BMC Bioinformatics

BioMed Central

Research

Open Access

ComPhy: prokaryotic composite distance phylogenies inferred from whole-genome gene sets

Guan Ning Lin[1], Zhipeng Cai[2], Guohui Lin[2], Sounak Chakraborty[3] and Dong Xu*[1]

Address: [1]Digital Biology Laboratory, Informatics Institute, Computer Science Department and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA, [2]Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada and [3]Department of Statistics, University of Missouri, Columbia, MO 65211, USA

Email: Guan Ning Lin - gnln66@mizzou.edu; Zhipeng Cai - zhipeng@cs.ualberta.ca; Guohui Lin - ghlin@cs.ualberta.ca; Sounak Chakraborty - chakrabortys@missouri.edu; Dong Xu* - xudong@missouri.edu

* Corresponding author

from The Seventh Asia Pacific Bioinformatics Conference (APBC 2009)
Beijing, China. 13–16 January 2009

Published: 30 January 2009

BMC Bioinformatics 2009, **10**(Suppl 1):S5 doi:10.1186/1471-2105-10-S1-S5

This article is available from: http://www.biomedcentral.com/1471-2105/10/S1/S5

Abstract

Background: With the increasing availability of whole genome sequences, it is becoming more and more important to use complete genome sequences for inferring species phylogenies. We developed a new tool ComPhy, 'Composite Distance Phylogeny', based on a composite distance matrix calculated from the comparison of complete gene sets between genome pairs to produce a prokaryotic phylogeny.

Results: The composite distance between two genomes is defined by three components: Gene Dispersion Distance (GDD), Genome Breakpoint Distance (GBD) and Gene Content Distance (GCD). GDD quantifies the dispersion of orthologous genes along the genomic coordinates from one genome to another; GBD measures the shared breakpoints between two genomes; GCD measures the level of shared orthologs between two genomes. The phylogenetic tree is constructed from the composite distance matrix using a neighbor joining method. We tested our method on 9 datasets from 398 completely sequenced prokaryotic genomes. We have achieved above 90% agreement in quartet topologies between the tree created by our method and the tree from the Bergey's taxonomy. In comparison to several other phylogenetic analysis methods, our method showed consistently better performance.

Conclusion: ComPhy is a fast and robust tool for genome-wide inference of evolutionary relationship among genomes. It can be downloaded from http://digbio.missouri.edu/ComPhy.

Background

The systematic classification of bacteria has been a long-standing problem because very limited morphological features are available. For a long time researchers could only group together similar bacteria for practical determinative needs [1]. Woese and collaborators initiated molecular phylogeny of prokaryotes by making use of the small subunit (SSU) ribosomal RNA (rRNA) sequences [2]. The ssu-rRNA trees [3] have been considered as the standard Tree of Life by many biologists.

Attempts to explicate the phylogeny of prokaryotes based on the ssu-rRNA have been by-and-large successful [3,4]. However, although such molecules have proved to be very useful phylogenetic markers, mutational saturation is a problem due to the restricted length and limited number of mutation sites [5]. Another well-known problem is that phylogenetic trees, constructed on single gene families may show conflicts [6] due to a variety of causes, such as LGT (Lateral Gene Transfer) [7], hybridization, lineage-sorting, paralogous genes [8], and pseudogenes [9]. With the increasing availability of whole genome sequences, methods using vast amounts of phylogenetic information contained in complete genome sequences are becoming more and more important for inferring species phylogenies. Because phylogenetic information extracted from whole genomes is based on the maximum genetic information, the resulting phylogenetic tree should be the best reflection of the evolutionary history of the species, assuming this history is tree-like [7,10]. Phylogenomics, i.e. using entire genomes to infer a species tree, represents the state of art for reconstructing phylogenies [11,12].

A relatively obvious approach to phylogenetic analysis of whole genomes uses multiple sequence alignments with certain evolutionary models [9,13]. However, the multiple sequence alignment strategy may not work for whole genomes and the evolutionary models may not always be applicable. Multiple sequence alignment could be misleading due to gene rearrangements, inversion, transposition and translocation at the genome level [7,8], unequal lengths of sequences, LGT, etc. On the other hand, reliable statistical evolution models are yet to be suggested for complete genomes.

To address these issues, Sankoff and Blanchette [18] defined an evolutionary edit distance as the number of inversions, transpositions and deletions/insertions required to change the gene order from one genome into another. Similar distance measures using rearrangement, recombination, breakpoint, comparative mapping and gene order have been extensively studied for whole genome phylogeny [14-21]. These approaches are computationally expensive, and in general do not produce correct results on events such as non-contiguous copies of a gene on the genome or non-decisive gene order.

Gene content was proposed as distance measure in whole genome phylogeny where "the similarity between two species is defined as the number of genes they have in common divided by their total number of genes" [22]. This idea was further extended to use lists of nucleotide segment pairs in comparison instead of lists of genes [23]. Such method fails when the gene contents of organisms are very similar, such as bacteria in closely related families, or chimerical genomes.

Overlapping gene information was also used to infer the genome phylogenies [24]. Overlapping genes are defined as pairs of adjacent genes of which the coding sequences overlap partly or entirely. Although overlapping genes have been shown to be a consistent and conserved feature across all microbial genomes sequenced to date [25], the limited amount of overlapping genes is usually not enough for evaluating a large number of genomes.

Some other methods infer phylogenies based on protein structural domain information [26-31], which considers LGT. However, they assume some proteome evolution models with lateral structural domain transfer events, which may not be accurate. Also, the method readjusts the protein structural domain graph each time when a LGT event is introduced, the complexity of model testing increases substantially when large number of lateral structural domain transfer events been assumed.

In this paper, we introduce a new tool 'ComPhy', which utilizes a robust and much less complex strategy, called 'Gene Composite Distance'. It combines different aspects of evolutionary relationships among genomes to produce a phylogenetic tree from a given set of whole genome sequences. We have applied this approach to 398 prokaryotic genomes, which were downloaded from NCBI [32]. More precisely, composite distance measure starts with an all-against-all pairwise genome comparison using BLASTP [33]. In the second step, a distance matrix is calculated from three components, i.e., GDD (Gene Dispersion distance), GBD (Genome Breakpoint distance) and GCD (Gene Content distance). This distance matrix is then fed to a distance-based algorithm, Neighbor-Joining (NJ) [34,35], using a third-party tool 'Phylip' [47] to produce a phylogenetic tree. In our current study, we do not consider LGT. Our goal is to have mathematically tractability and to develop a generalized phylogenetic distance model and a phylogenetic tree construction platform that can be easily applied to any species. Furthermore, using the completely sequenced genomes allows the construction of a phylogeny less sensitive to inconsistencies, such as LGT, unrecognized paralogy, and highly variable rates of evolu-

tion among different regions in a genome. The result phylogenetic trees are more representative of whole-genomes than those from single-gene trees.

