Whole-genome prokaryotic phylogeny

Stefan R. Henz¹, Daniel H. Huson¹,∗, Alexander F. Auch¹, Kay Nieselt-Struwe¹ and Stephan C. Schuster²

¹Center for Bioinformatics Tübingen (ZBIT), Sand 14, Tübingen 72076, Germany and ²Max-Planck-Institute for Development Biology, Spemannstrasse 35, Tübingen, 72076, Germany

Received on September 26, 2003; revised on March 5, 2004; accepted on April 8, 2004

ABSTRACT

Current understanding of the phylogeny of prokaryotes is based on the comparison of the highly conserved small ssu-rRNA subunit and similar regions. Although such molecules have proved to be very useful phylogenetic markers, mutational saturation is a problem, due to their restricted lengths. Now, a growing number of complete prokaryotic genomes are available. This paper addresses the problem of determining a prokaryotic phylogeny utilizing the comparison of complete genomes. We introduce a new strategy, GBDP, ‘genome blast distance phylogeny’, and show that different variants of this approach robustly produce phylogenies that are biologically sound, when applied to 91 prokaryotic genomes. In this approach, first BLAST is used to compare genomes, then a distance matrix is computed, and finally a tree- or network-reconstruction method such as UPGMA, Neighbor-Joining, BioNJ or Neighbor-Net is applied.

Contact: huson@informatik.uni-tuebingen.de

INTRODUCTION

Our current understanding of the phylogeny of prokaryotes is based on the comparison of the highly conserved small ssu-rRNA subunit (ssu-rRNA) as originally proposed by Carl Woese and colleagues (Woese, 1987). This RNA subunit has all the characteristics of a good phylogenetic marker, namely: (1) it is universally present in all organisms under consideration, (2) it was derived from a common ancestor and thus the sequences are homologous, (3) it is an essential and complex gene and, (4) because of its low rate of substitution, it is genetically very stable.

Attempts to elucidate the phylogeny of prokaryotes based on the ssu-rRNA have been quite successful (Huynen and Bork, 1998; Olsen et al., 1994). However, saturation is a problem due to the restricted length of the molecule and functional restrictions limiting the number of mutable sites. This issue can be addressed to some degree by using additional phylogenetic markers, such as 23S rRNA, the β-subunit of F₁F₀ ATPase, and also the elongation factor Tu (Ludwig and Schleifer, 1999, http://www.vermicon.de/english/news/science/KHS99111.htm).

Another well-known problem associated with this type of approach is that the evolutionary history of any single gene may differ from the phylogenetic history of the whole organism from which the corresponding molecule was isolated.

∗To whom correspondence should be addressed.

Generally, the resolution of deep divergences such as the phylogeny of the archae and (eu)bacteria is very difficult since information fades as one goes deeper into the past. Indeed, it has been shown that the sequence requirement grows exponentially with time as one attempts to resolve deeper and deeper divergences (Sober and Steel, 2002). Therefore, it seems important to make use of as much genetic information as available for the reconstruction of the prokaryotic phylogeny.

Now, a growing number of complete prokaryotic genomes is available and the question arises how to derive phylogenies based on the whole genomic information of organisms rather than based on a small number of genes (Wolf et al., 2001; Mirkin et al., 2003). One such approach is to determine a genome phylogeny based on gene content (Snel et al., 1999, Huson and Steel, submitted for publication).

In this paper, we introduce a new strategy, called GBDP, ‘genome blast distance phylogeny’, that combines different pieces of standard methodology to produce a phylogenetic tree or network from a given set of whole-genome data. We have applied this approach to 91 prokaryotic genomes, of which 90 were obtained from NCBI (NCBI, 2003, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_micr.html) and EBI (EBI, 2003, http://www.ebi.ac.uk/cgibin/genomes/genomes.cgi?genomes=Bacteria) and one additional genome (Wolinella succinogenes) that was recently sequenced in the laboratory of the last-named author (Baar et al., 2003).

