

# GO4genome: A Prokaryotic Phylogeny Based on Genome Organization

Rainer Merkl · Arnim Wiezer

Received: 27 June 2008 / Accepted: 3 April 2009 / Published online: 13 May 2009  
© Springer Science+Business Media, LLC 2009

**Abstract** Determining the phylogeny of closely related prokaryotes may fail in an analysis of rRNA or a small set of sequences. Whole-genome phylogeny utilizes the maximally available sample space. For a precise determination of genome similarity, two aspects have to be considered when developing an algorithm of whole-genome phylogeny: (1) gene order conservation is a more precise signal than gene content; and (2) when using sequence similarity, failures in identifying orthologues or the in situ replacement of genes via horizontal gene transfer may give misleading results. GO4genome is a new paradigm, which is based on a detailed analysis of gene function and the location of the respective genes. For characterization of genes, the algorithm uses gene ontology enabling a comparison of function independent of evolutionary relationship. After the identification of locally optimal series of gene functions, their length distribution is utilized to compute a phylogenetic distance. The outcome is a classification of genomes based on metabolic capabilities and their organization. Thus, the impact of effects on genome organization that are not covered by methods of molecular phylogeny can be studied. Genomes of strains belonging to *Escherichia coli*, *Shigella*, *Streptococcus*, *Methanosarcina*,

and *Yersinia* were analyzed. Differences from the findings of classical methods are discussed.

**Keywords** Whole-genome phylogeny · Bacterial genomes · Computational molecular biology · Prokaryotic phylogenomics · Gene ontology

## Introduction

The classical approach of phylogenetic categorization relies on the analysis of rRNA sequences as introduced by C. Woese (Woese and Fox 1977). However, if the sequences are too similar, it is not possible to determine an evolutionary relationship precisely. This is frequently the case when studying closely related species. For these applications, genome-based phylogenies are superior: the number of mutations separating species will increase with the number of genes analyzed. In addition, methods exploiting a larger number of genes are less affected by horizontal gene transfer (HGT), variable mutation rates, or misalignments (Snel et al. 1999; Fitz-Gibbon and House 1999). For these reasons, phylogenomic methods that use a large set of sequences have become the de facto standard for reconstructing phylogenies (Ciccarelli et al. 2006; Daubin et al. 2002), especially for closely related species (Oshima and Nishida 2007). The algorithms for genome-based phylogeny can be grouped according to their concepts. There are methods that compare genomic DNA sequences on the whole and methods that evaluate gene content or gene order. So far, sequence methods have been used most frequently. In this case, genomes are compared pairwise at the DNA level (Kurtz et al. 2004; Darling et al. 2004). These methods can be extended to construct phylogenetic trees (Henz et al. 2005).

---

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-009-9233-6) contains supplementary material, which is available to authorized users.

---

R. Merkl (✉)  
Institut für Biophysik und Physikalische Biochemie,  
Universität Regensburg, 93040 Regensburg, Germany  
e-mail: rainer.merkl@biologie.uni-regensburg.de

A. Wiezer  
Göttingen Genomics Laboratory, Grisebachstrasse 8,  
37077 Göttingen, Germany

The classical gene content methods are, compared to the above, more complex and require several steps. First, orthology of genes has to be determined. Then the occurrence or absence of genes has to be evaluated and used to infer a phylogenetic tree (see Snel et al. 1999; Baptiste et al. 2004; Lin et al. 2009). Alternatively, the sequences of a small set of genes (Ciccarelli et al. 2006; Konstantinidis et al. 2006) or of a core genome can be concatenated prior to phylogenetic analysis. It has been shown that 20 genes were sufficient for a phylogenetic analysis of eight yeast taxa (Rokas et al. 2003). However, these approaches have been criticized: For a set of 205  $\gamma$ -proteobacterial core genes it has been demonstrated that their history is unknown in many cases and that these genes rarely favor one phylogenetic tree (Susko et al. 2006; Baptiste et al. 2008).

In addition to content, gene order can be utilized to compare genomes. Quite sophisticated theoretical concepts have been developed for the assessment of genomic rearrangements (Sankoff 1992; Hannenhalli et al. 1995) and implemented for their analysis (Tesler 2002; Dalevi and Eriksen 2008). However, to the best of our knowledge, the only approach used so far for comparison of complete microbial genomes is the SHOT server, which exploits the occurrence of gene pairs (Korbel et al. 2002).

One goal of whole-genome comparison is the determination of the true evolutionary distance, i.e., the actual number of mutational events separating two genomes. Unfortunately, this distance cannot be inferred. As a substitute, an edit distance may be computed. The edit distance is the minimal number of evolutionary events selected from a predefined set of operations that transform one genome into the other one. However, even the computation of an edit distance is known to be NP-hard for genomes with unequal content (Xin et al. 2005). Definitely, this time complexity is a severe hindrance for using exact methods to analyze native genomes. Exact methods suffer from a second restriction: these algorithms consider the problem of determining the identity of genes as being solved. As a prerequisite, each gene has to be labeled with a number indicating the orthology class it belongs to. If sequence comparison is the basis to identify orthologues, it may fail: due to gene duplication, several paralogues may exist. In these cases, sequence similarity is no clear indicator of evolutionary relationship.

In addition, it has been proved plausible that HGT is a major force shaping the content of microbial genomes (see, e.g., Ochman et al. 2000 and references therein). Lawrence and Ochman (1997) proposed that at least 15% of the *E. coli* genome is atypical and may have arisen by recent transfer events. It has been concluded that 25% of the *Thermotoga maritima* genes are more closely related to archeal genes and may signal gene transfer between these

lineages (Nelson et al. 1999). For *M. mazei* it has been postulated that up to 30% of the genome may have been acquired via HGT (Deppenmeier et al. 2002). However, the extent and the long-term impact of HGT on individual genomes are still a matter of debate (see, e.g., Kurland et al. 2003).

It has been shown that genes may be replaced in situ with nonorthologous ones (Omelchenko et al. 2003). In these cases, the function of the gene product remains the same, which may not be detectable when comparing sequences. How such an event must be assessed with respect to phylogenetic analysis is debatable.

Due to these arguments, we introduce and apply a new paradigm for genome comparison: we consider genomes as a set of gene series implementing certain functions. The algorithm, which we named GO4genome, is based on the pairwise comparison of gene function and gene order. For assessment of function, it does not consider homology, i.e., evolutionary relationship. Instead, the algorithm utilizes gene ontology. For comparison of gene order, we introduce a heuristic approach. The algorithm identifies the longest series of genes possessing the most similar function. The number and length of these series are then used to compute pairwise genomic distances, which are the basis for phylogenetic inference. Thus, GO4genome comprises additional events like genomic rearrangements for comparison of genomes, which are beyond those exploited by methods of molecular phylogeny. We demonstrate for several groups of microbes that the inferred phylogenetic relationship is, in most cases, in agreement with the outcome of classical methods. A novel grouping of species was observed, e.g., in rapidly evolving genomes like those of *Yersinia pestis* strains or in *Shigella*.

## Materials and Methods

### Datasets

For all analyses, entries downloaded from the Genome Reviews database (<http://www.ebi.ac.uk/GenomeReviews/>) were utilized; it provides comprehensively annotated data, including gene ontology (GO) terms. The following datasets were used (accession numbers in parentheses).

### *Escherichia coli* Dataset

The *E. coli* dataset was comprised of *E. coli* EDL933 (AE005174\_GR.gbk), *E. coli* K-12 (U00096\_GR.gbk), *E. coli* Sakai (BA000007\_GR.gbk), *E. coli* UTI89 (CP000243\_GR.gbk), *E. coli* O1K1 / APEC (CP000468\_GR.gbk), *E. coli* CFT073 (AE014075\_GR.gbk), *E. coli* 536 (CP000247\_GR.gbk), *S. boydii* strain Sb227(CP000036\_GR.

gbk), *S. dysenteriae* strain Sd197 (CP000034\_GR.gbk), *S. flexneri* ATCC 700930 (AE014073\_GR.gbk), *S. flexneri* strain 301 (AE005674\_GR.gbk), *S. flexneri* strain 8401 (CP000266\_GR.gbk), *S. sonnei* strain Ss046 (CP000038\_GR.gbk), *S. typhimurium* LT2 (AE006468\_GR.gbk), *B. aphidicola* (BA000003\_GR.gbk), *R. conorii* (AE006914\_GR.gbk), *R. prowazekii* (AJ235269\_GR.gbk), *Y. pestis* Antiqua (CP000308\_GR.gb), and *A. pernix* (BA000002\_GR.gbk).

