*, 2008:808–18.

Makarenkov, V. and Leclerc, B. 1997. Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees. In: *Mirkin B., Morris FR., Roberts F, Rzhetsky A.*, eds. *Mathematical hierarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. Providence: Amer. Math. Soc., 183–208.

Makarenkov, V. and Leclerc, B. 2000. Comparison of additive trees using circular orders. *J. Computational Biol.*, 7:31–44.

Makarenkov, V., Kevorkov, D. and Legendre, P. 2006. Phylogenetic network construction approaches. Applied Mycology and Biotechnology, International Elsevier Series, vol. 6. *Bioinformatics*, 61–97.

Mihæscu, R., Levy, D. and Pachter, L. 2006. Why neighbourhood joining works. arXiv cs.DS/0602041, Accessed 20 Mai 2007, http://arxiv.org/PS_cache/cs/pdf/0602/0602041v3.pdf

Nakhleh, L., Warnow, T. and Linder, C.R. 2004. Reconstructing Reticulate Evolution in Species- Theory and Practice. *Recomb’04* March 27–31 2004, San Diego: ACM337–46.

Pauplin, Y. 2000. Direct calculation of a tree length using a distance matrix. *J. Mol. Biol.*, 51:41–7.

Robinson, W. 1951. A method for chronologically ordering archaeological deposits. *American Antiquity*, 16:293–301.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for constructing phylogenetic trees. In: *Saitou, N. and Nei, M.*, eds. *Molecular Biology and Evolution*, 4:406–25.

Snel, B., Bork, P. and Huynen, M.A. 1999. Genome phylogeny based on gene content. *Nature Genetics*, 21:108–10.

Thuillard, M. 2001. Wavelets in Soft Computing. Singapore: World Scientific.

Thuillard, M. 2004. Adaptive multiresolution search: how to beat brute force? *International Journal Approximate Reasoning*, 35(3):223–38.

Thuillard, M. 2007. Minimizing contradictions on circular order of phylogenetic trees. *Evolutionary Bioinformatics*, 3:267–77.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y. and Martin, W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, 5:123–35.

Wang, L.S., Warnow, T., Moret, B.E.M., Jansen, R.K. and Raubeson, L.A. 2006. Distance-based Genome Rearrangement Phylogeny. *Journal of Molecular Evolution*, 63:473–83.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V. and Koonin, E.V. 2002. Genome trees and the tree of life. *Trends Genetics*, 18:472–9.

Yang, S., Doolittle, R.F. and Boume, P.E. 2005. Phylogeny determined by protein domain content. *Proceedings National. Academy of Sciences U.S.A.*, 102:373–8.

Yushmanov, S.V. 1984. Construction of a tree with p leaves from 2p–3 elements of its distance matrix (Russian). *Matematicheskie Zametki*, 35:877–87.
Annex: List of Taxa Corresponding to the Best Order of Figure 3