More precisely, GBDP starts with an all-against-all pairwise comparison of genomes using BLASTN (Altschul et al., 1990). In a second step, a distance matrix is calculated from the resulting HSPs (High Speed Products). Here, we studied a number of variants which are described in detail below. This distance matrix is then processed by a distance-based phylogenetic method to produce a phylogenetic tree or network. We considered UPGMA (Sokal and Michener, 1958), Neighbor-Joining (NJ) (Saitou and Nei, 1987; Studier and Keppler, 1988), BioNJ (Gascuel, 1997), and NEIGHBORNET (Bryant and Moulton, 2002).

Comparison of our results with the NCBI taxonomy (NCBI, 2003, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_micr.html) (Fig. 4) and the phylogeny reported in (Woese, 2000) shows that GBDP is a robust method for determining the phylogeny of a collection of whole genomes. Of the four reconstruction methods considered, BioNJ and NEIGHBORNET produce the most biologically sound phylogenies. Additionally, the latter method can lead to useful insights as the reticulations in the NEIGHBORNET graph explains, in
isolated cases, why some individual taxa are placed ambiguously in a number of the phylogenies reported.

**METHODS**

**Phylogenetic networks**

The deep divergence of the prokaryotes makes the resolution of the different bacterial groups a difficult task and several of the inter-species and intra-species branching orders remain unclear.

In theory, the representation of the ambiguous branching orders by a phylogenetic network rather than a tree should be helpful here. One of the most popular methods for computing phylogenetic networks, in this case so-called splits graphs, is split decomposition (Bandelt and Dress, 1992; Huson, 1998).

Splits graphs, is split decomposition (Bandelt and Dress, 1992; Huson, 1998). In such a splits graph, bipartitions or splits of the taxa are represented by bands of parallel edges and the graph can be interpreted as representing multiple phylogenetic branching patterns simultaneously. Unfortunately, the split decomposition method is very conservative and so, given the level of diversity studied here, the resulting graph is almost completely unresolved. The deep divergence of the prokaryotes makes the resolution of the different bacterial groups a difficult task and several of the inter-species and intra-species branching orders remain unclear.

In our study, the phylogenetic network produced by NeighborNet (Bryant and Moulton, 2002) is a new agglomerative algorithm for computing (planar) splits graphs that produces a much better resolution. In theory, the representation of the ambiguous branching orders by a phylogenetic network rather than a tree should be helpful here. One of the most popular methods for computing phylogenetic networks, in this case so-called splits graphs, is split decomposition (Bandelt and Dress, 1992; Huson, 1998).

In such a splits graph, bipartitions or splits of the taxa are represented by bands of parallel edges and the graph can be interpreted as representing multiple phylogenetic branching patterns simultaneously. Unfortunately, the split decomposition method is very conservative and so, given the level of diversity studied here, the resulting graph is almost completely unresolved. NeighborNet (Bryant and Moulton, 2002) is a new agglomerative algorithm for computing (planar) splits graphs that produces a much better resolution.

In our study, the phylogenetic network produced by NeighborNet (Fig. 3) indicates, in some cases, why some individual taxa are ambiguously placed in the trees reported. For example, the *Neisseria meningitidis* sequences are placed differently in all three phylogenetic trees. However, two significant splits separating the group in the network produced by NeighborNet clearly indicate the ambiguous evolutionary signal. Another example is the grouping of *Borrelia burgdorferi* and *Fusobacterium nucleatum* within the Firmicutes in all three trees. NeighborNet, too, clusters them within the Firmicutes group and, again, conflicting splits indicate the uncertainty of this position.

It is also interesting to note that much fewer conflicting splits appear in the archaeal phylogeny than in the bacterial phylogeny.

**The GBDP strategy**

The first step in the GBDP strategy is an all-against-all pairwise comparison of all genomes using a sequence comparison method such as BLAST (Altschul et al., 1990). For any two genomes X and Y, this produces a list of high-scoring segment pairs (HSPs), each consisting of a pair of sequence segments in X and Y (or in either of the opposite strands) of very similar length and whose significance is indicated by an E-value and/or score.