Methods

Taxon selection

398 single-chromosome prokaryotic genome protein sequences were downloaded in the Fasta format from the NCBI [32] ftp server in September 2007. The physical gene location files of these genomes were also downloaded from NCBI in a tab-delimited format. We represent the species biological names as defined in Bergey's code [1]. For example, a lineage in the Bergey's Manual of Systematic Bacteriology or its online outline is abbreviated as B13.3.2.6.2 = Phylum BXIII (Firmicutes), Class III (Bacilli), Order II (Lactobacillales), Family VI (Streptococcaceae), Genus II (Lactococcus). Table 1 shows the taxon statistics of the 432 prokaryotic genomes, including 34 multi-chromosomal species that we do not consider in this paper.

In order to test the performance of our method, 9 datasets with different combinations of the whole 398 genomes are formed for different purposes. Dataset 1 is formed by 52 randomly selected species from the Bergey's taxonomy tree to test for robustness. We like to test the performance of our method on this broad range of species. Dataset 2 has 53 species, half of them are randomly picked from Archaea genomes (A in Bergey's code) and half are randomly picked from Baceteria genomes (B in Bergey's code). These 53 species from two major clades have a clear taxonomy structure with two clusters. Dataset 3 has 82 species, half of them are from Phylum B12 since half of 398 genomes are actually from B12 Phylum, and the other half is randomly selected from all the other types of genomes, i.e., half of them have a tight cluster and the other half are diverse. Dataset 4 includes all the 398 single chromosome genomes. Since many prokaryotes are from Phylum B12 and Phylum B13, therefore, we form datasets 5 and 6 from all the 181 Phylum B12 genomes and 96 Phylum B13 genomes, respectively, to test for the effects of the co-linearity of datasets on phylogeny construction. Dataset 7 is a union of datasets 5 and 6, again with two tight clusters of species. Dataset 8 has 165 prokaryotic species obtained from BPhyOG [24] in order to compare the performance between our method and the overlapping gene based phylogeny used by BPhyOG. Dataset 9, with 54 prokaryotic species was obtained from Deeds [31] for comparing performance between our method and the structural domain based phylogeny construction. Both dataset 8 or 9 contain a subset of the 398 complete prokaryotic data. We believe these 9 diverse datasets allow us to test the robustness of our method comprehensively.

Table 1: Taxon statistics of the 432 prokaryotic complete genomes

Phylum	C	O	F	G	S	str
A1	1	3	4	4	7	7
A2	8	9	12	18	23	23
A3	1	1	1	1	1	1
Subtotal 3	10	13	17	23	31	31
B1	1	1	1	1	1	1
B2	1	1	1	1	1	1
B4	1	2	2	2	3	4
B6	1	1	1	1	2	2
B10	1	3	3	8	15	19
B11	1	1	1	2	4	4
B12	5	33	53	99	157	208
B13	3	7	14	22	58	96
B14	3	9	15	16	31	35
B15	1	1	1	1	1	1
B16	1	1	2	3	7	11
B17	1	1	2	3	7	9
B19	1	1	1	2	2	2
B20	3	3	5	5	6	7
B21	1	1	1	1	1	1
Subtotal 15	25	66	103	167	296	401
Total 18	35	79	120	190	327	432

P = Phylum, C = Class, O = Order, F = Family, G = Genus, S = Species, str = Strain. A = Archaea and B = Bacteria.

Identification of orthologs

The initial step in the phylogenetic analysis methods is to determine which genes are to be compared between species. Since the ultimate goal is determining the distance between every two genomes, intuitively we use pairwise orthology for every pair of genomes. So far there is no existing database containing the orthologous groups for all the genomes that we are studying.

Here, we define orthologs by performing an all-against-all BLAST between every pair of protein sequences for each pair of species. The reciprocal best BLAST hits are used to determine the list of orthologs between every pair of species. Additional filtering methods have also been applied to refine the list of orthologs between the pair of genomes, such as pairs of genes to be considered orthologs must satisfy BLAST hit with E-values below 10^{-3} and sequence identity higher than 30%. The tests of variations of ortholog definition will be shown in the Result and Discussion section. In ComPhy, we also give the users the flexibility to apply their definitions of orthologs.

Composite distance phylogeny

This strategy is to compute a distance between any two genomes *X* and *Y* based on the set of orthologs obtained in the previous step. This new systematic composite dis-

tance measurement takes into account both similarities and dissimilarities between every pair of genomes using their entire gene sets. ComPhy utilizes this composite distance formulation to determine the phylogeny for given genomes. The formulation will be discussed in three separated calculation steps, GDD, 'Gene Dispersion Distance', GBD, 'Genome Breakpoint Distance' and GCD, 'Gene Content Distance'.

GDD (Gene dispersion distance)

The first component, GDD, is to quantify the extent of dispersion of orthologs from one genome to another. Our assumption is that the closer the two species in the evolutionary tree, the more similar the physical arrangements of corresponding orthologs are. The dispersion of orthologs from one genome to another can be seen as how far orthologs move away from their physical locations during evolution due to events such as rearrangement, recombination, insertion and deletion. The further the evolution distance of the two species, the more dispersed the orthologs between two species are. In other words, the distance separations of pairs of orthologs are more conserved between a pair of closely related genomes. To simplify the problem, we consider only pairs of orthologs that are right next to each other. The gene dispersion distance of an ortholog pair from genome A to genome B can be then formulated as

$$d_{GDD}(A, B) = \frac{\sum l_{i,i+1}}{n^2}$$

where $l_{i,\ i+1}$ is the distance separation between the i-th ortholog and the $(i+1)$-th ortholog of genome B in genome A. For example, if orthologs b_i and b_{i+1} are next to each other on genome B, but their corresponding orthologs a_x and a_y in genome A are l orthologs apart (counting 1 ortholog separation as 1 distance unit), then the distance between this ortholog pair is l. n is number of orthologs between genomes A and B, which is the maximum dispersion distance. For normalization, one n is needed to normalize the size of a genome or the total number of orthologs (n). Another n is to normalize the dispersion distance against the maximum distance between two orthologs, which is n also. Thus, n^2 is needed for normalization. In fact, the normalization factors, such as n, n^2, and n^3, have been tested to see performances. Our study has shown that using n^2 as the normalization factor, in terms of range of dispersion distance and number of shared orthologs, has the optimal results.