#### *Streptococcus* Dataset

The *Streptococcus* dataset included *S. agalactiae* III (AL732656\_GR.gbk), *S. agalactiae* Ia (CP000114\_GR.gbk), *S. agalactiae* V (AE009948\_GR.gbk), *S. mutans* (AE014133\_GR.gbk), *S. pneumoniae* NCTC7466 (CP000410\_GR.gbk), *S. pneumoniae* R6 (AE007317\_GR.gbk), *S. pneumoniae* TIGR4 (AE005672\_GR.gbk), *S. pyogenes* M12 MGAS2096 (CP000261\_GR.gbk), *S. pyogenes* M12 MGAS9429 (CP000259\_GR.gbk), *S. pyogenes* M2 MGAS10270 (CP000260\_GR.gbk), *S. pyogenes* M4 MGAS10750 (CP000262\_GR.gbk), *S. pyogenes* M5 (AM295007\_GR.gbk), *S. pyogenes* M1 MGAS5005 (CP000017\_GR.gbk), *S. pyogenes* M1 ATCC700294 (AE004092\_GR.gbk), *S. pyogenes* M18 MGAS8232 (AE009949\_GR.gbk), *S. pyogenes* M28 MGAS6180 (CP000056\_GR.gbk), *S. pyogenes* M3 MGAS315 (AE014074\_GR.gbk), *S. pyogenes* M3 SSI\_1 (BA000034\_GR.gbk), *S. pyogenes* M6 MGAS10394 (CP000003\_GR.gbk), *S. sanguinis* (CP000387\_GR.gbk), *S. suis* 05ZYH33 (CP000407\_GR.gbk), *S. suis* 98HAH33 (CP000408\_GR.gbk), *S. thermophilus* LMG18311 (CP000023\_GR.gbk), *S. thermophilus* LMD9 (CP000419\_GR.gbk), and *S. thermophilus* CNRZ1066 (CP000024\_GR.gbk).

#### *Methanosarcina* Dataset

The *Methanosarcina* dataset comprised *M. mazei* (AE008384\_GR.gbk), *M. barkeri* (CP000099\_GR.gbk), *M. acetivorans* (AE010299\_GR.gbk), *M. thermophila* (CP000477\_GR.gbk), *M. hungatei* (CP000254\_GR.gbk), *M. marisnigri* (CP000562\_GR.gbk), *M. labreanum* (CP000559\_GR.gbk), *T. acidophilum* (AL139299\_GR.gbk), *T. volcanium* (BA000011\_GR.gbk), *P. horikoshii* (BA000001\_GR.gbk), *P. abyssi* (AL096836\_GR.gbk), and *P. furiosus* (AE009950\_GR.gbk).

#### *Yersinia* Dataset

The *Yersinia* dataset included *Y. enterocolitica* (AM286415\_GR.gbk), *Y. pestis* Antiqua (CP000308\_GR.gb), *Y. pestis* Nepal516 (CP000305\_GR.gbk), *Y. pestis* Mediaevalis 91001 (AE017042\_GR.gbk), *Y. pestis* Orientalis CO-92 (AL590842\_GR.gbk), *Y. pestis* Pestoides F (CP00068\_GR.gbk), *Y. pestis* Mediaevalis KIM5 (AE009952\_GR.gbk), and *Y. pseudotuberculosis* (BX936398\_GR.gbk).

#### Computing *funSim* Values for Genomes

For each genome, a file was created containing, in multiple FASTA format, GO terms for each gene separated according to the three GO categories “cellular component,” “biological process” (BP), and “molecular function” (MF). The program *funSim* (Schlicker et al. 2006) version 1.0 was used to compare genomes pairwise. The score was deduced from the categories BP and MF. The output of *funSim* is a distance matrix storing for each pair of genes  $a_i, b_j$  the value  $funSim(a_i, b_j)$ . For each pair of genomes  $G_k, G_l$  belonging to a dataset under study, such a matrix (*GkGl\_matrix*) was computed.

#### Computing Phylogenetic Distance Matrices by Means of GO4genome

According to the dataset  $G_1 \dots G_n$  to be analyzed, GO4genome reads the respective *GkGl\_matrices* and computes for each pair of genomes a  $Dist_{GO}$  value as described under Results and according to Formula (6). The set of  $Dist_{GO}$  values is written to a file in Nexus format (Maddison et al. 1997). The source code for the generation of Nexus-formatted distance matrices and the yersinia dataset can be downloaded from <http://www-bioinf.uni-regensburg.de>.

#### Creating Neighbor Nets

For the visualization of results, we utilized the program SplitsTree4 (version 4.8) (Huson and Bryant 2006). The output of GO4genome was fed into SplitsTree4. Neighbor nets were created by using default parameters.

## Results

### Toward a Novel Algorithm of Genome Comparison Based on Gene Ontology Annotations

As it was our aim to develop a method for the comparison of genomes which exploits encoded function, we first focused on an adequate scoring scheme. So far, gene content methods have been based exclusively on the concept of homology. For assessment of this approach, the following characteristics have to be considered. (1) This categorization of genes (gene products) is a binary one. Definitely, a scoring scheme with finer granularity supports a more precise comparison of genomic content, which is less error prone also. (2) The

classification may fail on paralogues. It was shown that gene duplication is an important factor in genome evolution (Snel et al. 2002). (3) This classification is based on a common evolution of respective genes. In cases where a nonorthologous in situ replacement of a gene via HGT preserves function, the analysis of homology will report disparate genome content.

With the advent of gene ontology (GO), this binary classification scheme can easily be replaced by a continuous one. GO is a standardized vocabulary permitting a coherent annotation of gene products. It is now common to supply genes and gene products with a set of GO terms annotating, e.g., function or their involvement in biological processes. Recently, methods for comparing sets of GO terms have been introduced (Del Pozo et al. 2008; Schlicker et al. 2006). The latter method relies on two similarity measures; one, named *funSim*, can be used to characterize the functional similarity of gene products. It has been shown that this identification of functionally related proteins is independent of their evolutionary relationship (Schlicker et al. 2006). The outcome of *funSim* is, for each pair of genes  $a_i, b_j$ , a score  $0.0 \leq \text{funSim}(a_i, b_j) < 1.0$ . For the following, we assume that genome  $G1$  consists of  $n$  genes  $a_1, a_2, \dots, a_n$ , and genome  $G2$  of  $m$  genes  $b_1, b_2, \dots, b_m$ , being annotated with GO terms. In addition, it is assumed that  $m \leq n$ , which can always be ascertained by changing indices, if necessary. A matrix  $GO\_S[a_1 \dots a_n][b_1 \dots b_m]$  can be computed, which harbors all *funSim*( $a_i, b_j$ ) values. In analogy to classical scoring matrices,  $GO\_S$  constitutes a basis for the comparison of  $G1$  and  $G2$  in gene function.

For analyses described below, we utilized the annotations deposited in the Genome Reviews database of the EBI (see “Materials and Methods”), which provides comprehensively annotated genomes. A typical example is *Escherichia coli* K-12 (accession number U00096\_GR.gbk). This dataset contained 4277 genes; 3496 have been annotated with GO terms. Of the remaining 781 genes, 462 have been described as “hypothetical” or “uncharacterized”; most of the other annotations are nonspecific. Therefore, one can assume that the largest fraction of shared genes has been provided with GO terms, putting an analysis on a sound basis.

As explained above, current algorithms for genome comparison are based on a binary classification of genes. Additionally, those classical algorithms for sequence comparison (Smith and Waterman 1981) which can utilize a scoring system cannot deal with inversions. However, this kind of genetic rearrangement occurs quite frequently, even in closely related genomes (Hughes 2000; Belda et al. 2005). Therefore, we propose a novel method which rests

on the identification of high-scoring segments as BLAST does (Altschul et al. 1990).

