Phylogeny from Whole Genome Comparison*

Graziela S. Araújo and Nalvo F. Almeida Jr. **

Departamento de Computação e Estatística
Universidade Federal de Mato Grosso do Sul
CP 549 - 79070-900 Campo Grande MS, Brazil
{nalvo,gsa}@dct.ufms.br

Abstract. In this paper, we propose and compare several distance-based phylogenetic measures from whole genome comparison which take into account gene content and gene order conservation. We show that it is possible to infer phylogenetic issues from whole genome comparison, since gene content and gene order are conserved between closely related species.

1 Introduction

The increasing availability of complete genomes has created the need for tools to analyze and compare them, and at same time has provided a useful data set for phylogenetic reconstruction. In general, phylogenetic inference is based on comparison of homologous sequences. However, whole genome comparison can be useful by providing information about relationships involving their genes, or more precisely, their predicted proteins.

Here, we present a study of how pairwise whole genome comparison can contribute to phylogenetic inferences. Specifically, we test for some distance-based measures from gene content and gene order conservation of two prokaryote genomes. Our main goal is to gain a better understanding of changes in gene organization between different species, and how these changes can elucidate some issues about evolution, since we know that gene content and gene order are conserved between closely related prokaryote species [22, 23].

This paper is structured as follows. In section 2 we describe what kind of data we want from whole genome comparison. In section 3 we propose six distance-based measures for phylogeny inference, based in two main approaches. In section 4 we show how we combine the information from whole genome comparison with a distance-based method for phylogeny reconstruction. Finally, in section 5, we show our results and make some final remarks.

2 Genome comparison data

Whole genome comparative analysis, specifically involving gene content and gene order conservation, is a powerful tool for studies of genomic evolution. Many

* This research is partly founded by FUNDECT-MS.
** To whom correspondence should be addressed.

studies have been proposed about whole genome comparison and its benefits [13, 15, 18, 20, 21, 24]. For our purposes, the main idea is to infer phylogenetic information from pairwise whole genome comparison at level of their predicted proteins to gain a better understanding of changes in gene organization between different species.

In this section, we will describe what kind of data we are interested from pairwise whole genome comparison to build distance-based phylogenetic data.

We will get data from a tool developed by Almeida *et.al* [1, 2]. That tool, called EGG (Extended Genome-Genome comparison), finds all pairs of orthologous genes by using BLASTP program [3, 4]. The comparison takes all the predicted proteins of both genomes into account, following all-against-all fashion. After that, a bipartite graph is built, where an edge represents a pair of orthologous genes. An edge of this graph is called a *match*. Formally, a match is pair (g, h) of genes whose BLASTP e-value is not greater than 10^{-5} and the alignment includes at least 60% of each sequence. Note that a gene can be in several matches. When a gene h is the best BLASTP hit found by g and vice versa, we have a *bi-directional best hit (BBH)*. Thus, a gene can participate at most in one BBH. Actually, BBHs try to minimize interference from paralogous genes. EGG has been used successfully in some genome projects [9, 27].

After graph construction, EGG looks for organization structures in it, called *orthologous regions*. Basically, an orthlogous region is a region (in both genomes) of closely matches (more details can be seen in [1]).

Obviously, the similarity between two species can be measured in terms of the existence of matches, BBHs and orthologous regions, since such structures can be interpreted in terms of evolutionary events, like gene loss, for example. Thus, we will use these three structures to make some estimations for whole genome distance-based phylogenetic measures.

3 Two approaches for whole genome distance-based phylogeny

In order to understand evolutionary issues of some organisms, we focus on two levels of organization of the genomes, leading to two different approaches for building evolutionary distance data from whole genome comparison.