Note, the gene dispersions, from A to B and from B to A, are not necessarily symmetric. We can define the dispersion distance measure in three different ways, namely, we can use the average $D(A, B) = [d(A, B) + d(B, A)]/2$ or use one of the two directions, either $D(A, B) = d(A, B)$ or $D(A, B) = d(B, A)$. Our study indicates that this directionality does have some impact on the overall performance and averaging over both directions produces better and more consistent results. Thus, the dispersion distance between genome A and B is defined as

$$D(A, B) = [d(A, B) + d(B, A)]/2$$

Figure 1 shows two different dispersions of orthologs pairs between genome B and genome A, and between genome B and genome C as a hypothetical example. To calculate the distance between two genomes, we need to calculate each distance separation of pair of neighboring orthologs of one genome in another. In Figure 1, given there are 13 orthologs, the distance separation for orthologs pair of b_1b_2 is a_4a_{11}, so distance separation of b_1b_2 is $l_{b_1b_2} = a_{11} - a_4 = 7$, and distance separation of b_2b_3 is $l_{b_2b_3} = a_{11} - a_5 = 6$, etc. The total distance separations of all ortholog pairs b_ib_{i+1} of genome B in A is $\sum l_{b_i,b_{i+1}} = 66$, and hence, $d_{GDD}(B, A) = 66/13^2 = 0.391$ for GDD from genome B to A. Using same calculation, we can get $d_{GDD}(A, B) = 47/13^2 = 0.278$, $d_{GDD}(C, B) = 38/13^2 = 0.225$ and $d_{GDD}(B, C) = 37/13^2 = 0.219$. We then get $D(A, B) = 0.34$ and $D(B, C) = 0.22$. As a result, genome B is closer related to genome C than to genome A. In other words, the ortholog pair distance separations are more conserved between B and C than between A and B.

We will also use a real example of three bacterial species, *Pyrobaculum aerophilum str. IM2* (A1.1.1.1.3), *Pyrobaculum islandicum DSM 4185* (A1.1.1.1.3), and *Thermus thermophilus HB27* (B4.1.2.1.1) as an example to show the gene dispersion distance idea. By putting the dispersion distances of ortholog pairs, between *P. aerophilum* and *P. islandicum* and between *P. aerophilum* and *T. thermophilus*, in different distance bins, Figure 2 demonstrates the conservation of the dispersion distance of ortholog pairs between closely related species. *P. aerophilum* and *P. islandicum*, both belong to Thermoproteaceae family in Thermoprotei order, so in the figure the black bins show uneven distribution of frequencies and most dispersion distances are falling into the smallest distance bin. This result agrees with the experimental finding that these two species are highly similar in terms of contents of genes and overall genome organization [48]. In contrast, *T. thermophilus* is a member of Thermaceae family in Deinococci order, as a result the white bins show more evenly distributes of dispersion distances among distance bins.

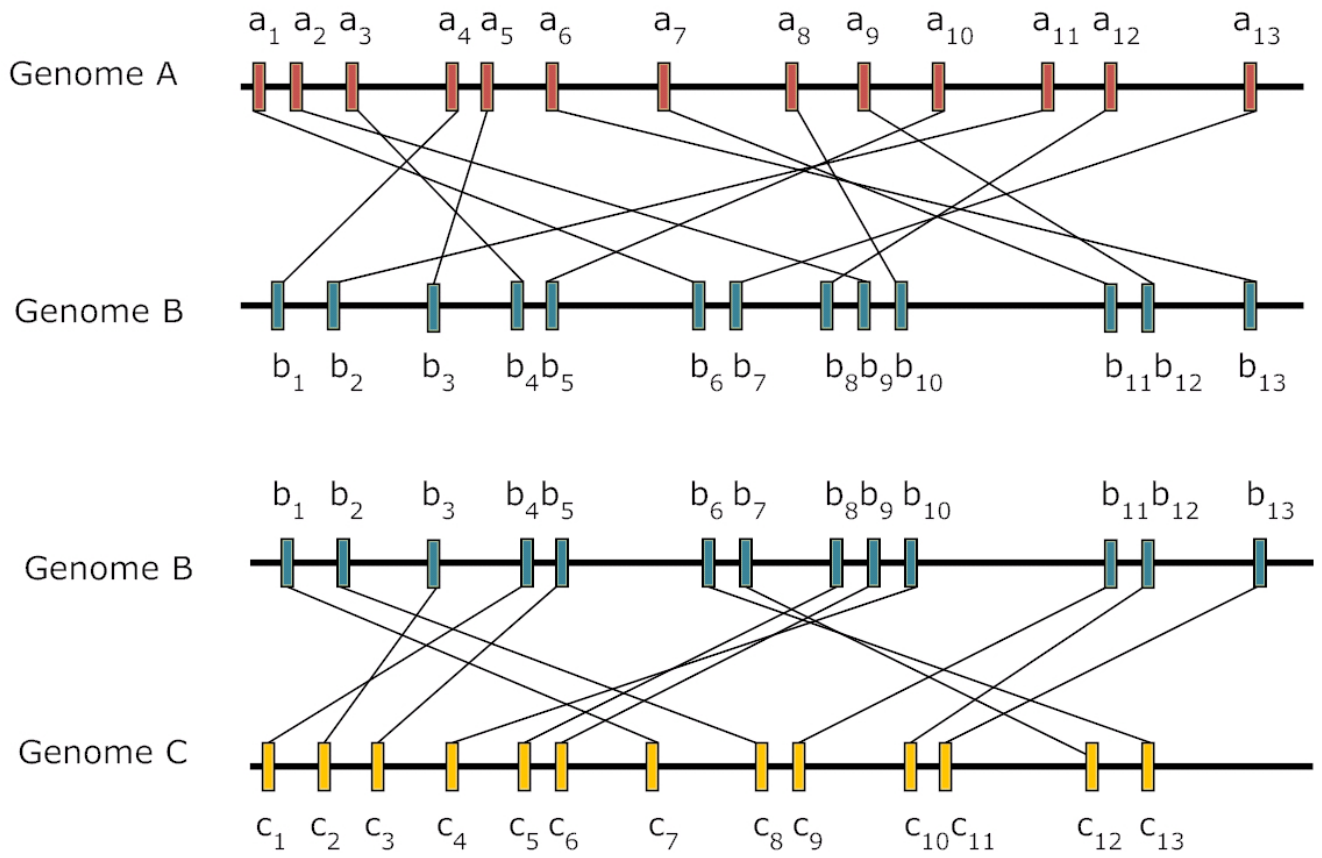

Figure 1
Gene dispersions example between pairs of genomes. The horizontal dark lines represent hypothetical genomes *A*, *B* and *C*. Each vertical box on the line is an ortholog and all orthologs are indexed according to their physical locations. The line connecting two boxes from one genome to another represents corresponding orthologs between two genomes.