### Identifying Gene Series of Maximal Length with the Most Similar Function

An approximation for computing an edit distance is the construction of a *cover* (Swenson et al. 2008). A *cover* consists of a series of genes that exist in both genome  $G1$  and genome  $G2$ . A cover is said to be *optimal* if it corresponds to the minimal number of edit operations needed to transform  $G1$  into  $G2$ . However, the computation of an optimal cover is NP-hard (see Swenson et al. 2008). Therefore, a *minimal cover* that consists of the smallest number of series is used as a surrogate (Swenson et al. 2008).

Here we propose an algorithm that identifies a *functionally minimal cover* for the genomes  $G1$  and  $G2$ . The algorithm utilizes the matrix  $GO\_S[a_1 \dots a_n][b_1 \dots b_m]$ .  $GO\_S$  values were used to identify high scoring 3-tuples of genes (called *HS3Ts* or  $A\_HS3Ts$ ). We selected tuples of length 3, as these are the shortest  $n$ -mers allowing the identification of local optima. *HS3Ts* were determined according to the following rules and stored in a matrix  $TG$  of size  $n \times m$ :

$$HS3T[i, j] = \begin{cases} 1 & \text{if } \text{diag}(i, j) = \text{true} \\ 0 & \text{if } \text{diag}(i, j) = \text{false} \end{cases} \quad (1)$$

The value of  $\text{diag}(i, j)$  originated from the following expression (compare Supplementary Fig. S1, Panel A):

$$\begin{aligned} & (GO\_S[a_i, b_j] \geq GO\_S[a_i, b_{j+1}]) \wedge (GO\_S[a_i, b_j] \\ & \geq GO\_S[a_i, b_{j-1}]) \\ & \wedge (GO\_S[a_i, b_j] \geq GO\_S[a_{i-1}, b_j]) \wedge (GO\_S[a_i, b_j] \\ & \geq GO\_S[a_{i+1}, b_j]) \\ & \wedge (GO\_S[a_{i-1}, b_{j-1}] \geq GO\_S[a_{i-1}, b_j]) \wedge (GO\_S[a_{i-1}, b_{j-1}] \\ & \geq GO\_S[a_i, b_{j-1}]) \\ & \wedge (GO\_S[a_{i+1}, b_{j+1}] \geq GO\_S[a_{i+1}, b_j]) \wedge (GO\_S[a_{i+1}, b_{j+1}] \\ & \geq GO\_S[a_i, b_{j+1}]) \end{aligned} \quad (2)$$

For  $HS3T[i, j] = 1$ , three neighboring elements of  $TG$  were set to 1 according to  $TG[i, j] = TG[i + 1, j + 1] = TG[i - 1, j - 1] = 1$ .

Analogously, stretches indicating genomic inversions were identified:

$$A\_HS3T[i, j] = \begin{cases} 1 & \text{if } A\_diag(i, j) = \text{true} \\ 0 & \text{if } A\_diag(i, j) = \text{false} \end{cases} \quad (3)$$

$A\_diag(i,j)$  is the result of the following term (compare Supplementary Fig. S1, Panel B):

$$\begin{aligned} & (GO\_S[a_i, b_j] \geq GO\_S[a_i, b_{j+1}]) \wedge (GO\_S[a_i, b_j] \\ & \geq GO\_S[a_i, b_{j-1}]) \\ & \wedge (GO\_S[a_i, b_j] \geq GO\_S[a_{i-1}, b_j]) \wedge (GO\_S[a_i, b_j] \\ & \geq GO\_S[a_{i+1}, b_j]) \\ & \wedge (GO\_S[a_{i+1}, b_{j-1}] \geq GO\_S[a_{i+1}, b_j]) \wedge (GO\_S[a_{i+1}, b_{j-1}] \\ & \geq GO\_S[a_i, b_{j-1}]) \\ & \wedge (GO\_S[a_{i-1}, b_{j+1}] \geq GO\_S[a_i, b_{j+1}]) \wedge (GO\_S[a_{i-1}, b_{j+1}] \\ & \geq GO\_S[a_{i-1}, b_j]) \end{aligned} \quad (4)$$

If  $A\_HS3T[i,j]$  was 1, the content of  $TG$  was altered according to  $TG[i,j] = TG[i-1, j+1] = TG[i+1, j-1] = 1$ .

It is reasonable to prevent the further assessment of a pair of genes  $a_i, b_j$  that do not have similar function. Therefore, we introduced a lower limit  $GO\_cut\_off$  when filling  $GO\_S$ . Besides unrelated function, low  $funSim$  values might originate from inadequate annotation quality, from inconsistencies in the ontology, or from errors in the  $funSim$  implementation. To assess  $funSim$  values, we utilized GO4genome to compare all genes  $a_j$  of those 19 genomes  $G_i$  constituting the *E. coli* dataset (see below) with themselves and determined the distribution of  $funSim_{G_i G_i}(a_j, a_j)$  values. Altogether 48,746 gene pairs were analyzed; less than 5% had  $funSim$  values  $< 0.59$ , and more than 90% a  $funSim$  value  $\geq 0.87$ . Therefore, we selected  $GO\_cut\_off = 0.59$ . These results also confirmed that the annotations as deposited in the Genome Reviews database as well as the implementation of  $funSim$  are of high quality. We confirmed that the outcome of GO4genome does not depend critically on this parameter. Supplementary Fig. S2 allows comparison of analyses of the *E. coli* dataset based on  $GO\_cut\_off$  values of 0.59, 0.68, and 0.75.

If  $HS3Ts$  overlapped, longer diagonal elements  $diag(a_i, b_j, a_k, b_l)$  resulted, extending from position  $i, j$  to position  $k, l$ . The same could be the case for  $A\_HS3Ts$ . All diagonal elements occurring in  $TG$  were sorted according to their length and stored in a list,  $DIAG\_LIST$ . In the next step, an optimal set of diagonal elements was selected in order to label genes  $b_1$  to  $b_m$ . Starting with the element  $diag(a_i, b_j, a_k, b_l)$  of maximal length, genes  $b_j$  to  $b_l$  were labeled. In addition, all elements of any  $diag_m$  belonging to the corresponding intervals  $a_i \dots a_k$  or  $b_j \dots b_l$  were removed. Entries in  $DIAG\_LIST$  were processed until all genes  $b_1$  to  $b_m$  were labeled or until  $DIAG\_LIST$  was empty. The result of this process is a set of diagonal elements (a *functionally minimal cover*)  $S\_DIAG$  that contains all genes  $b_j$  of  $G_2$  possessing a significant functional similarity to genes of  $G_1$ . Please note that, due to this filter, gene pairs  $a_i, b_j$  possessing the highest  $funSim$  values are not necessarily

elements of  $S\_DIAG$ . This set may contain crosswise-arranged elements, which could be separated by gaps of arbitrary lengths (compare Supplementary Fig. S3). Figure 1 shows that the set of genes constituting  $S\_DIAG$  and those sequences aligned by MUMmer or generated by a pairwise BLAST analysis overlap to a great extent. Using the above results, a distance  $Dist_{GO}$  for  $G_1$  and  $G_2$  was calculated according to the following formulae:

$$sim_{GO}(diag_k) = \sum_{a_i, b_j \in diag_k} funSim(a_i, b_j) \quad (5)$$

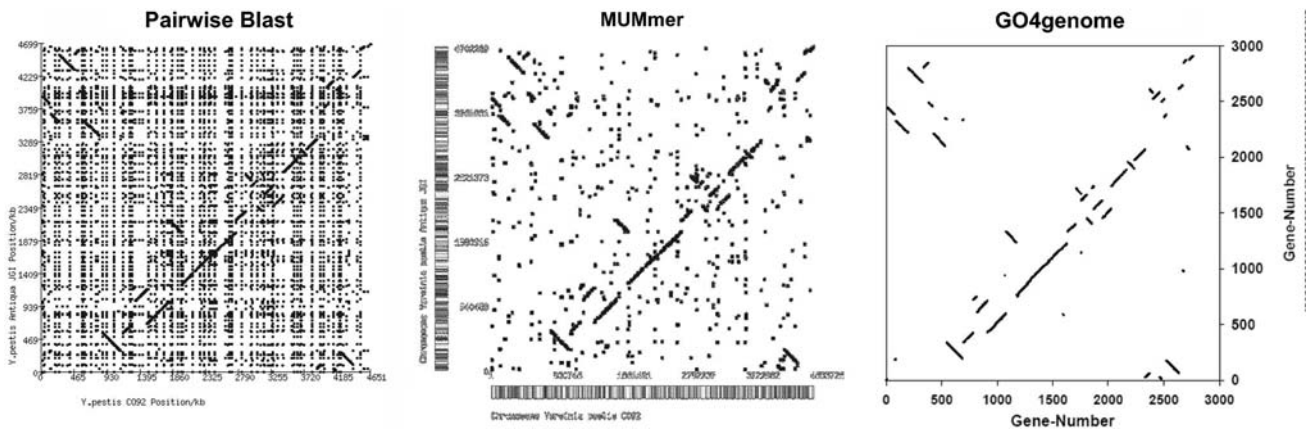
$$\begin{aligned} Dist_{GO}(G_1, G_2) \\ = -\log \left( \sum_{diag_k \in S\_DIAG} \left( \frac{sim_{GO}(diag_k)}{weighted\_gsize(G_1, G_2)} \right)^\lambda \right) \end{aligned} \quad (6)$$

$$weighted\_gsize(G_1, G_2) = \frac{\sqrt{2} \cdot size(G_1) \cdot size(G_2)}{\sqrt{size(G_1)^2 + size(G_2)^2}} \quad (7)$$