In our study, we used standard nucleotide–nucleotide BLASTN. This is the most time consuming part of the process and given 91 prokaryotic genomes, this step took about 170 CPU hours. We are currently evaluating the use of tBLASTX for this step.

The second step of the GBDP strategy is the application of a distance transformation that computes a distance between any two genomes, whose significance is indicated by an E-value and/or score. This is the most time consuming part of the process and given 91 prokaryotic genomes, this step took about 170 CPU hours. We are currently evaluating the use of tBLASTX for this step.

The second step of the GBDP strategy is the application of a distance transformation that computes a distance between any two genomes X and Y, based on the set of HSPs obtained in the first step. The first distance scoring segment pairs (HSPs), each consisting of a pair of sequence segments in X and Y (or in either of the opposite strands) of very similar length and

\[ d_{cov}(X, Y) := -\log \left( \frac{|X_{cov}| + |Y_{cov}|}{|X| + |Y|} \right), \quad (1) \]
where \(|X|\) and \(|Y|\) denote the length of genomes \(X\) and \(Y\), respectively, and \(|X_{\text{cov}}|\) and \(|Y_{\text{cov}}|\) denote the number of base pairs in \(X\) and \(Y\), respectively, that are covered by some HSP between \(X\) and \(Y\). The simple rationale behind Equation (1) is that more closely related genomes will be more covered by a larger number of HSPs.

However, we observed that application of this transformation can lead to a problematic relative placement of two genomes when one genome is essentially a subset of a much larger second one. For example, in our data set this is true of a *Buchnera* genome, which is essentially a subset of the *Escherichia coli* genome, approximately 14% the size of it (Moran and Mira, 2001). Although a substantial amount of *Buchnera* is covered by HSPs between *Buchnera* and *E.coli*, *E.coli* has a large amount of additional DNA that remains uncovered, thus obscuring the subset relationship. To address this problem, we introduce a modification of the first distance transformation:

\[
d'_{\text{cov}}(X, Y) := -\log \frac{|X_{\text{cov}}| + |Y_{\text{cov}}|}{2 \min(|X|, |Y|)}
\]

In practice, this transformation performs slightly better than the first one and throughout this paper, the distance matrix obtained using (2) is referred to as the coverage distance.

Repeats of significant size or number can result in a coverage distance that makes two different genomes seem to be more closely related than they actually are. To address this problem, we consider a second distance transformation, called the matched distance. This is obtained by selecting a maximum subset of HSPs that are non-overlapping both in the \(X\) and in the \(Y\) sequence. In other words, the set of selected HSPs has the property that any nucleotide position in either genome is contained in at most one selected HSP.

Moreover, we detected little difference between the greedy and greedy-with-trimming approaches. Throughout this paper, the term matched distances will denote the distances obtained using the greedy selection by match length.
The choice whether to use the coverage distance or matched distance can make quite a difference. For example, a BLASTN comparison of *N. meningitidis* serogroup A and *N. meningitidis* serogroup B gives rise to 350,000 HSPs that are clearly repeat induced (Parkhill et al., 2000). In this case, the covered distance is 0.0538, whereas the matched distance is 0.182.

A different distance transformation is based on the concept of breakpoints (Sankhoff and Blanchette, 1997; Sankhoff et al., 2000; Wang et al., 2003, http://www.smi.stanford.edu/projects/helix/pbs02/wang.pdf). A breakpoint can be detected when two HSPs (from a set of non-overlapping HSPs obtained, e.g. by one of greedy heuristics described above) map consecutive intervals in one genome onto non-consecutive intervals in the other. More precisely, we define a breakpoint as follows: Consider a HSP $h_1$ covering the interval $[X_1, X_2]$ in genome $X$ and the interval $[Y_1, Y_2]$ in genome $Y$. Now, consider the next interval in $X$ that is covered by an HSP $h_2$, with coordinates $[X_2, X_3]$. (For this discussion we assume that all matches are in the forward strands of both genomes. We omit the discussion of the other cases, which is straightforward but slightly tedious.) Let $[Y_2, Y_3]$ denote the corresponding interval in $Y$. We count a breakpoint, if there exists a third, intervening HSP $h_3$ that covers a position between $Y_2$ and $Y_3$ and thus destroys the co-linearity of $h_1$ and $h_2$. We define the breakpoint transformation as