GBD (genome breakpoint distance)

This distance transformation is based on the concept of breakpoints, where two sequence segments map consecutive intervals in one genome onto non-consecutive intervals in the other [18,40]. We simplified it by considering a breakpoint as where two ortholog sets map consecutively in one genome but not in the other. In other words, a breakpoint defined here is where the consecutive mapping of a set of orthologs between two genomes stops. Figure 3 gives a hypothetical example. There are two separated sets of consecutive orthologous genes and they are (a_3b_2, a_4b_3, a_5b_4, a_6b_5) and (a_8b_{13}, a_9b_{12}, $a_{10}b_{11}$), the consecutive mapping between genome *A* and *B* stops at positions a_2a_3, a_6a_7 and $a_{10}a_{11}$ in genome *B*. Therefore, there would be three breakpoints between genome *A* and genome *B*. Let X_{AB} be the number of breakpoints between genome *A* and genome *B*, and N_{AB} be the total number of orthologs between genome *A* and genome *B*. We then define a breakpoint similarity between *A* and *B* as

$$D_{GBD}(A,B) = 1 - \frac{X_{AB}}{N_{AB}}$$

Using Figure 3 as an example, there are 3 breakpoints and there are 13 orthologs, and hence, $D_{GBD}(A, B)$ is 1–3/13 = 0.769.

GCD (gene content distance)

The last component of composite distance is calculated using the idea of gene content [22] to show the similarity between two genomes. Here, we define the distance as

$$D_{GCD}(X,Y) = 1/(\frac{2^*N_c}{N_X+N_Y})$$

where N_c is the number of orthologs between *X* genome and *Y* genome, N_X is the number of genes in genome *X*, and N_Y is the number of genes in genome *Y*.

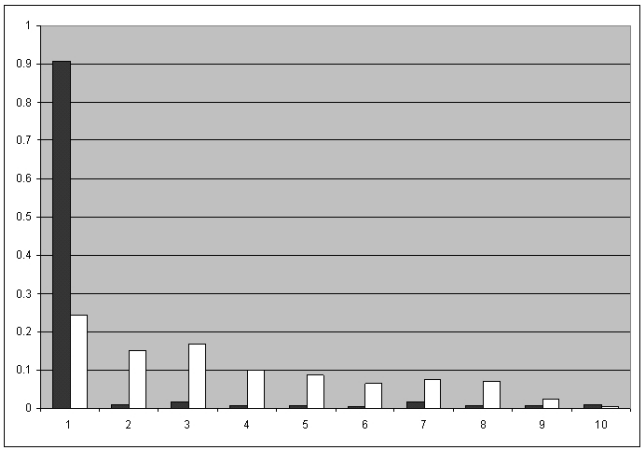

Figure 2
Histogram of dispersion distance counts from *P. islandicum* to *P. aerophilum* and from *T. thermophilus* to *P. aerophilum*. We divide the range of dispersion distance between ortholog pairs into 10 bins on X-axis between *P. aerophilum* and *P. islandicum* and between *P. aerophilum* and *T. thermophilus*. The lower indexed bin contains shorter dispersion distance of ortholog pairs. The height of the bin represents the frequency of dispersion distance falling into the bin. Using *P. aerophilum* as the target genome, a black bar represents the frequency of the dispersion distance between genomes *P. aerophilum* and *P. islandicum*, and a white bar represents the frequency of dispersion distance between *P. aerophilum* and *T. thermophilus*.

Composite distance formulation

The composite distance used in this paper for genome distance calculation has three distance components described above, where they represent three different aspects of genomes. GDD describes conservation of relative physical separation distances of orthologs, where this conservation can be thought as evolution timestamps. GBD utilizes the ordering of genes between a pair of genomes. Although it has some correlation with the first component, GBD reflects more local synteny (such as micro-synteny) instead of large-scale genome rearrangement characterized by GDD. GCD shows the level of similarity shared from genome composition without considering gene locations. It acts as an adjustment for different sizes of genomes, which could be thought as a normalization factor. Preliminary experiments showed that they are all very informative (Table 2). Therefore, the composite distance is defined as following:

$$D(X, Y) = \log D_{GDD} + \log D_{GBD} + \log D_{GCD}$$

We apply logarithm to the formula for retaining precision of computing and correcting the saturation effects in sequence data [41]. By considering the three distances, we generalize the conservation of gene order into an accurate and robust measurement.

Other genome distance measures

To compare with methods developed by other researchers, we also implemented several other distance measures for phylogenetic analysis.

(1) Overlapping gene phylogenetic distance

Overlapping genes (OG) [24] are defined as pairs of adjacent genes of which the coding sequences overlap partly or entirely. The distance between genomes i and j is defined as:

$$D_{ij} = 1 - \frac{x_{ij} + x_{ji}}{2 * \min(x_i, x_j)}$$

where x_i is the number of OG pairs in genome i and x_{ij} is the number of OG pairs in genome i that have their respective orthologs in genome j, and *vice versa* for other subscripts.

(2) Structural domain phylogenetic distance

The structural domain based distance method [31] uses the idea of Protein Domain Universe Graph (PDUG), a graph in which a nonredundant set of all known protein structural domains [42,43] are represented as nodes, and the structural similarity between domains is used to define edges between them. The distribution of edges per node in the graph was shown markedly different from random graph [44]. By combining with information of probability of LGT events, a similarity distance between species could be constructed. For example, degree distribution is calculated by comparing domain graph against known PDUG as follows:

$$p_N(k) = \sum_{s=k}^{Maxk_{N_0}} \binom{s}{k} \left(\frac{N}{N_0}\right) \left(1 - \frac{N}{N_0}\right)^{s-k}$$

where $Maxk_{N_0}$ is the degree of the maximally connected node in the underlying graph (such as PDUG) with N_0 nodes and a species graph with N nodes.

LGT is modeled as the movement of a node from a proteome in which that node exists into a proteome in which it does not. A transfer does not remove the node from the "donor" organism, but it may replace (thus, "erase") one of the nodes in the acceptor organism to preserve the proteome size. The donor and acceptor organisms are chosen randomly, and the transferred node is chosen randomly from the set of nodes in the donor proteome that do not exist in the acceptor proteome. The acceptor node that is replaced is also chosen at random.

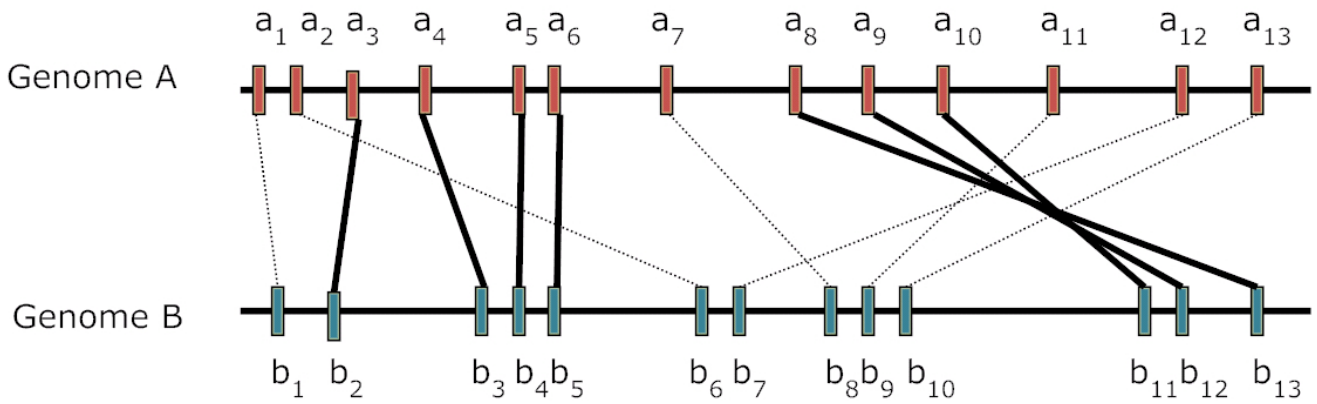

Figure 3
Identification of Breakpoints. The horizontal lines represent hypothetical genomes *A* and *B*. Each vertical box on the line is an ortholog and all orthologs are indexed according to their physical locations. The line connecting two boxes from one genome to another represents corresponding orthologs between two genomes. The solid lines connecting orthologs belong to some consecutive ortholog sets. The dotted lines connecting orthologs are orthologs pairs that are not in any consecutive ortholog set.