In the first level, we take into account the presence-absence of each individual gene. The idea is to evaluate the phylogenetic relationships of two genomes by looking for pairs of orthologous genes between them. For this approach, we can use matches and BBHs, as seen in section 2. In the other level, we look for clusters of closely pairs orthologous genes which would correspond to genomic elements that have been conserved during evolution. For this approach we can use orthologous regions, also described in section 2. We then test a total of six evolutionary genome distance measures from a whole genome comparison, where four of them are for the first level of organization and other two are for the second one.

Let us consider two genomes G and H with $|G|$ and $|H|$ genes, respectively. Let M denote the set of matches, B denote the set of BBHs and R denote the set of all orthologous regions between genomes G and H.

For the first approach, we begin with a measure, D_1, based on the number of matches between G and H, namely the ratio between $|M|$ and $|G| \cdot |H|$ (the maximum number of possible matches). The second measure, D_2, is based on the ratio between B and $\min\{|G|, |H|\}$ (the maximum number of BBHs). Note that M and B tend to be greater as the phylogenetic distance decreases. In order to fix this, we need to invert the estimations. Thus,

$$D_1 = \frac{1}{\frac{|M|}{|G| \cdot |H|}} \quad \text{and} \quad D_2 = \frac{1}{\frac{|B|}{\min\{|G|, |H|\}}}.$$

The third estimation of distance of two genomes is based on the BLAST scores of the all matches found, whereas the fourth one is based on the same kind of score, but now considering only BBHs. Thus,

$$D_3 = \frac{1}{\sum_{(g,h) \in M} s(g,h)} \quad \text{and} \quad D_4 = \frac{1}{\sum_{(g,h) \in B} s(g,h)},$$

where $s(g,h)$ is the BLAST alignment score of genes g and h.

For the last two measures, let us denote as m_r and b_r respectively the number of matches and BBHs of a region r found between G and H. Thus,

$$D_5 = \frac{1}{\sum_{r \in R} m_r} \quad \text{and} \quad D_6 = \frac{1}{\sum_{r \in R} b_r}.$$

Considering that gene content and gene order conservations are good clues for elucidate relationships between two genomes [7, 22–24], we believe that these measures can be useful for phylogenetic inferences.

4 Methodology

In order to test those measures presented in the previous section, we have chosen a set of six prokaryote publicly available genomes, *Escherichia coli K-12 MG1655* (EC), *Pseudomonas aeruginosa PA01* (PA), *Staphylococcus aureus MU50* (SM), *Staphylococcus aureus N315* (SN), *Salmonella typhi* (ST), and *Salmonella typhimurium LT12* (SL). We have made such a choice in order to have at least two pairs of very related genomes, namely SN-SM and ST-SL.

We built six distance matrices, one for each measure, by comparing all-against-all six genomes and calculating, for each pair of them, $D_1, \ldots, D_6$.

For constructing the corresponding trees, we used the well-known *Neighbor-Joining* algorithm [19], available in the Phylip package [10, 11]. Neighbor-Joining algorithm is one of the most popular distance-based methods for building phylogenetic trees. Roughly, Neighbor-Joining method calculates a score for each

pair of species. Each step of the algorithm consists in joining the pair with the minimum score, creating a new internal node; this new node replaces the two chosen species in the matrix and the distances to this new node are recomputed. The output is an unrooted tree.

Several studies have showed that Neighbor-Joining method has a reliable effectiveness (accuracy and performance) for distance-based data [5, 8, 17, 25, 26]. That is the main reason why we have chosen it for testing our measures.

To evaluate the measures proposed in section 3, first we have built a distance matrix for six 16S ribosomal RNA (rRNA) sequences, one of each chosen species. All the 16S rRNA sequences were obtained from Genbank [6], available at NCBI website. The method used for calculating that matrix was DNADIST [12], a program to compute distance matrix from nucleotide sequences, also available in the Phylip package. Below, the 16S matrix we have calculated.