$sim_{GO}(diag_k)$  is the sum of all  $funSim$  values for those gene pairs constituting one element  $k \in S\_DIAG$ . If two neighboring elements  $diag_k, diag_l$  occupied the same diagonal line,  $sim_{GO}$ -values were merged (see Supplementary Fig. S3). For the computation of a distance,  $sim_{GO}(diag_k)$  values were divided by the weighted average genome size,  $weighted\_gsize(G_1, G_2)$ , in analogy to Korbelt et al. (2002). For Formula (6) we propose to use a  $\lambda$  which is  $> 1.0$ . In this case, any combination of two or more normalized  $sim_{GO}(diag_k)$  values (indicating rearrangements) will sum up to a value which is  $< 1.0$ . The comparison of trees deduced for the *E. coli* dataset (data not shown) proved that  $\lambda = 1.05$  is appropriate.

The evolutionary distance  $Dist_{GO}(G_1, G_2)$  was deduced from the estimated similarity by applying the negative logarithm, as proposed by Korbelt et al. (2002). Please note that short fragments contribute only marginally to the distance value; see Formula (6). Therefore, we did not consider elements consisting of fewer than three gene pairs; compare Formulae (2) and (4). For the *E. coli* dataset (see below), the number of elements making up individual sets  $S\_DIAG$  varied between 1 and 208.

For a set of genomes  $G_1 \dots G_n$ , the outcome of all pairwise comparisons  $G_i, G_j$  is a distance matrix of size  $n \times n$ . A frequently used method for the construction of a tree is some variant of a neighbor joining algorithm (Saitou and Nei 1987). The resulting tree will be free of ambiguities, if the distance matrix is additive. However, for the general case, we did not expect additive matrices when comparing several genomes via GO4genome. If conflicting signals (i.e., distances) exist, a neighbor net can be used for indication. We utilized the version implemented with SplitsTree4 (Huson and Bryant 2006).



**Fig. 1** Whole-genome comparison of *Y. pestis* CO-92 and *Y. pestis* Antiqua using three different methods. To identify genomic regions showing maximal synteny, three plots were generated. These originated from pairwise BLAST hits (left column), MUMmer (middle column), and GO4genome (right column). *Y. pestis* genomes contain a large number of transposases, contributing to the regular pattern in the BLAST plot and the “noise” in the MUMmer plots.

#### GO4genome Deduced a Sound Phylogeny for *E. coli* and Close Relatives

As the first case, we analyzed a dataset containing GO terms of all completely sequenced *E. coli* genomes, those of *Salmonella typhimurium*, *Shigella boydii*, *Shigella dysenteriae*, three strains of *Shigella flexneri*, *Shigella sonnei*, *Yersinia pestis*, *Buchnera aphidicola*, *Rickettsia prowazekii*, *Rickettsia conorii*, and *Aeropyrum pernix*. Figure 2 shows the resulting neighbor net. The net indicates that some conflicting signals exist. However, for *E. coli* species and close relatives, their phylogenetic relation could be resolved unambiguously. The uropathogenic strains *E. coli* 536, *E. coli* UTI89, and *E. coli* CFT073 and the avian pathogenic strain *E. coli* O1:K1/APEC form a subtree as well as the two enterohemorrhagic strains *E. coli* O157:H7/EDL933 and *E. coli* O157:H7/str. Sakai and *E. coli* K-12. The relationship of *E. coli* K-12, *E. coli* O157:H7, and *E. coli* CFT073 is in agreement with findings deduced from the comparison of DNA sequences (Elena et al. 2005) and tRNA genes (Withers et al. 2006). The observation that the genome composition of the avian *E. coli* O1:K1 strain is most similar to that of UTI89 followed by *E. coli* 536, *E. coli* CFT073, and *E. coli* K-12 is in agreement with results deduced from genome content (Johnson et al. 2007).

The position of *S. flexneri* and *S. typhimurium* corresponds to previous findings: *S. flexneri* is assumed to originate from an ancestral *E. coli* strain (Rolland et al. 1998). According to a phylogenetic analysis of *gyrB* gene sequences, *S. flexneri* is a closer relative of *E. coli* than of *S. typhimurium* (Fukushima et al. 2002). The relation of *S. flexneri* and the last-mentioned *E. coli* strains is in agreement with a whole-genome tree and an average nucleotide

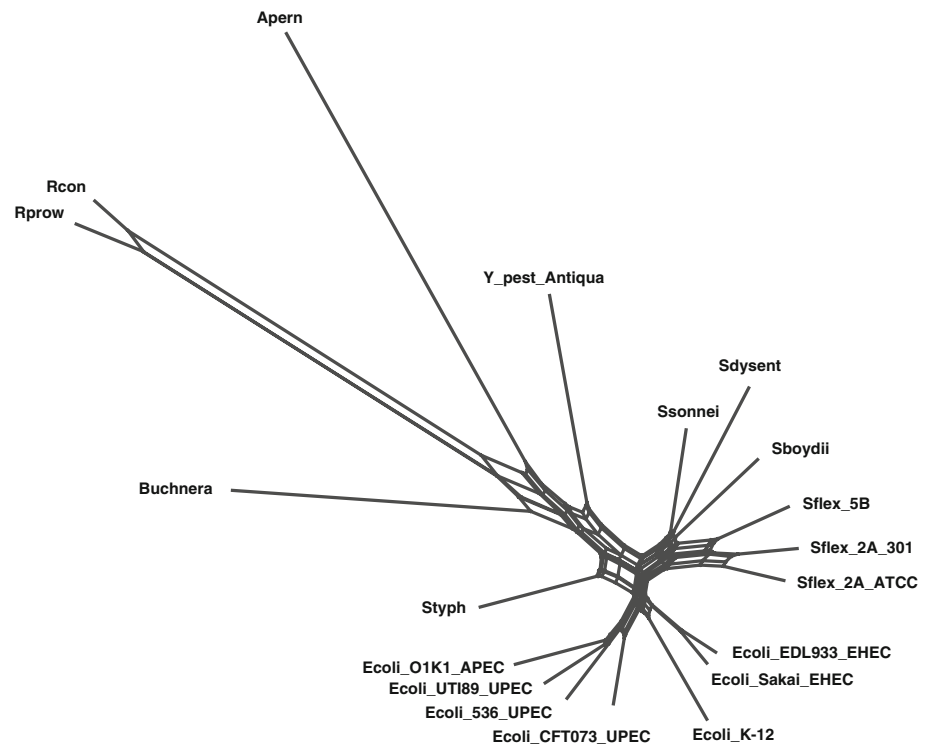
identity tree (see Konstantinidis et al. 2006). All *Shigella* genomes were grouped together; the three *S. flexneri* strains cluster in one distinct group. *S. boydii*, *S. dysenteriae*, and *S. sonnei* constitute a second cluster. A phylogenetic analysis of shigellae, based on smaller sets of gene sequences, resulted in inconsistent phylogenies (see Yang et al. 2007); see the “Discussion”.