(3) CCV/CV-based phylogenetic distance

Gao et al. [45] used all the string appearance frequencies (strings of length k) to represent each genome, that is, each whole genome can be regarded as a high-dimension vector, where each vector component is the frequency of a particular combination of nucleotides (A, T, C, G). Then the pairwise distance between two genomes can be calculated as the Euclidean distance between the corresponding two vectors. Wu et al. [46] extended this idea to use all the string appearance frequencies to define a CCV-based phylogenetic distance. In this distance, strings of length from 1 to k are all employed.

Phylogenetic inference

All of methods mentioned above, including composite distance method, were used to generate a distance matrix between all pairs of genomes. Phylogenetic trees were then generated using the Neighbor-Joining [34,35] algorithm in Phylip (version 3.67) [47] (likelihood based approaches will be applied in future studies).

Performance measurement

There is no official standard for prokaryotic taxonomy. However, it is widely believed by microbiologists that the classification scheme in Bergey's Manual of Systematic Bacteriology [1] is the best approximation available. To measure the performances of a distance method, we calculated the percentage of agreed quartet topologies between the tree created by the method and the tree from the Bergey's taxonomy. A quartet topology is a subtree structure of the subset of 4 taxa (called a quartet). Given a quartet of taxa, a, b, c, and d, there are 3 possible ways to connect the taxa as terminals (Figure 4). Note that, the tree from Bergey's taxonomy is not a binary tree; therefore, one node may have more than two children. As a result, some of the quartets do not have any of the three quartet topol-

Table 2: Accuracy for different components combinations of our proposed distance method

Data sets	Number of species	GDD (%)	GCD (%)	GBD (%)	GCD*GDD (%)	GCD*GBD (%)	GDD*GBD (%)	GCD*GDD*GBD (%)
Dataset1	52	85.12	86.44	84.54	91.45	90.29	90.29	90.29
Dataset2	53	87.76	86.40	84.45	90.65	90.74	90.74	90.74
Dataset3	82	80.37	92.58	84.19	94.46	95.93	96.06	98.46
Dataset4	398	83.73	86.56	81.23	89.93	87.07	87.28	90.07
Dataset5	181	95.04	89.74	90.20	94.30	95.67	98.16	98.30
Dataset6	96	87.39	85.45	84.88	99.36	99.26	99.36	99.26
Dataset7	277	88.70	84.04	86.75	88.71	89.71	88.23	90.71
Dataset8	165	85.36	77.98	77.03	94.44	94.38	94.47	94.38
Dataset9	54	89.31	87.34	83.76	92.31	92.31	92.37	96.55

GDD = gene dispersion distance, GCD = gene content distance, GBD = gene breakpoint distance. GCD*GDD is the combination of two distance components and GCD*GDD*GBD is the combination of all three terms. All distances are logarithmically transformed.

ogies in Figure 4. We call a quartet is resolved if no more than three of taxa share the same parents. To measure the performance of our method, we will collect agreed quartets in which the quartets have the same topology between the tree from Bergey's taxonomy and a binary phylogenetic tree. The accuracy is percentage of the agreed quartets.

Results and discussion

Variance of orthlog definition

To test the robustness of our ortholog definition, different variations of E-value cut-offs and sequence identities have been selected for performance evaluation. Our results showed that E-value cut-offs lower than 10^{-3} do not have significant effects on the results. This is probably because the reciprocal Blast hit process would ensure the majority of the homolog selections as long as E-values are small enough. However, different selections of percentage identities have some impacts on the performance. A very high percentage identity would be too stringent to obtain enough number of orthologs, while a very low percentage identity would have too many false positive ortholog pairs. We used dataset 9 as representative dataset, described in the Method section, for testing orthologs selection with E-value below 10–3 and percentage identities of 10%, 20%, 30% and 40%. Accuracies of 76%, 82%, 89% and 85% were obtained using GDD method, respectively. Therefore, we selected the E-value cut-off at 10–3 and percentage identity of 30% for the optimal ortholog definition.

Comparison to single gene trees

Although attempts to explain the phylogeny of prokaryotes based on the ssu-rRNA have been quite successful [3,4], a well-known problem associated with this type of single-gene 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. We were able to obtain single-gene phylogenetic trees of 13 bacterial species [51] for comparing with our method. Figure 5 shows three different trees based on single-gene selection and one tree based on the whole-genome gene sets using our method. Using the accuracy measurement described in the Method section, accuracies of 88%, 81% and 83% are obtained for tree (a), (b) and (c), respectively, while tree (d) based on the whole-genome gene sets has a significantly higher accuracy of 91%. As an example, in tree (b) and (c), *Salinnibacter rubber*, which is a member of Bacteroidetes (B20), is placed outside its own phylum towards the protebacteria phylum (B12). In tree (d), *Rhodopirellula baltica* is placed closer to phyla Bacteroidetes (B20) and Chlorobi (B11), which is correct [52]. Huerta-Cepas *et al*. [53] found degrees of topological variations among single-gene phylogenies were much greater than previously thought. Their conclusions, although based on eukaryotes, may be applicable to the whole tree of life, and are probably even more important to the prokaryote phylogeny given more LGTs in prokaryotes.

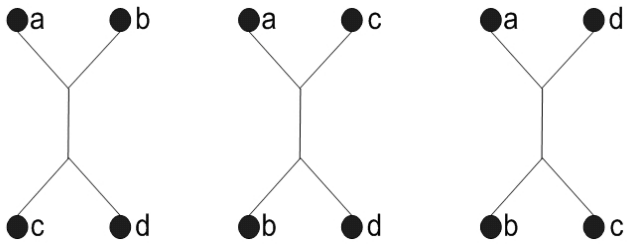

Figure 4
Three possible quarter topologies. Given four taxa (a, b, c, and d), there are only three unique ways to connect the taxa as terminals.