$$d_{break}(X, Y) = -\log \left(1 - \frac{B_X + B_Y}{M_X + M_Y}\right),$$

where $B_X$, or $B_Y$, is the number of breakpoints in $X$, or $Y$, respectively, and $M_X$, or $M_Y$, is the number of matched intervals in $X$, or $Y$, respectively.

In our study, the breakpoint transformation performed very poorly within the framework of GBDP (not shown here). This is perhaps not surprising as the order of the matching segments of two prokaryotic genomes is usually extremely different, whereas it is commonly assumed that breakpoint methods lead to good results only if the genomes have preserved a substantial amount of co-linearity.

By definition, BLAST is not symmetric (Altschul et al., 1990) and this potentially leads to a number of different variants of GBDP. Let $d(X, Y)$ denote a distance based on blasting genome $X$ against $Y$. We can define a distance matrix in three different ways, namely, we can use the average $D(X, Y) := \frac{d(X, Y) + d(Y, X)}{2}$ or use one of the two directions, either $d(X, Y) := D(Y, X)$ or $d(Y, X) := D(Y, X)$.

Our studies seem to indicate that this directionality does have some impact on the performance of GBDP methods and averaging over both directions produces better results. Thus, all results shown here are based on the latter method.

The third and final step of the GBDP strategy consists of applying a standard distance-based phylogenetic tree- or network reconstruction method. In our study, we used UPGMA (Sokal and Michener, 1958), NJ (Saitou and Nei, 1987; Studier and Keppler, 1988), as implemented in the release 3.57c of the Phylip package (Felsenstein, 1989), and BioNJ (Gascuel, 1997), in order to reconstruct phylogenetic trees. We used two different network methods, split decomposition (Bandelt and Dress, 1992) as implemented in Huson and Bryant, (manuscript in preparation) and NeighborNet (Bryant and Moulton, 2002) as implemented in Huson and Bryant, (manuscript in preparation) to compute phylogenetic networks.
Comparison of variants of GBDP

In our studies we compared the topology of the trees computed with the topology of the NCBI taxonomy (NCBI, 2003, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_micr.html), which we regard as the ‘correct’ one, although it is not totally resolved. In the following, we will refer to this as the NCBI tree, (Fig. 4).

We could compare any computed phylogenetic tree or network with the NCBI tree $T_0$ simply by counting the number of non-trivial splits in $T$ that are not contained in $T_0$, and vice versa, thus obtaining a number of false positives and false negatives. However, because the NCBI tree $T_0$ is not fully resolved, proceeding in this way would over-count false positives. To avoid this problem, we only count a non-trivial split from $T_0$ as a false positive, if it is missing from $T_0$ and it is incompatible with $T_0$ (Bandelt and Dress, 1992).

For purposes of a first evaluation, we define the compatibility-score ($c$-score) of the reconstruction as:

$$c\text{-score} := \frac{\Sigma(T_{\text{comp}})}{\Sigma(T)}, \quad (5)$$

where $\Sigma(T)$ and $\Sigma(T_{\text{comp}})$ denote the set of all non-trivial splits in $T$, and the set of all non-trivial splits in $T$ that are compatible with $T_0$, respectively.

We used this $c$-score to evaluate how well a reconstructed tree or network might reflect the ‘correct’ biological evolution. The matched distances variants of GBDP produced the highest $c$-scores and the resulting phylogenies are displayed in Figures 1–3. The $c$-scores for a number of different variants of GBDP are summarized in Table 1. Note that the $c$-scores reported for NEIGHBORGEO are very low. This reflects the fact that the data is not very tree-like and the computed phylogenetic networks display range of conflicting phylogenies, thus necessarily producing false positives.