*Buchnera*, *Y. pestis*, *Rickettsia*, and *A. pernix* were more distant from the other species. The positioning of *Buchnera* is a specific challenge, as the genome of this endosymbiont has undergone massive genome reduction since the divergence from a free-living  $\gamma$ -proteobacterial ancestor. High substitution rates and biased nucleotide patterns have been the reason for the deviant tree topologies computed for individual sequences. A tree deduced from a concatenation of 205 protein sequences gave the same relationship as shown in Fig. 2 for *E. coli*, *S. typhimurium*, *Y. pestis*, and *Buchnera* (Lerat et al. 2003). In summary, these consistencies demonstrate that the above method of analyzing gene function and order generates a sound phylogeny, which is in most cases consistent with classical methods. As expected, the topology of the GO4genome net is less resolved for distantly related species (compare Fig. 2). Gene order conservation is lost rapidly when comparing species which are less related (Tamames 2001).

Streptococci form Distinct Groups

The genus *Streptococcus* is one of the most diverse and important human and agricultural pathogens. The genomes of streptococci exhibit extreme levels of evolutionary plasticity accompanied by a high level of gene gain and loss. It has been shown that recombination is an important

**Fig. 2** A neighbor net of *E. coli*, shigellae, and several other microbial species deduced from encoded gene function. GO4genome was used to compute a distance matrix. SplitsTree4 was utilized to generate and display a neighbor net. A local net-like structure indicates ambiguities. Thus, regions of unclear topology can be visualized. See “Materials and Methods” for species names



factor in the evolution of *Streptococcus* genomes (Lefébure and Stanhope 2007). Based on gene gain, loss, and duplication, core-based phylogenies have been determined for *Streptococcus* and, more specifically, for *S. agalactiae* and *S. pyogenes* strains (Lefébure and Stanhope 2007). According to this approach, *S. pyogenes* and *S. agalactiae* are closely related, as well as *S. pneumoniae* and *S. suis*. Additionally, a tree for *Streptococcus* has been deduced from a joint analysis of 504 single-copy genes (Anisimova et al. 2007). In this case, *S. pyogenes* and *S. agalactiae* have been most similar, as well as *S. thermophilus* and *S. pneumoniae*. Genome organization as deduced by GO4genome is in agreement with these findings and additionally identifies the genome structure of *S. sanguinis* as most similar to that of *S. pneumoniae* and *S. suis*; see Fig. 3. In addition, the net topology is concordant with findings deduced from an analysis of *dnaJ* and *gyrB* sequences (Itoh et al. 2006).

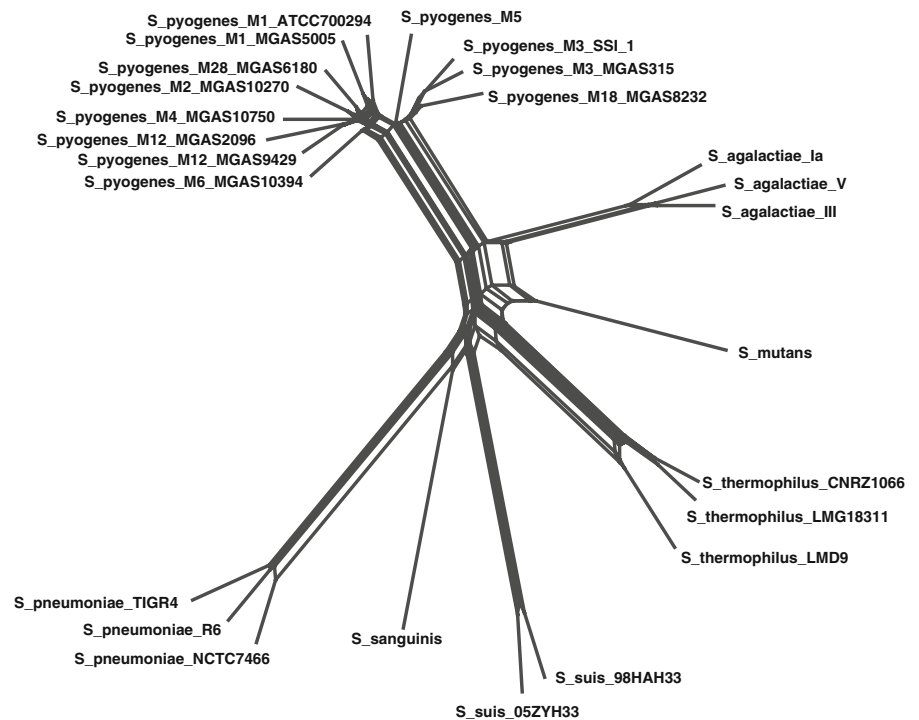
According to Lefébure and Stanhope (2007), among *S. pyogenes* strains the pairs (MGAS9429, MGAS2096; MGAS315, SSI-1) and (M1 GAS, MGAS5005) are most related. GO4genome predicted the same relationship; compare Fig. 3. However, for some species, like (MGAS8232, MGAS10394), the predictions differ. Additionally, the net indicates that the serovars M3 and M18 form one group, and M1, M2, M4, M12, and M28 a second group, which is less homogeneous. M5 and M6 lie isolated. In summary, the phylogenetic net showed a relatively low level of ambiguities. The analysis of genome organization

clearly separated individual *Streptococcus* species and allowed the grouping of serovars. As can be seen, gene gain and loss had no major impact on the overall genome organization of the species.

#### Horizontal Gene Transfer Has Little Effect on the Genome Organization of *Methanosarcina*

So far, three genomes of *Methanosarcina* have been analyzed. The genomes differ significantly in size: the genome of *M. mazei* contains 3370 genes; that of *M. barkeri*, 3606 genes; and that of *M. acetivorans*, 4540 genes. It has been postulated that up to 30% of the *M. mazei* genes have been acquired via HGT (Deppenmeier et al. 2002). For *M. mazei*, 8.1% of its genes constitute larger genomic islands with atypical codon usage; for *M. acetivorans* this fraction is 10.8% (Merkl 2004). Thus, these genomes represent an appropriate set for testing the robustness of GO4genome against HGT and variations in genome size. We compiled a dataset consisting of the above *Methanosarcina* and *Methanosaeta thermophila* (a distantly related methanosarcinales), *Methanospirillum hungatei*, *Methanoculleus marisnigri*, *Methanocorpusculum labreanum* (three methanomicrobiales), three pyrococci, and two thermoplasmata. Figure 4 shows the resulting neighbor net. All species belonging to the same order were grouped in distinct subnets; the only exception was *M. thermophila*. It is known that the evolutionary relationship to *Methanosarcina* is a distant one: analysis of the 16S RNA gave the

**Fig. 3** A phylogenetic classification of streptococci based on encoded gene function and gene location. GO4genome was used to compute a distance matrix according to Formula (6). By means of SplitsTree4, a neighbor net was generated and plotted. Among *S. pyogenes* strains, several clusters are discernible

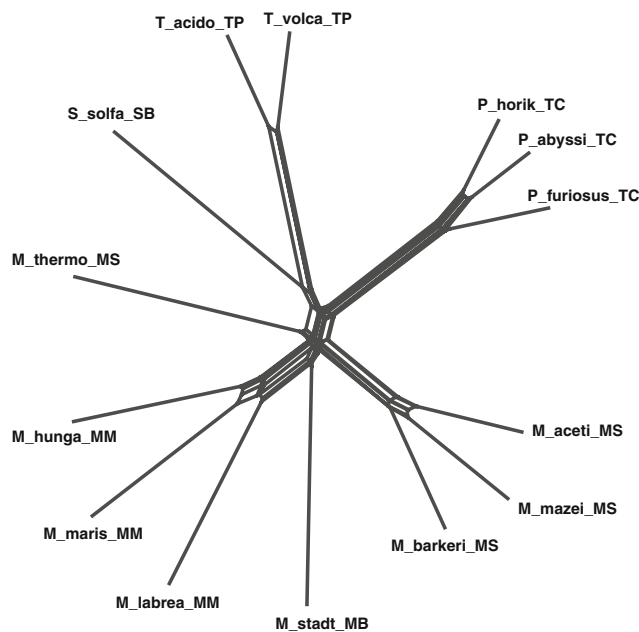


same local topology as shown in Fig. 4 for *M. mazei*, *M. thermophila*, and *M. hungatei* (Sekiguchi et al. 1998). Notably, the *Methanosarcina* species form a distinct subgroup, indicating that variations in genome size and larger

amounts of HGT have only a minor effect on the resolving power of GO4genome.