Composite distance as optimal distance measurement

We used 9 datasets, described in the Method section, for performance evaluations. In comparison to several other phylogenetic analysis methods, our composite distance method showed consistently better performance (see Tables 2 and 3). We have achieved above 90% accuracy comparing to Bergey's taxonomy system for all the selected datasets.

Table 2 shows the performance results of 7 different combinations of three composite distance components using the 9 datasets based on performance evaluation method described in the Method section. Table 3 compares the performances of 5 different methods including our proposed method. The accuracy is defined as the percentage of agreed quartets between the tree created by a distance method and the tree from the Bergey's taxonomy system. Due to the complexity of obtaining accurate LGT events and protein structural models for all the species to apply the Structural Domain method on the genome distance calculation, we could not obtain the genome phylogeny trees for all the 9 datasets except dataset 9, whose result can be directly downloaded from [31].

Results from Tables 2 and 3 suggest the optimal distance calculation is the composite distance, which combines the advantages of GDD, GCD and GBD. We use dataset 9 by Deeds [31] for detailed discussion. Dataset 9 consists of 8 Archaea (A2) species, 1 Bacteria Aquificae (B1) species, 1 Bacteria Thermotogae (B2) species, 2 Bacteria Cyanobacteria (B10) species, 21 Bacteria Proteobacteria (B12) species, 15 Bacteria Firmicutes (B13) species, 5 Bacteria Actinobacteria (B14) species, and 1 Bacteria Fusobacteria (B21) species. In this dataset, *Thermotoga maritima* (B2.1.1.1.1), a rod-shaped bacterium belongs to the order

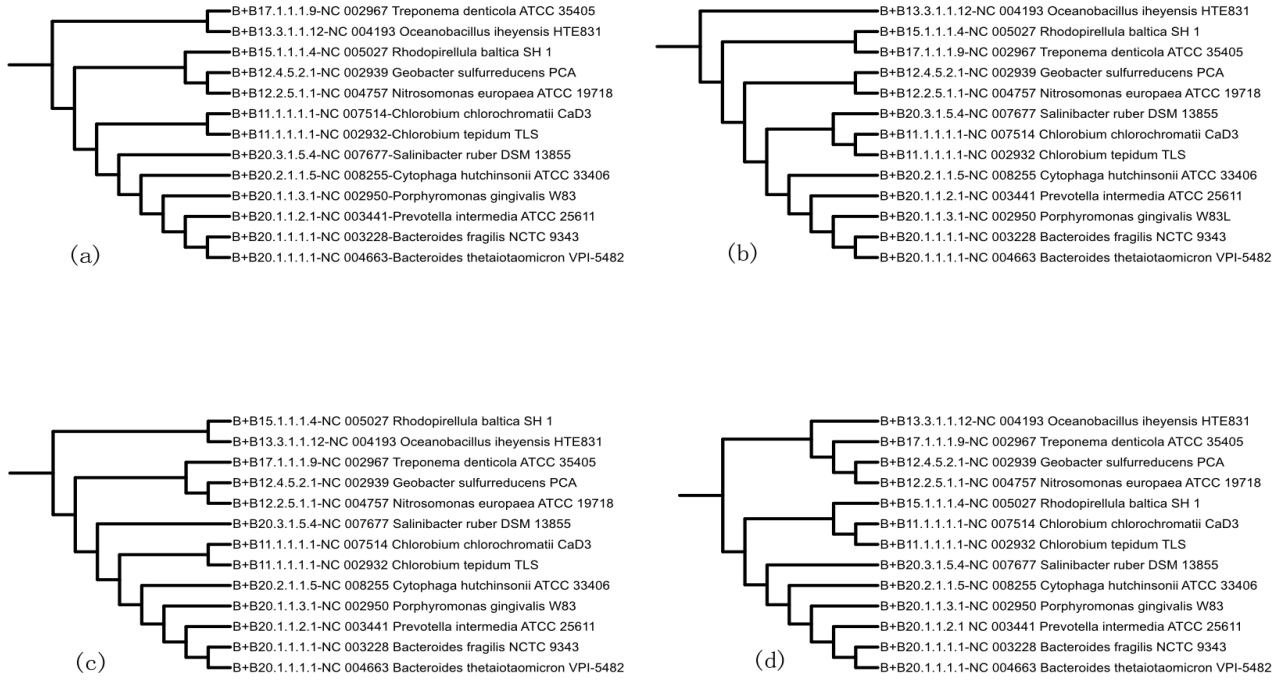

Figure 5

Phylogenetic trees based on different gene selections. (a) Phylogenetic tree of 13 bacterial species based on CTP synthase (*pyrG*) affiliates with *Bacteroidate*s; (b) phylogenetic tree of 13 bacterial species based on *glyA* affiliates with *Chlorobi*; (c) phylogenetic tree of 13 bacterial species based on Chaperonim Hsp 60 (*groEL*) affiliates with the superphylum *Bacteroidates-Chlorobi*. (d) phylogenetic tree of 13 bacterial species based on the whole-genome gene sets.

Thermotogales, contains 1,877 predicted coding regions, but it has only about 110 genes that have orthologs in the genomes of other thermophilic Eubacteria and Archaea. Completion of its genome has revealed a high degree of similarity with Archaea in terms of contents of genes and overall genome organization, where almost one quarter of the genome is Archaea in nature, instead of other bacterial phylum. Conservation of gene order between *T. maritima* and the Archaea species in many of the clustered regions suggests that LGT may have occurred between thermophilic Eubacteria and Archaea [36]. When classifying this species, our composite distance method moves *T. maritima* closer to the Archaea clade on the tree, which reflects this biological property. In contrast, other methods, such as CCV method and overlapping gene method, put *T. maritima* closer to either the Proteobacteria (B12) clade or the Firmicutes (B13) clade on the phylogenetic tree. Out of these three methods that do not model LGT events specifically, the composite distance method still demonstrates good sensitivity of the proposed composite distance method.

By comparing different trees generated from different distance measures, we found that GDD and GBD are more sensitive on deeper-lever of the trees. However, GCD is more accurate on higher levels, such as clades of the tree. For example, the gene content method puts five species of Diplococci (class) of Fiemicutes (B13) together with its neighboring class Bacilli of Fiemicutes, while GDD misclassifies those five species as members of the Proteobactria (B12) clade. As we can see from Table 2, by combining all three distance components together as the composite distance measure, the advantages of individual components are also combined together to achieve the optimal results. Other methods, such as structural domain method, position Archaea at a higher level phylum than Bacteria phylum on the phylogeny tree, which contradicts the taxonomy where they are at the parallel levels on the tree. In contrast, our composite distance method is consistent with the taxonomy. A possible reason could be the consideration of LGT events, as the structural domain method overly emphasizes the role of LGT genes from Archaea to Bacteria. The overlapping gene method however, has problem of using limited information to estimate the distance. Overlapping genes are conserved, but they represent a small number of genes in the genome and hence, the statistical errors may be large. This is shown by mis-classifying *Pyrococcus furiosus* (A2.6.1.1.3) into clade