DISCUSSION

In recent years the determination-driven classification of microbial organisms has been given up in order to make room for a phylogenetic system (Woese, 1987; NCBI, 2003, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_micr.html). The most valuable phylogenetic marker for the establishment of a general phylogenetic system turned out to be the ssu-rRNA. Using this marker, the three domains of life are recognizable, namely archa, eukaryotes and bacteria. The aim of this study is to extend the set of tools available for the phylogenetic comparison of prokaryotes beyond the already published ssu-rRNA (Woese, 1987) and gene content approaches (Snel et al., 1999, Hason and Steel, 2004).

Using the GBDP strategy, i.e. a genome wide homology search using BLASTN followed by a distance matrix computation and tree construction (see below for algorithmic details), we were able to construct phylogenetic trees and networks that resemble the previously published phylogenies in their most important aspects, with some noteworthy deviations.
with the current NCBI taxonomy

<table>
<thead>
<tr>
<th>Distance transformation</th>
<th>Phylogenetic method</th>
<th>c-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>covered</td>
<td>UPGMA</td>
<td>0.716</td>
</tr>
<tr>
<td>greedy length</td>
<td>UPGMA</td>
<td>0.716</td>
</tr>
<tr>
<td>greedy score</td>
<td>UPGMA</td>
<td>0.716</td>
</tr>
<tr>
<td>greedy E-value</td>
<td>UPGMA</td>
<td>0.716</td>
</tr>
<tr>
<td>covered</td>
<td>NJ</td>
<td>0.705</td>
</tr>
<tr>
<td>greedy length</td>
<td>NJ</td>
<td>0.716</td>
</tr>
<tr>
<td>greedy score</td>
<td>NJ</td>
<td>0.716</td>
</tr>
<tr>
<td>greedy E-value</td>
<td>NJ</td>
<td>0.716</td>
</tr>
<tr>
<td>covered</td>
<td>NeighborNet</td>
<td>0.4</td>
</tr>
<tr>
<td>greedy length</td>
<td>NeighborNet</td>
<td>0.433</td>
</tr>
<tr>
<td>greedy score</td>
<td>NeighborNet</td>
<td>0.433</td>
</tr>
<tr>
<td>greedy E-value</td>
<td>NeighborNet</td>
<td>0.433</td>
</tr>
</tbody>
</table>

Table 1. The c-score of different GBDP phylogenies, based on a comparison with the current NCBI taxonomy

In the first column, the words covered and greedy indicate which distance transformation was used, i.e., the covered distances, Equation (2), or matched distances, Equation (3), respectively. In the latter case, length, score, and E-value indicates which variable the greedy selection was based on. The second column indicates the phylogenetic reconstruction method used. The third column reports the c-score as defined in Equation (5).

Most of the variants of GBPD that we studied identify the archaean domain as a monophyletic unit. (Figs 1–4). Furthermore, the topology within this domain described by the ssu-rRNA approach is also observable. Such a clustering of extreme thermophile, methanogen or extreme halophile archaean species can be demonstrated in the presented phylogenetic models. A major difference between both phylogenies can be observed for organisms that branch off early from the bacterial phylum, such as the Thermotogales and the Aquificales. Both are found in the GBPD approach as branching off from the archaean phylum, despite being considered (for good reasons) as bacterial organisms.

In contrast, other deep branching bacterial phyla such as the Deinococci, the Green bacteria, Actinobacteria, Spirochaetes and Cyanobacteria are grouping within the bacterial monophyletic branch, albeit from different branch points than those described in the ssu-rRNA phylogenies. Particular robustness with respect to the different variants of GBPD that we considered is displayed by the phyla of the Actinobacteria, Cyanobacteria and Chlamydiae, all considered to be deep branching species. Interestingly, the same pattern can be observed for the group of the e-Proteobacteria, which also do not branch off the general proteobacterial branch, but rather remain distinct.