GO4genome Groups *Yersinia* in a Novel Way

*Yersinia pestis* is a Gram-negative bacterium and the causative agent of plague. *Y. pestis* is considered a recently emerged clone of *Y. pseudotuberculosis*, which evolved during the last 9000–40,000 years (Achtman et al. 2004). Originally, yersinia were grouped into a “nonclassical” subspecies (containing *Microtus*) and three “classical” biovars, based on their ability to reduce nitrate and utilize glycerol: Antiqua (positive for both markers), Mediaevalis (do not reduce nitrate but utilize glycerol), and Orientalis (positive for nitrate reduction but do not utilize glycerol). Due to the latest analytical methods and molecular relatedness, *Y. pestis* strains were split into three major branches (Achtman et al. 2004; Auerbach et al. 2007). Branch 0 contains *Y. pestoides* isolates and the *Microtus* isolate 91001. 1.ORI subsumes bacteria related to Orientalis strains, classical Mediaevalis strains are referred to 2.MED, and Antiqua isolates are split into two distinct groups, 1.ANT and 2.ANT, which were isolated in Africa and East Asia, respectively. A MLVA analysis suggested that 2.MED and 2.ANT represent sister clades (Achtman et al. 2004). Based on the analysis of several parameters like SNPs and the genome-specific inactivation of genes, it has been postulated that the Antiqua and CO-92 strains belong to one branch, and KIM and Nepal516 to the second one. According to this analysis, 1.ANT is closely related to the



**Fig. 4** A whole-genome phylogeny for methanosarcinales and other archaea. GO4genome was used to determine a distance matrix. SplitsTree4 was utilized for computation of a neighbor net and visualization. The suffixes indicate the lineage: *MS* methanosarcinales, *MM* methanomicrobiales, *TP* thermoplasmatales, *TC* thermococcales. See “Materials and Methods” for species names



Orientalis strain CO-92, while 2.ANT (represented by the Asian Antiqua strain Nepal516) is more closely related to the Mediaevalis strain KIM (Chain et al. 2006). Figure 5 shows that GO4genome proposed a different topology: one split separated *Y. pestis* Mediaevalis KIM5, *Y. pestis* biovar Microtus 91001, and *Y. pestis* Orientalis CO-92; a second one, the two Antiqua strains *Y. pestis* Antiqua and *Y. pestis* Nepal516; and a third, *Y. pseudotuberculosis* and *Y. enterocolitica*. This result indicates that the genome organization of biovars Mediaevalis (including Microtus) and Orientalis (represented by CO-92) is most similar; the same holds for the two representatives of the Antiqua biovar. For strain 91001, evolution from an ancient *Y. pestis* strain in a different lineage has been postulated (Song et al. 2004). According to GO4genome, its genome organization most resembles CO-92 and KIM5.

### Conclusions that can Be Drawn from Genome Organization

The analyses introduced above exemplify the application of GO4genome and indicate the types of problems that can be studied. In the following, we summarize some results. The crenarchaeon *Sulfolobus solfataricus* and the euryarchaeon *Thermoplasma acidophilum* inhabit the same ecological niche. There is evidence for a large amount of HGT between these species (Ruepp et al. 2000); many genes are closely related (e.g., *trpA* and *trpB* [Merkl 2007]). However, Fig. 4 clearly indicates that the genome composition of these species is quite dissimilar. For *M. mazei*, 8.1% of its genes constitute larger genomic islands with atypical codon usage; for *M. acetivorans* this figure is 10.8% (Merkl

2004). Figure 4 shows, that despite these islands, their overall genome composition is still highly similar. Both findings suggest that HGT restructures genome content only locally.

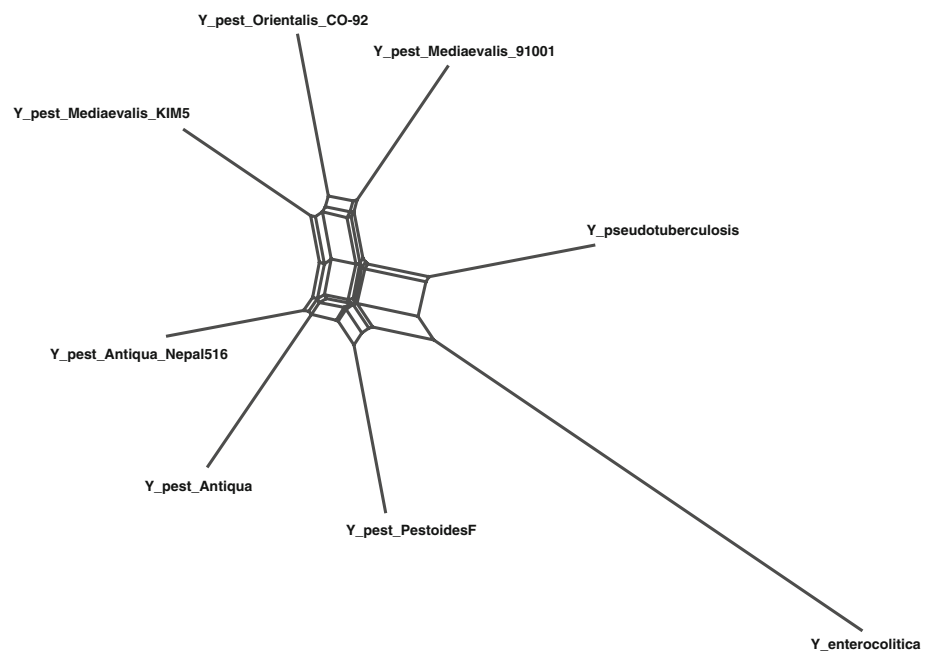
Shigellae do not have a single evolutionary origin; however, many of their characteristics indicate convergent evolution (Pupo et al. 2000). Figure 2 makes clear that convergent evolution can be seen on the level of genome organization. More specifically, genome composition separates the three *S. flexneri* strains from *S. boydii*, *S. sonnei*, and *S. dysenteriae*, which constitute a separate cluster. In the case of *Y. pestis*, the similarity of genome organization proposes a convergent evolution of the Antiqua strains. The effect is detectable on the genome level; compare Fig. 5.

### Discussion

#### What Is the Outcome of Classical Methods for the Cases Considered?

At first glance, it seems trivial to deduce the relationship of closely related prokaryotes. However, a comparison of the outcome of state-of-the-art methods makes clear that this is not always a trivial task. Several cases are discussed below. The first example is the *E. coli* group. According to the analysis of tRNA genes (Withers et al. 2006) and 36 randomly chosen genomic regions (Elena et al. 2005), *E. coli* O157:H7 is a closer relative of *E. coli* K-12 than *S. flexneri*. However, maximum likelihood analyses of core genomes and the ANI method identify *S. flexneri* as being more

**Fig. 5** A whole-genome phylogeny for *Yersinia* strains. GO4genome was used to determine a distance matrix. SplitsTree4 was utilized for computation of a neighbor net and visualization. The net indicates that the genomes of the two Mediaevalis strains and of CO-92, as well as those of the two Antiqua strains, are most similar, respectively, when compared regarding gene function and their location



closely related to *E. coli* K-12 than to *E. coli* O157:H7 (Konstantinidis et al. 2006). These differences might be due to the specific fate of individual genes. As has been pointed out, not all  $\gamma$ -proteobacterial core genes bear a similar phylogenetic signal supporting the same tree topology (Susko et al. 2006).

Shigellae have long been known to be closely related to *E. coli*. Due to biotyping, the genus has been divided into the four species *S. boydii*, *S. dysenteriae*, *S. flexneri*, and *S. sonnei*. Based on the analysis of eight housekeeping genes, it has been postulated that shigellae do not have a single evolutionary origin, which indicates convergent evolution of phenotypic properties (Pupo et al. 2000). An analysis of 23 housekeeping genes (Yang et al. 2007) has confirmed the clustering of shigellae into three main clusters, C1, C2, and C3. Clusters C1 and C2 consisted of *S. dysenteriae* and *S. boydii* strains; most of the strains (like F2a used here) constituting C3 were *S. flexneri*. The *S. boydii* strain Sb277 (used here) belonged to C1. The *S. sonnei* strain Ss046 (used here) was a direct neighbor of C1. The *S. dysenteriae* strain Sd197 (used here) laid isolated; the closest neighbors were *E. coli* EDL933 and *E. coli* Sakai. Contrariwise, an analysis of four chromosomal genes which were particularly polymorphic grouped Sd197 close to C1 and Ss046 close to C2 and C3 (Yang et al. 2007).