Table 3: Accuracy comparison between our composite distance method and other methods

Data sets	Number of species	GCD (%)	OG (%)	CCV (k <= 5) (%)	CCV(k = 5) (%)	SDD (%)	Composite Distance (%)
Dataset1	52	86.44	83.93	87.82	88.29	NA	90.29
Dataset2	53	86.40	85.49	86.27	87.92	NA	90.74
Dataset3	82	92.58	84.35	95.97	91.54	NA	98.46
Dataset4	398	86.56	85.52	79.03	78.86	NA	90.07
Dataset5	181	89.74	80.34	87.19	87.19	NA	98.30
Dataset6	96	85.45	87.22	99.00	99.07	NA	99.26
Dataset7	277	84.04	81.89	83.28	83.19	NA	90.71
Dataset8	165	77.98	87.87	85.20	82.78	NA	94.38
Dataset9	54	87.34	88.27	91.39	91.47	81.57	96.55

GCD = gene content distance, OG = overlapping gene distance, SDD = structural domain distance.

of Bacilli (B13), instead of Archaea (A2) using the overlapping gene method.

It is interesting to note that GBD method by itself performed poorly for large-scale comparison of prokaryote genomes, which is in accordance with the commonly held view that breakpoint methods lead to reliable results only if the genomes are sufficiently co-linear, such as in datasets 5 and 6.

Efficiency comparison

Besides comparing the accuracies of different distance measures, we also consider computational efficiency. Most whole-genome phylogeny construction methods, including ours, require a process of defining orthologs through time-consuming BLAST. Assuming there are m genomes and each genome has roughly $\sim n$ orthologs, then the complexity for reciprocal BLAST hits in ortholog identifications would be $O(m^2n^2)$. Excluding this process, our method, which computes in linear time, takes much less computing time than other methods, especially the breakpoint distance measure [18]. Structural domain method, in another way, considers LGT events. It approximates the genome distance by continuously readjusting the protein structural domain graph when applying each LGT event (see the Method section for more details), and the running time could easily take up to hours if not days for a large genome data set. The CCV method, although not requiring a process of defining orthologs, considers every possible string of length up to k for whole genome sequences. This method requires even higher computational resources in terms of memory and CPU cycles. Overall, the composite distance measure shows not only the higher accuracy but also fast speed.

Further discussion

Given pairs of prokaryotic species, there could be situations where one genome is essentially a subset of another much larger genome. For example, in our datasets this is true of *Buchnera aphidicola* genome, which is essentially a subset of the *Escherichia coli* genome, with approximately 14% the size of it [49]. Shared genome sequences could make two different genomes seem to be more closely related than they actually are. We have tried to model this case by modifying the GCD formula since this method uses all the genes in genome pair as normalization factor, not just the orthologs. We used the smaller genome gene set size instead of summing two genome gene set sizes for normalization. This modified formula would consider the similar segments of genome at most once. However, the performance of this modified GCD formula decreased significantly from around 85% to 60% for most of the datasets. Given that the situation where one genome is part of another genome is rare, it appears that considering this in our distance calculation lost the generality of the method.

Although all our trees would be generated as binary trees, or phylograms, with two leaf species for each node and they are hard to compare to the taxonomy, which is usually not binary, we find our results are consistent with most of the taxonomy in Bergey's system based on the percentage of agreed quartet topologies. Nevertheless, we still mis-classify some species on the tree. For example, *Treponema pallidum* in class Spirochaetes (B17.1) is placed as a sib of the class phylum Diplococci (B14.1), which does not agree with current classifications. This may be because the genome *T. pallidum* has high number of LGTs, which is as high as 32.6% [37]. It would be hard for the current method to deal with such an extreme. Other cases, such as chimerical genomes [37] or paralogous genes, are not modeled in this study, but could have misleading effects on our classifications. Chimeric genomes, which could have happened due to LGT, would produce false lineage for the interested genomes. Paralogous genes would artificially shorten the distance between two genomes if there are many paralogs. Future developments of the tool would include events of LGT, gene copy number, and conservation of overlapping genes, as well as exclude genes with abnormal evolution rates. We can see that ComPhy provides a framework for incorporating more relevant biological aspects for distance measurement.

Conclusion

ComPhy, a stand-alone phylogeny construction tool, provides a robust and easy-to-use tool for biologists. It does not require multiple sequence alignment and is fully automated. ComPhy implements a composite distance method, which does not depend on any type of evolution models in calculating the distance between two genomes besides the protein sequences and gene physical locations. It allows users to infer phylogenies for any set of genomes of interest to study their evolutionary relationships by either generating a phylogram tree or a Newick format tree file for further study. Although the tool is built for complete-genome gene sets phylogeny, users can provide pre-defined ortholog sets to build the phylogeny according their criteria. The process takes less than a minute from given protein sequence files and protein location files to the outputs of trees for hundreds of species if excluding the BLASTP for generating orthologs. Although in the current stage of the application, our method works only for species with single chromosome, we will extend ComPhy to study eukaryotic genomes with improved methods working on multi-chromosomes. We believe this is a timely development as the whole-genome phylogeny becomes dominant with the arrival of more complete genome sequences, especially from the metagenomic analyses of microbial communities [38,39].

Availability

ComPhy, all the datasets (including genome sequences, gene location files and Bergey's code), and the phylogenetic trees generated in this study are available at http://digbio.missouri.edu/ComPhy.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GNL carried out the phylogeny constructions and drafted the manuscript. ZC and GL designed the datasets and provided the performance evaluation codes. SC provided some ideas and formulations. DX conceived and coordinated the study. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by NSF/ITR-IIS-0407204. Some computation of this project was carried out using the UMBC Computing Resources at the University of Missouri. We like to thank Profs. Bolin Hao and Chris Pires for helpful discussions, and JianJiong Gao for proofreading.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 1, 2009: Proceedings of The Seventh Asia Pacific Bioinformatics Conference (APBC) 2009. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S1