	SM	SN	EC	PA	SL	ST
SM	0.0000	0.0019	4.7443	5.9752	5.2363	4.9180
SN		0.0000	4.7443	5.9752	5.2363	4.9180
EC			0.0000	4.1879	0.0314	0.0287
PA				0.0000	4.0768	4.2328
SL					0.0000	0.0066
ST						0.0000

Although our analysis is independent of phylogenies based on the level of sequence identity of individual genes, the reason why we have decided to use 16S rRNA sequences for evaluating our measures is because 16S rRNA sequences are present in almost all currently available genomes, and mainly because they are recognized and used as potential markers for phylogenetic inferences [16].

Secondly, we again used Neighbor-joining algorithm for building that we called 16S tree, showed in figure 1. Note that the algorithm for 16S sequences was able to join the more related species accordingly, like SM and SN, and SL and ST. We assume that the 16S tree is the true one for our set of species.

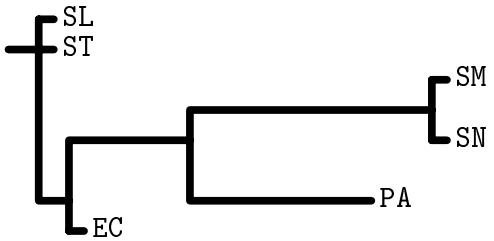

Fig. 1. Tree obtained from 16S rRNA sequences. Branch sizes represent relative distances between the species.

5 Results and discussion

Now we will show the trees $T_1, \ldots, T_6$, obtained from the distance measures proposed in section 3, $D_1, \ldots, D_6$, respectively. Although the branch size represents an estimate for the matrix distance, we believe that the main factor in evaluating the measures is the ability to topologically join closely related species. So, we are more concerned about the topology of the trees than about the branch sizes, since one good measure should at least lead to a tree whose topology agrees with the true one.

Figure 2(a) shows tree T_2. Trees T_3, T_4 and T_5 have the same topology, that agrees with the topology of 16S tree. Tree T_1 is showed in figure 2(b). The same topology was obtained for T_6. Although both trees T_1 and T_6 were able to join the pairs of closely related species, their topology is slightly different than 16S one.

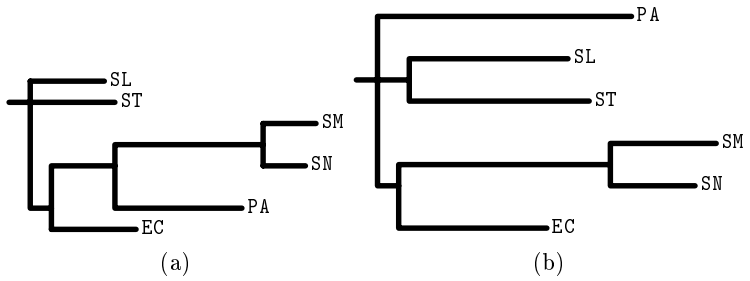

Fig. 2. Tree T_2 (a), whose distance matrix is based on the number of BBHs between two genomes, and tree T_1 (b), where distance data is based on the number of pairs of matches.

Trees of the first approach for distance-based phylogeny, based on gene content, was able to get the desired topology, except that one based on the number of the matches, namely T_1. This fact give us two clues about phylogeny based in whole genome comparison. Firstly, gene content is in fact a reliable instrument for building whole genome distance-based phylogeny measures. Secondly, BBHs tend to be more helpful in such kind of inference, since a match is a weaker relationship, which can occur even in pair of distant genomes.

For the second approach, where the distances are based on orthologous regions, the opposite took place. The tree based on the number of BBHs per region is slightly different than 16S tree, whereas that one based on the number of matches per region was good. This makes sense, since we could detect, specially between closely related genomes, an impressive number of small regions with large number of matches inside of them. Beside that, lots of isolated BBHs (outside orthologous regions) were detected. These BBHs do not contribute for that kind of measure.