The vast majority of the remaining species are categorized within the Firmicutes (gram-positive) and the Proteobacteria. For the phylogenetic network constructed by the NeighborNet and matched distance variant of GBDP (Fig. 3, see below for algorithmic details), the entire group Firmicutes is found to cluster nicely together, closely resembling published single-gene phylogenies based on 16S rRNA, 23S rRNA, the β-subunit of F_{0}F_{1}-ATPase, and also the elongation factor (Ludwig and Schleifer, 1999, http://www.vermicon.de/english/news/science/KHS99111.htm). However, the order within the Firmicutes is disturbed by the presence of the two species B. burgdorferi, a Spirochaete, and Enucleatum, a Fusobacterium. This fact is true for all variants of GBPD that we studied, and could therefore be indicative of the need to further substantiate their phylogenetic classification. This notion is supported by the fact that the second Spirochaetae, Treponema pallidum, is rather deep branching and never clusters with B. burgdorferi.

A second group of organisms that does not follow the ssu-rRNA phylogeny are those of the heavily degraded genomes. On one side are the genomes of two Rickettsia species, which originally have been classified within the α-Proteobacteria. Both genomes are known to have recently undergone massive mutation-driven, genome-wide deletions (Andersson et al., 1998). It might therefore be possible that the same mutational saturation that can be observed for ssu-rRNA markers is also affecting the comparison to the other members of the α-Proteobacteria. This might also affect the classification of the two Buchnera species, which in several of the presented phylogenies do not group together with their next closest relative Escherichia coli, a γ-Proteobacterium. In addition to ssu-rRNA based studies, the phylogenetic relationship between Buchnera and E. coli has also been the subject of detailed syntenic studies, which establish the time frame of divergence between the two Buchnera strains to be 50 million years (Tamas et al., 2000).

The group of organisms comprising of Haemophilus, Pasteurella, Escherichia, Salmonella and Yersinia form a robust phylum within the γ-Proteobacteria that holds up in all phylogenies presented. The larger group of γ-Proteobacteria, which also includes Xanthomonas, Xylella and Pseudomonas, constitutes a more diffuse phylogeny, which is interspersed by the β-Proteobacteria N.meningitidis and Ralstonia solanacearum. However, this positioning of the β-Proteobacteria in the cluster of γ-Proteobacteria also appears to be robust among the various results.

An ambiguous grouping is also found for the firmucute Thermoaerobacter tengcongensis, which clusters in three of four methods (NJ, BIONJ and NeighborNet) with other thermophile species, albeit archaean or bacterial. This might reflect a genomic convergence that is driven by environmental factors (temperature) rather than a shared ancestral history.

An interesting observation of the study is that the phylogenies computed for the archaean phylum are much better in agreement with the ones from the ssu-rRNA analysis than those from the bacterial phylum. Biological reasons for this may be found in the much slower growth rate of archaean organisms that leads to a slower rate of mutation, as well as a lower degree of lateral gene transfer compared to bacterial species. In how far lateral gene transfer has an influence on the performance of GBDP methods will have to be examined by comparisons in which the flexible moieties of the genomes are omitted.

It is also important to note that the applied methods are based on DNA similarities. Nevertheless, it can be demonstrated that even in cases where the average GC-content of the genome differs by as much as 20% (as between Camphylobacter and Wolinella) a phylogenetically correct placement can be obtained.

In summary, the presented variants of the GBPD strategy are, within limits, robust and recover the general features of the ssu-rRNA phylogeny, which is the molecular basis of the current microbial classification. The observed differences should be used as a starting point both for a reinvestigation of the phylogenetic position of organisms and also for an analysis of the weaknesses of the methods presented.
We found that the accuracy of the trees computed depends on having a sufficiently large sampling of taxa and is therefore expected to increase further as more whole genome sequences become available.

We are currently comparing the proteomes of these species by employing tblASTx in the comparison stage of the GBDP strategy. We suspect that this will yield a small gain in the biological plausibility of the phylogeny obtained, while substantially increasing the computational cost involved.

REFERENCES