References

1. Bergey's Manual Trust: **Bergey's Manual of Determinative Bacteriology.** 9th edition. *Williams & Wilkins, Baltimore. MD*; 1994.
2. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci* 1977, **74:**5088-5090.
3. Olsen GJ, Woese CR: **The wind of (evolutionary) change: breathing new life into microbiology.** *J Bacteriol* 1994, **176:**1-6.
4. Huynen MA, Bork P: **Measuring genome evolution.** *Proc Natl Acad Sci USA* 1998, **95:**5849-5856.
5. Ludwig W, Schleifer K-H: **Phylogeny of bacterial beyond the 16S rRANA standard.** *ASM News*; 1999:752-757.
6. Teichmann SA, Mitchison G: **Is there a phylogenetic signal in prokaryote proteins?** *J Mol Evol* 1999, **49:**98-107.
7. Doolittle WF: **Phylogenetic classification and universal tree.** *Science* 1999, **284:**2124-2129.
8. Daubin V, *et al.*: **Phylogenetics and cohesion of bacterial genomes.** *Science* 2003, **301:**829-832.
9. Rokas A, Williams AL, King N, Carroll SB: **Genome-scale approaches to solving incongruence in molecular phylogenies.** *Nature* 2003, **425:**798-804.
10. Ge F, *et al.*: **The cobweb of life revealed by genome-scale estimates of horizontal gene transfer.** *PLoS Biol* 2005, **3:**e316.
11. Ciccarelli FD, *et al.*: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311:**1283-1287.
12. Daubin V, *et al.*: **A phylogenetic approach to bacterial phylogeny: evidence of a core of genes sharing a common history.** *Genome Res* 2002, **12:**1080-1090.
13. Goremykin VV, Hellwig FH: **Evidence for the most basal split in land plants diving Bryophyte and Tracheophyte lineages.** *Plant Syst Evol* 2005, **254:**93-103.
14. Hannenhalli S, Pevzner PA: **Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals.** *JACM* 1999, **46:**1-27.
15. Kececioglu J, Ravi R: **Of mice and men. Evolutionary distances.** *Proceedings of the 6th ACM-SIAM Symposium on Discrete Algorithms* 1995:604-613.
16. Kececioglu J, Ravi R: **Reconstructing a history of recombinations from a set of sequences.** *Discrete Appl Math* 1998, **88:**239-260.
17. Boore JL, Brown WM: **Big trees from little genomes: mitochondrial gene order as a phylogenetic tool.** *Curr Opin Genet Dev* 1998, **8:**668-674.
18. Sankoff D, Blanchette M: **Multiple genome rearrangement and breakpoint phylogeny.** *J Comput Biol* 1998, **5:**555-570.
19. Sankoff D: **Geneome rearrangement with gene families.** *Bioinformatics* 1999, **15:**909-917.
20. Sankoff D: **Comparative mapping and genome rearrangement.** *FromJay Lush to Genomics: VisionsFor Animal Breeding and Genetics* 1999:124-134.
21. Berman P, Hannenhalli S, Karpinski M: **Approximation algorithm for sorting by reversals.** In *Technical Report TR01-047 ECCC*; 2001.
22. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nat Genet* 1999, **21:**108-110.
23. Henz SR, Huston DH, Auch AF, Nieselt-Struwe K, Schuster SC: **Whole Genome-based Prokaryotic Phylogeny.** *Bioinformatics* 2004, **21:**2329-2335.
24. Luo Y, *et al.*: **BPhyOG: An interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes.** *BMC Bioinformatics* 2007, **8:**266.
25. Johnson ZI, Chisholm SW: **Properties of overlapping genes are conserved across microbial genomes.** *Genome Res* 2004, **14:**2268-2272.
26. Lin J, Gerstein M: **Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels.** *Genome Res* 2000, **10:**808-818.
27. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: **Universal trees based on large combined protein sequence data sets.** *Nat Genet* 2001, **28:**281-285.
28. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades.** *BMC Evol Biol* 2001, **1:**8.
29. Korbel JO, Snel B, Huynen MA, Bork P: **SHOT: A web server for the construction of genome phylogenies.** *Trends Genet* 2002, **18:**158-162.

30. Mirkin BG, Fenner TI, Galperin MY, Koonin EV: **Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes.** *BMC Evol Biol* 2003, **3:**2.
31. Deeds EJ, Heneessey H, Shakhnovich EI: **Prokaryotic phylogenies inferred from protein structural domains.** *Genome Res* 2005, **15:**393-402.
32. NCBI: **Microbia complete genomes taxonomy.** 2007 [ftp://ftp.ncbi.nih.gov/genomes/Bacteria/].
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
34. Saitou N, Nei M: **The neighbor-Joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
35. Studier JA, Keppler KJ: **A note on the neighbor-joining algorithm of Saitou and Nei.** *Mol Biol Evol* 1988, **5:**729-731.
36. Worning , Peder , *et al.*: **Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritime.** *Nucleic Acids Res* 2000, **28:**706-709.
37. Gross , Jeferson , Meurer , Jörg , Bhattacharya , Debashish : **Evidence of a chimeric genome in the cyanobacterial ancestor of plastid.** *BMC Evolutionary Biology* 2008, **8:**117.
38. Pope PB, Patel BKC: **Metagenomic analysis of a freshwater toxic cyanobacteria bloom.** *FEMS Microbiology Ecology* 2008, **64(1):**9-27.
39. Jones BV, Marchesi JR: **Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome.** *Nature Methods* 2007, **4:**55-61.
40. Wang LS, Jansen RK, Moret BME, Raubeson LA, Warnow T: **Fast phylogenetic methods for the analysis of genome rearrangement data: an empirical study.** *Pac Symp Biocomput.* 2002:524-535.
41. Felsenstein J: *Inferring phylogenies Sinauer Associates, Mass*; 2004:158-159.
42. Holm L, Sander C: **The FSSP database: Fold classification based on structure-structure alignment of proteins.** *Nucleic Acids Res* 1996, **24:**206-209.
43. Dietmann S, Holm L: **Identification of homology in protein structure classification.** *Nat Struct Biol* 2001, **8:**953-957.
44. Albert R, Barabasi A-L: **Statistical mechanics of complex networks.** *Rev Mod Phys* 2002, **74:**47-97.
45. Gao L, Qi J, Sun J, Hao B: **Prokaryote phylogeny meets taxonomy: an exhaustive Comparison of composition vector trees with systematic.** *Science in China* 2007, **50:**587-599.
46. Wu X, Cai Z, Wan XF, Hoang T, Goebel R, Lin G: **Nucleotide composition string selection in HIV-1 subtyping using whole genomes.** *Bioinformatics* 2007, **23(14):**1744-1752.
47. Felsenstein J: **PHYLIP – Phylogeny inference package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
48. Feinberg L, Srikanth R, Vachet R, Holden J: **Constraints on Anaerobic Respiration in the Hyperthermophilic Archaea *Pyrobaculum islandicum* and *Pyrobaculum aerophilum*.** *Appl Environ Microbiol* 2008, **74:**396-402.
49. Moran NA, Mira A: **The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*.** *Genome Biol* 2001, **2:**1-12.
50. Woese CR: **Baterial evolution.** *Microbiol Rev* **51:**221-272.
51. Soria-Carrasco V, Valens-Vadell M, Peña A, Antoìn J, Amann R, Castresana J, Rosselloì-Mora R: **Phylogenetic position of Salinibacter ruber based on concatenated protein alignments.** *Systematic and Applied Microbiology* 2007, **30(3):**171-179.
52. Glockner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzym K, Heitmann K, Rabus R, Schlesner H, Amann R, Reinhardt R: **Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1.** *Proceedings of the National Acedemy of Sciences* 2003, **100:**8298-8303.
53. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T: **The human phylome.** *Genome Biol* 2007, **8:**R109.