| Taxon                           | Accession or Strain          |
|--------------------------------|-----------------------------|
| 1. RTYP-144-01-Rickettsia typhi ATCC VR-144 |                          |
| 2. RPRO-MAD-01-Rickettsia prowazekii Madrid E |                          |
| 3. RCON-MAL-01-Rickettsia conorii str. Malish 7 |                          |
| 4. WPIP-WME-01-Wolbachia piipientis wMel |                          |
| 5. BJAP-USD-01-Bradyrhizobium japonicum USDA110 |                          |
| 6. RPAL-009-01-Rhodopseudomonas palustris CGA009 |                          |
| 7. RCON-MAL-01-Rickettsia conorii str. Malish 7 |                          |
| 8. BQUI-TOU-01-Bartonella quintana Toulouse |                          |
| 9. BHEN-HOU-01-Bartonella henselae Houston-1 |                          |
| 10. BMEL-M16-01-Brucella melitensis M16 |                          |
| 11. BSUI-133-01-Brucella suis str. 1330 |                          |
| 12. ATUM-C58-01-Agrobacterium tumefaciens C58 |                          |
| 13. SMEL-102-01-Sinorhizobium meliloti strain 1021 |                          |
| 14. CCRE-XXX-01-Caulobacter crescentus CB15 |                          |
| 15. XAXO-306-02-Xanthomonas axonopodis pv. citri str. 306 |                          |
| 16. XCAM-AT3-01-Xanthomonas campestris pv. campestris ATCC33913 |                          |
| 17. XVAS-XPD-01-Xyella fastidiosa PD |                          |
| 18. XFAS-9A5-01-Xyella fastidiosa 9A5 |                          |
| 19. NEUR-718-01-Nitrosomonas europaea ATCC19718 |                          |
| 20. CVIO-472-01-Chromobacterium violaceum ATCC12472 |                          |
| 21. NMEN-Z24-01-Neisseria meningitidis Z2491 |                          |
| 22. NMEN-MC5-01-Neisseria meningitidis MC58 |                          |
| 23. BPER-251-01-Bordetella pertussis NCTC-13251 |                          |
| 24. BBRO-252-01-Bordetella bronchiseptica NCTC-13252 |                          |
| 25. BPAR-253-01-Bordetella parapertussis NCTC-13253 |                          |
| 26. RSOL-XXX-01-Ralstonia solanacearum |                          |
| 27. PSYR-DC3-01-Pseudomonas syringae pv. tomato DC3000 |                          |
| 28. PPUT-KT2-01-Pseudomonas putida KT2440 |                          |
| 29. PAER-PAO-01-Pseudomonas aeruginosa PAO1 |                          |
| 30. CBUR-RSA-01-Coxiella burnetii RSA 493 |                          |
| 31. WGLO-BRE-01-Wiglesworthia glossinidia brevipalpis |                          |
| 32. BAPH-XSG-01-Buchnera aphidicola SG |                          |
| 33. BUCH-APS-01-Buchnera sp. APS |                          |
| 34. BAPH-XBP-01-Buchnera aphidicola Bp |                          |
| 35. BFLO-XXX-01-Blochmannia floruman |                          |
| 36. ECAR-043-01-Erwinia carotovora subsp. atroseptica SCR1043 |                          |
| 37. YPES-COH-01-Yersinia pestis CO92 |                          |
| 38. YPES-KIM-01-Yersinia pestis KIM |                          |
| 39. SFLE-457-01-Shigella flexneri 2457T |                          |
| 40. SFLE-301-01-Shigella flexneri 301 |                          |
| 41. ECOL-RIM-01-Escherichia coli 0157H7 RIMD0509952 |                          |
| 42. ECOL-EDL-01-Escherichia coli O157H7 EDL933 |                          |
| 43. ECOL-MGI-01-Escherichia coli MG1655 |                          |
| 44. ECOL-CFT-01-Escherichia coli CFT073 |                          |
| 45. SENT-CT1-02-Salmonella enterica Typhi CT18 |                          |
| 46. SENT-LT2-01-Salmonella enterica serovar Typhi LT2 |                          |
| 47. PLUM-TO1-01-Photorhabdus luminescens TTO1 |                          |
| 48. HINF-KW2-01-Haemophilus influenzae KW20 |                          |