Based on the analysis of SNPs, a new nomenclature has been proposed for yersiniae (Achtman et al. 2004); see above. It has been postulated that Antiqua and CO-92 belong to one branch, and KIM and Nepal516 to a second one (Chain et al. 2006). A DNA microarray analysis of 22 strains of *Y. pestis* indicated that the two biovar strains Antiqua and Mediaevalis showed the most divergence from the CO-92 strain, and KIM and Nepal516 were clustered together (Hinchliffe et al. 2003). An analysis of CRISPR elements suggested that the Orientalis lineage branched out of the Antiqua strain earlier than the Mediaevalis biovar; the relative position of African Antiqua strains could not be fixed (Vergnaud et al. 2007). In summary, the above examples indicate that the phylogenetic signals studied highlight different aspects of genome evolution. This observation is in agreement with recent findings deduced from several methods of whole-genome phylogeny (McCann et al. 2008).

#### What Distinguishes Whole-Genome Analysis from Traditional Methods?

Several aspects of genome organization are not covered by classical methods. In many cases, bacteriophages are involved in the transfer of genomic islands. For *S. flexneri* 2a, 314 IS elements have been identified, which is more than sevenfold the content of *E. coli* K-12 (Jin et al. 2002).

A comparison of the *Y. pseudotuberculosis* genome with CO-92 and KIM10+ indicated that an extraordinary expansion of IS families has occurred since their divergence. It was deduced that the least common ancestor of CO-92 and KIM10+ carried 109 IS elements. Since their divergence from *Y. pseudotuberculosis*, KIM10+ and CO-92 have undergone 10 or 18 rearrangements, respectively (Chain et al. 2004). Thus, it is quite likely that the insertion elements and/or the subsequent rearrangements they have generated played an important role in the speciation of *Y. pestis* strains (Chain et al. 2004). *Y. pestis* is actively undergoing reductive evolution and there is some evidence for convergent evolution (Chain et al. 2006).

In addition to the acquisition of novel genes by means of HGT, genetic rearrangements alter the position, the orientation, or the coding strand with respect to the origin of replication. As a consequence, gene dosage may be affected, as has been demonstrated for inversions in the genome of *E. coli* (Hill and Gray 1988). Depending on position, the effects of such rearrangements differ drastically (Esnault et al. 2007). Compared to *E. coli* sequences, 13 translocations and inversions of size >5 kb have been identified in the genome of *S. flexneri*. It has been assumed that these rearrangements allow reoptimization of promoters in order to cope with selective pressure (Jin et al. 2002). The impact of rearrangements and their high frequency indicated above demand whole-genome analysis. In contrast to this approach, the analysis of a few genes or of SNPs covers a different aspect of phylogenomics, namely, the historical lineage of genes or genomes.

As has been shown, analysis of the common gene content has disadvantages as a measure for determination of phylogenies (Tamames 2001). In contrast, gene order conservation defines the course of evolution more precisely. In addition, its analysis does not depend on the presence of a certain set of genes. Along these lines, GO4genome supports a completely different aspect of “genome similarity,” supplementing sequence-based methods and those elucidating the evolution of genes and genomes. As our approach assesses genomic signals which are influenced by more and different parameters than those related to the fate of single molecules, the grouping of species that differs from an analysis of classical markers is no surprise and does not judge the quality of any method. The networks resulting from GO4genome trace the evolutionary process of speciation based not on mutational events but on signal similarities in genome organization. As shown above for yersiniae, the genome organization of *Y. pestis* Antiqua and *Y. pestis* Antiqua Nepal516 is most similar; the same holds for KIM5, CO-92, and Mediaevalis 91001. Among *Shigella*, the genome organization of *S. flexneri* strains differs from that of *S. boydii*, *S. dysenteriae*, and *S. sonnei*, which form a cluster. Most likely,

effects which shape genomes above the gene level are responsible for these similarities.

As is the case for many other algorithms, we cannot prove the liability of our approach *sensu strictu*. However, the concordance of a great portion of the net topologies with well-established phylogenetic relations makes our findings highly plausible. We have demonstrated for several cases of inconsistencies that independent findings indicate them as well. In addition, it is unlikely that, just by chance, the genomes of (say) the Antiqua strains or of shigellae cluster in the pattern observed.

#### Limitations and Further Improvements

For prokaryotes, the organization of their genes in operons (Jacob and Monod 1961) and uberoperons (Lathe et al. 2000) is well established and it is known that the degree of genomic rearrangements increases constantly with the time of divergence (Suyama and Bork 2001). This holds even though there are discordant processes like HGT or varying rates of evolution or gene loss. However, these processes have been shown to add noise rather than a directional bias (Dutilh et al. 2004). In summary, these findings argue for analysis of genome organization. The above method is the first one utilizing the overall genome structure for determination of phylogenetic trees. So far, gene order has been exploited for gene pairs (Korbel et al. 2002) or rearrangements have been studied for a reduced set of genes in  $\gamma$ -proteobacterial genomes (Belda et al. 2005). The approach introduced with GO4genome eliminates some of the pitfalls of sequence-based phylogenies by comparing genes on function. For pairwise comparison of the genomes, it is not necessary to compare the respective sequences, which avoids false assignments. Due to the “fuzzy” scoring function, the selection of paralogues has only a minor effect on the identification of conserved genomic segments. As ontology is exploited, in situ replacements of genes maintaining the function of gene products have little impact on the phylogenetic distance. We believe that assessing HGT events in this way is at least a considerable alternative. Microbial genomes may contain a substantial number of duplicated genes, which argues for filtering (cf. Fig. 1). The above findings show that the proposed processing is appropriate to identify relevant gene series which can surrogate a *cover*.

Optimal applications for GO4genome are the study of serovars (see Fig. 3) or of closely related species (see Fig. 2).

Several improvements of GO4genome are conceivable. So far, the algorithm assesses gene function and location but not gene orientation. When comparing two genomes, the transcriptional orientation of each gene pair can be the same (positive polarity) or different (negative polarity).

However, how to integrate this signal into Formula (6) is unclear. The algorithm considers the length and size of rearrangements but not their location, e.g., with respect to the origin of replication. To do this, it would be necessary to model gene dosage for each species.

Above, we have focused on genomes consisting of a single chromosome. An analysis of several chromosomes is trivial; how to consider plasmids is unclear. Unfortunately, approaches exploiting gene order cannot be utilized for higher organisms: gene order is poorly conserved in eukaryotes (Huynen et al. 2001). The ultimate goal would be the comparison of all completely sequenced microbial genomes in order to compare genome organization. Due to the modular concept of our approach, such an analysis is feasible.

**Acknowledgment** The project was carried out within the BiotechGenoMik-Network Göttingen financed by the German Federal Ministry of Education and Research (BMBF).

#### References

- Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, Chenal-Francisque V, Worsham P, Thomson NR, Parkhill J, Lindler LE, Carniel E, Keim P (2004) Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci USA 101:17837–17842
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
- Anisimova M, Bielawski J, Dunn K, Yang Z (2007) Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. BMC Evol Biol 7:154
- Auerbach RK, Tuanyok A, Probert WS, Kenefic L, Vogler AJ, Bruce DC, Munk C, Brettin TS, Eppinger M, Ravel J, Wagner DM, Keim P (2007) *Yersinia pestis* evolution on a small timescale: comparison of whole genome sequences from North America. PLoS ONE 2:e770
- Baptiste E, Boucher Y, Leigh J, Doolittle WF (2004) Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol 12: 406–411
- Baptiste E, Susko E, Leigh J, Ruiz-Trillo I, Bucknam J, Doolittle WF (2008) Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. Mol Biol Evol 25:83–91
- Belda E, Moya A, Silva FJ (2005) Genome rearrangement distances and gene order phylogeny in  $\gamma$ -proteobacteria. Mol Biol Evol 22:1456–1467
- Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, Brubaker RR, Fowler J, Hinnebusch J, Marceau M, Medigue C, Simonet M, Chenal-Francisque V, Souza B, Dacheux D, Elliott JM, Derbise A, Hauser LJ, Garcia E (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. Proc Natl Acad Sci USA 101:13826–13831
- Chain PS, Hu P, Malfatti SA, Radnedge L, Larimer F, Vergez LM, Worsham P, Chu MC, Andersen GL (2006) Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. J Bacteriol 188:4453–4463

- Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287
- Dalevi D, Eriksen N (2008) Expected gene-order distances and model selection in bacteria. *Bioinformatics* 24:1332–1338
- Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14:1394–1403
- Daubin V, Gouy M, Perrière G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* 12:1080–1090
- Del Pozo A, Pazos F, Valencia A (2008) Defining functional distances over gene ontology. *BMC Bioinformatics* 9:50
- Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martínez-Arias R, Henne A, Wiezer A, Bäumer S, Jacobi C, Brüggemann H, Lienard T, Christmann A, Bomeke M, Steckel S, Bhattacharyya A, Lykidis A, Overbeek R, Klenk HP, Gunsalus RP, Fritz H-J, Gottschalk G (2002) The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol* 4:453–461
- Dutilh BE, Huynen MA, Bruno WJ, Snel B (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J Mol Evol* 58:527–539
- Elena SF, Whittam TS, Winkworth CL, Riley MA, Lenski RE (2005) Genomic divergence of *Escherichia coli* strains: evidence for horizontal transfer and variation in mutation rates. *Int Microbiol* 8:271–278
- Esnault E, Valens M, Espéli O, Boccard F (2007) Chromosome structuring limits genome plasticity in *Escherichia coli*. *PLoS Genet* 3:e226
- Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res* 27:4218–4222
- Fukushima M, Kakinuma K, Kawaguchi R (2002) Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the *gyrB* gene sequence. *J Clin Microbiol* 40:2779–2785
- Hannenhalli S, Chappay C, Koonin EV, Pevzner PA (1995) Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics* 30:299–311
- Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC (2005) Whole-genome prokaryotic phylogeny. *Bioinformatics* 21:2329–2335
- Hill CW, Gray JA (1988) Effects of chromosomal inversion on cell fitness in *Escherichia coli* K-12. *Genetics* 119:771–778
- Hinchliffe SJ, Isherwood KE, Stabler RA, Prentice MB, Rakin A, Nichols RA, Oyston PC, Hinds J, Titball RW, Wren BW (2003) Application of DNA microarrays to study the evolutionary genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*. *Genome Res* 13:2018–2029
- Hughes D (2000) Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biol* 1:REVIEWS0006
- Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267
- Huynen MA, Snel B, Bork P (2001) Inversions and the dynamics of eukaryotic gene order. *Trends Genet* 17:304–306
- Itoh Y, Kawamura Y, Kasai H, Shah MM, Nhung PH, Yamada M, Sun X, Koyana T, Hayashi M, Ohkusu K, Ezaki T (2006) *dnaJ* and *gyrB* gene sequence relationship among species and strains of genus *Streptococcus*. *Syst Appl Microbiol* 29:368–374
- Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318–356
- Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, Zhang X, Zhang J, Yang G, Wu H, Qu D, Dong J, Sun L, Xue Y, Zhao A, Gao Y, Zhu J, Kan B, Ding K, Chen S, Cheng H, Yao Z, He B, Chen R, Ma D, Qiang B, Wen Y, Hou Y, Yu J (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res* 30:4432–4441
- Johnson TJ, Kariyawasam S, Wannemuehler Y, Mangiamiele P, Johnson SJ, Doetkott C, Skyberg JA, Lynne AM, Johnson JR, Nolan LK (2007) The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J Bacteriol* 189:3228–3236
- Konstantinidis KT, Ramette A, Tiedje JM (2006) Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Appl Environ Microbiol* 72:7286–7293
- Korbel JO, Snel B, Huynen MA, Bork P (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet* 18:158–162
- Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: a critical view. *Proc Natl Acad Sci USA* 100:9658–9662
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12
- Lathe WC 3rd, Snel B, Bork P (2000) Gene context conservation of a higher order than operons. *Trends Biochem Sci* 25:474–479
- Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 44:383–397
- Lefebvre T, Stanhope MJ (2007) Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol* 8:R71
- Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the  $\gamma$ -proteobacteria. *PLoS Biol* 1:E19
- Lin GN, Cai Z, Lin G, Chakraborty S, Xu D (2009) ComPhy: prokaryotic composite distance phylogenies inferred from whole-genome gene sets. *BMC Bioinformatics* 10(Suppl 1):S5
- Maddison DR, Swofford DL, Maddison WP (1997) NEXUS: an extensible file format for systematic information. *Syst Biol* 46:590–621
- McCann A, Cotton JA, McInerney JO (2008) The tree of genomes: an empirical comparison of genome-phylogeny reconstruction methods. *BMC Evol Biol* 8:312
- Merkl R (2004) SIGI: score-based identification of genomic islands. *BMC Bioinformatics* 5:22
- Merkl R (2007) Modelling the evolution of the archeal tryptophan synthase. *BMC Evol Biol* 7:59
- Nelson KE, Clayton RA, Gill SR et al (1999) Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399:323–329
- Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304
- Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ*. *Genome Biol* 4:R55
- Oshima K, Nishida H (2007) Phylogenetic relationships among mycoplasmas based on the whole genomic information. *J Mol Evol* 65:249–258
- Pupo GM, Lan R, Reeves PR (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci USA* 97:10567–10572
- Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804
- Rolland K, Lambert-Zechovsky N, Picard B, Denamur E (1998) *Shigella* and enteroinvasive *Escherichia coli* strains are derived

- from distinct ancestral strains of *E. coli*. *Microbiology* 144: 2667–2672
- Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* 407:508–513
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Sankoff D (1992) Edit distance for genome comparison based on nonlocal operations. In: Apostolico A, Crochemore M, Galil Z, Manber U (eds) Third annual symposium on combinatorial pattern matching, vol 644. Springer, Heidelberg, pp 121–135
- Schlicker A, Domingues FS, Rahnenführer J, Lengauer T (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* 7:302
- Sekiguchi Y, Kamagata Y, Syutsubo K, Ohashi A, Harada H, Nakamura K (1998) Phylogenetic diversity of mesophilic and thermophilic granular sludges determined by 16S rRNA gene analysis. *Microbiology* 144:2655–2665
- Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197
- Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. *Nature Genet* 21:108–110
- Snel B, Bork P, Huynen MA (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* 12:17–25
- Song Y, Tong Z, Wang J, Wang L, Guo Z, Han Y, Zhang J, Pei D, Zhou D, Qin H, Pang X, Han Y, Zhai J, Li M, Cui B, Qi Z, Jin L, Dai R, Chen F, Li S, Ye C, Du Z, Lin W, Wang J, Yu J, Yang H, Wang J, Huang P, Yang R (2004) Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res* 11:179–197
- Susko E, Leigh J, Doolittle WF, Baptiste E (2006) Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the  $\gamma$ -proteobacteria. *Mol Biol Evol* 23:1019–1030
- Suyama M, Bork P (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet* 17:10–13
- Swenson KM, Marron M, Earnest-DeYoung JV, Moret BME (2008) Approximating the true evolutionary distance between two genomes. *ACM J Exp Algorithm* 12:3.5
- Tamames J (2001) Evolution of gene order conservation in prokaryotes. *Genome Biol* 2:RESEARCH0020
- Tesler G (2002) GRIMM: genome rearrangements web server. *Bioinformatics* 18:492–493
- Vergnaud G, Li Y, Gorgé O, Cui Y, Song Y, Zhou D, Grissa I, Dentovskaya SV, Platonov ME, Rakin A, Balakhonov SV, Neubauer H, Pourcel C, Anisimov AP, Yang R (2007) Analysis of the three *Yersinia pestis* CRISPR loci provides new tools for phylogenetic studies and possibly for the investigation of ancient DNA. *Adv Exp Med Biol* 603:327–338
- Withers M, Wernisch L, dos Reis M (2006) Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. *RNA* 12:933–942
- Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* 74: 5088–5090
- Xin C, Jie Z, Zheng F, Peng N, Yang Z, Stefano L, Tao J (2005) Assignment of orthologous genes via genome rearrangement. *IEEE/ACM Trans Comput Biol Bioinform* 2:302–315
- Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, Zhu Y, Yu J, Jin Q (2007) Revisiting the molecular evolutionary history of *Shigella* spp. *J Mol Evol* 64:71–79