| 49. PMUL-P7-01-Pasteurella multocida Pm70 |                          |
| 50. VCHO-N16-01-Vibrio cholerae El Tor N19691 |                          |
| 51. VVUL-YJ0-01-Vibrio vulnificus YJ016 |                          |
| 52. VPAR-RIM-01-Vibrio parahaemolyticus RIMD2210633 |                          |
| 53. SONE-MR1-01-Shewanella oneidensis MR1 |                          |
| 54. CJEJ-NCT-01-Campylobacter jejuni NCTC11168 |                          |
| 55. HPYL-266-01-Helicobacter pylori 26695 |                          |
| 56. HPYL-J99-01-Helicobacter pylori J99 |                          |
| 57. HHEP-449-01-Helicobacter hepaticus ATCC51449 |                          |
| 58. WSUC-740-01-Wolinella succinogenes strain DSM 1730 |                          |
| 59. AAEO-VF5-01-Aquifex aeolicus VF5 |                          |
| 60. GSUL-PCA-01-Geobacter sulfurreducens PCA |                          |
| 61. DVUL-HIL-01-Desulfovibrio vulgaris strain Hildenborough |                          |
| 62. BBAC-100-01-Desulfovibrio bacteriovorus HD100 |                          |
| 63. CTEP-TLS-01-Chlorobium tepidum TLS |                          |
| 64. BTHE-VPI-01-Bacteroides thetaiotaomicron VPI5482 |                          |
| 65. PGIN-W83-01-Porphyrmonas gingivalis W83 |                          |
| 66. BBUR-B31-01-Brorrelia burgdorferi B31 |                          |
| 67. TDEN-405-01-Treponema denticola ATCC35405 |                          |
| 68. TPAL-NIC-01-Treponema pallidum Nichols |                          |
| 69. LINT-566-01-Leptospira interrogans strain H130 |                          |
| 70. LINT-130-01-Leptospira interrogans L1-130 |                          |
| 71. TMAR-MSB-01-Thermotoga maritima MSB8 |                          |
| 72. FNUC-ATC-01-Fusobacterium nucleatum ATCC25586 |                          |
| 73. TTEN-MB4-01-Thermoanaerobacter tengcongensis MB4 |                          |
| 74. CTET-E88-01-Clostridium tetani E88 |                          |
| 75. CPER-X13-01-Clostridium perfringens str. 13 |                          |
| 76. CACE-ATC-01-Clostridium acetobutylicum ATCC824 |                          |
| 77. LLAC-IL1-01-Lactococcus lactis IL1403 |                          |
| 78. SMUT-UA1-01-Streptococcus mutans UA159 |                          |
| 79. SAGO-260-01-Streptococcus agalactiae 2603 V/R |                          |
| 80. SAGO-NEM-01-Streptococcus agalactiae NEM316 |                          |
| 81. SPYO-SF3-01-Streptococcus pyogenes M1 SF370 |                          |
| 82. SPYO-MGA-01-Streptococcus pyogenes M18 MGAS8232 |                          |
| 83. SPYO-XM3-01-Streptococcus pyogenes M3 MGAS315 |                          |
| 84. SPYO-SSI-01-Streptococcus pyogenes M3 SSI-1 |                          |
| 85. SPYO-394-01-Streptococcus pyogenes MGAS10394 |                          |
| 86. SPNE-TIG-01-Streptococcus pneumoniae TIGR4 |                          |
| 87. SPNE-XR6-01-Streptococcus pneumoniae R6 |                          |
| 88. EFAE-V58-01-Enterococcus faecalis V583 |                          |
| 89. LPLA-WCF-01-Lactobacillus plantarum WCFS1 |                          |
| 90. LJOH-533-01-Lactobacillus johnsonii NCC533 |                          |
| 91. LMON-365-01-Listeria monocytogenes F2365 |                          |
| 92. LMON-819-01-Listeria monocytogenes H7858 |                          |
| 93. LMON-854-01-Listeria monocytogenes F6854 |                          |
| 94. LMON-EGD-01-Listeria monocytogenes EGD-e |                          |
| 95. SAUR-476-01-Staphylococcus aureus MSSA476 |                          |
| 96. SAUR-MW2-01-Staphylococcus aureus MRSA MW2 |                          |
| Minimum contradiction matrices in whole genome phylogenies |
|-----------------------------------------------------------|
| 99. SAUR-N13-01-Staphylococcus aureus MRSA N315 |
| 100. SAUR-MU5-01-Staphylococcus aureus VRSA Mu50 |
| 101. SAUR-252-01-Staphylococcus aureus MRSA252 |
| 102. BSUB-168-01-Bacillus subtilis 168 |
| 103. BANT-AME-01-Bacillus anthracis Ames |
| 104. BCER-987-01-Bacillus cereus ATCC 10987 |
| 105. BCER-579-01-Bacillus cereus ATCC 14579 |
| 106. OIHE-HET-01-Oceanobacillus iheyensis HET831 |
| 107. BHAL-C12-01-Bacillus halodurans C-125 |
| 108. MMYC-G1T-01-Mycoplasma mycoides subsp. mycoplasma SC strain PG1 |
| 109. MOMB-63K-01-Mycoplasma mobile 163K |
| 110. MPUL-UAB-01-Mycoplasma pulmonis UAB CTIP |
| 111. UURE-SV3-01-Ureaplasma urealyticum serovar 3 |
| 112. MGEN-G37-01-Mycoplasma genitalium G-37 |
| 113. MPNE-M12-01-Mycoplasma pneumoniae M129 |
| 114. MGAL-RLO-01-Mycoplasma penetrans HF2 |
| 115. PAST-XOY-01-Phytoplasma asteris OY |
| 116. CPAR-TII-01-Cryptosporidium parvum typeII |
| 117. PFAL-3D7-01-Plasmodium falciparum 3D7 |
| 118. ECUN-XXX-01-Encephalitozoon cuniculi |
| 119. NCRA-XX3-01-Neurospora crassa |
| 120. YLIP-B99-01-Yarrowia lipolytica CLIB99 |
| 121. AGOS-XXX-01-Ashbya gossypii |
| 122. KLAC-210-01-Kluyveromyces lactis CLIB210 |
| 123. SCER-S28-01-Saccharomyces cerevisiae S288C |
| 124. CGLA-138-01-Candida glabrata CBS138 |
| 125. DHAN-767-01-Debaryomyces hansenii CBS767 |
| 126. SPOM-XXX-01-Schizosaccharomyces pombe |
| 127. ATHA-XXX-01-Arabidopsis thaliana |
| 128. CMER-10D-01-Cyanidioschyzon merolae 10D |
| 129. CBRi-XXX-01-Caenorhabditis briggsae |
| 130. CELE-XXX-01-Caenorhabditis elegans |
| 131. DMEL-XXX-01-Drosophila melanogaster |
| 132. AGAM-PES-01-Anopheles gambiae PEST |
| 133. HSAP-XXX-03-Homo sapiens v15.33.1 |
| 134. MMUS-XXX-02-Mus musculus |
| 135. NEQU-N4M-01-Nanoarchaeum equitans Kin4-M |
| 136. APER-XK1-01-Aeropyrum pernix K1 |
| 137. PAER-IM2-01-Pyrobaculum aerophilum IM2 |
| 138. STOK-XX7-01-Sulfolobus tokodaii str. 7 |
| 139. SSOL-XP2-01-Sulfolobus solfataricus P2 |
| 140. TAC1-DSM-01-Thermoplasma acidophilum DSM1728 |
| 141. TVOL-GSS-01-Thermoplasma volcanium GSS1 |
| 142. PTOR-790-01-Picrophilus torridus DSM9790 |
| 143. PABY-GE5-01-Pyrococcus abyssi GE5 |
| 144. PHOR-OT3-01-Pyrococcus horikoshii OT3 |
| 145. MTHE-DEL-01-Methanobacterium thermoautotrophicum deltaH |
| 146. MJAN-DSM-01-Methanococcus jannaschii DSM 2661 |
| 147. MTUB-H37-01-Mycobacterium tuberculosis H37Rv |
| 148. MBOV-AF2-01-Mycobacterium bovis AF2122/97 |
| 149. MLEP-XTN-01-Mycobacterium leprae leprae |
| 150. CEFF-Y3S-01-Corynebacterium efficiens YS314T |
| 151. CDIP-129-01-Corynebacterium diphtheriae |
| 152. PACN-202-01-Propionibacterium acnes KPA171202 |
| 153. LXYL-B07-01-Leifsonia xyli subsp. xyli CTCB07 |
| 154. THWI-TW1-01-Tropheryma whipplei Twist |
| 155. THWI-TW0-01-Tropheryma whipplei TW02/27 |
| 156. SAVE-XXX-01-Steptomyces avermitilis |
| 157. TTHE-B27-01-Thermus thermophilus HB27 |
| 158. DRAD-XXR-01-Deinococcus radiodurans R1 |
| 159. GIVIO-421-01-Gloeobacter violaceus PCC 7421 |
| 160. SYNE-PCC-01-Synechocystis sp. PCC6803 |
| 161. TELO-BP1-01-Thermosynechococcus elongatus BP-1 |
| 162. NOST-PCC-01-Anabaena sp. strain PCC 7120 |
| 163. PMAR-SS1-01-Prochlorococcus marinus SS120 |
| 164. PMAR-MED-01-Prochlorococcus marinus MED4 |
| 165. SYCC-WH8-01-Synechococcus sp. WH8102 |
| 166. PMAR-MIT-01-Prochlorococcus marinus MIT9313 |
| 167. PIRE-ST1-01-Pirellula sp. strain 1 |
| 168. PCHL-E25-01-Parachlamydia sp. UWE25 |
| 169. CCADV-GPI-01-Chlamydia pneumoniae caviae GPIC |
| 170. CPNE-J13-01-Chlamydia pneumoniae J138 |
| 171. CPNE-CWL-01-Chlamydia pneumoniae CWL029 |
| 172. CPNE-AR3-01-Chlamydia pneumoniae AR39 |
| 173. CTRA-SVD-01-Chlamydia trachomatis serovar D |
| 174. CTRA-MOP-01-Chlamydia trachomatis serovar D |
| 175. CTRA-SVD-01-Chlamydia trachomatis serovar D |
| 176. CTRA-MOP-01-Chlamydia trachomatis serovar D |

Evolutionary Bioinformatics 2008:4