Although the tree based in BBHs per region was not so good, it is not a good idea just to reject this kind of measure, because EGG works in a sensitive way, in the sense that the orthologous regions are found basically by joining close matches, without any other criterion. This may be causing the determination of some small regions just by chance. That is another probable reason why the number of matches inside a region have worked well.

These results are currently being analyzed, and we are sure much remains to be done on the methodology that have been described. For example, a more detailed analysis of the quality of the regions found by EGG is necessary. Other possible ongoing work is about what kind of phylogenetic information we can extract from whole genome comparison. It seems that just a distance measure between two genomes is poor, compared to the large amount of information we can take from this kind of comparison. Gallut and colleagues [14], for example, explored gene order information by building character-based data for inferring phylogeny.

We hope to make new progress in this study by exploring these different approaches, but for now we can conclude that gene content and gene order conservations can be helpful to infer important issues about prokaryote phylogeny.

References

1. N.F. Almeida. *Tools for genome comparison.* PhD thesis, Institute of Computing, University of Campinas, May 2002. In Portuguese.
2. N.F. Almeida and J.C. Setubal. Egg: Extended genome-genome comparison. In preparation.
3. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
4. S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acid Research*, 25:3389–3402, 1997.
5. K. Atteson. The performance of the Neighbor-Joining method of phylogeny reconstruction. In *COCOON*, pages 101–110, 1997.
6. D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, B. Rapp, and D. Wheeler. Genbank. *Nucleic Acids Res.*, 28(1):15–18, 2000.
7. P. Bork, B. Snel, G. Lehmann, M. Suyama, T. Dandekar, W. Lathe III, and M. Huynen. Comparative genome analysis: exploiting the context of genes to infer evolution and predict function. In *Comparative genomics*, pages 281–294. Kluwer Academic Publishers, 2000.
8. M.E. Cosner, R. Jansen, B. Moret, L. Raubeson, L. Wang, and S. Wyman. A new fast heuristic for computing the breakpoint phylogeny and experimental analyses of real and synthetic data. In *Proceedings of the 8th Internatinal conference on intelligent systems for molecular biology*, pages 104–115, La Jolla, USA, 2000.
9. A.C. Rasera da Silva, J.C. Setubal, and N.F. Almeida et al. Comparison of the genomes of two *xanthomonas* pathogens with differing host specificities. *Nature*, 417(6887):459–463, 2002.
10. J. Felseinstein. Phylip – phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

11. J. Felseinstein. Phylip (phylogeny inference package) version 3.5c. Distributed by the author, 1993. Department of Genetics, University of Washington, Seattle.
12. J. Felsenstein. *DNADIST, a program to compute distance matrix from nucleotide sequences.*
13. W. Fujibuchi, H. Ogata, H. Matsuda, and M. Kanehisa. Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. *Nucleic Acids Research*, 28:4029–4036, 2000.
14. C. Gallut, V. Barriel, and R. Vignes. Gene order and phylogenetic information. In *Comparative genomics.* Kluwe Academic Publishers, 2000.
15. M. Huynen, B. Snel, W. Lathe III, and P. Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Research*, 10:1204–1210, 2000.
16. W. Ludwig and K. Schleifer. Phylogeny of bacteria beyond the 16S rRNA standard. *ASM News*, 1999.
17. L. Nakhleh, B.M. Moret, U. Roshan, K.St. John, J. Sun, and T. Warnow. The accuracy of phylogenetic methods for large datasets. In *proceedings of fifth pacific symp. of Biocomputing (PSB'02)*, pages 211–222, Hawaii, USA, 2002.
18. R. Overbeek, M. Fonstein, M. D'Souza, G. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *PNAS*, 96:2896–2901, 1999.
19. N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, 1987.
20. J.C. Setubal and N.F. Almeida. Detection of related genes in procaryotes using syntenic regions. In *DIMACS Workshop on Whole Genome Comparison.* DIMACS Center, Rutgers University, February 2001.
21. B. Snel, G. Lehmann, P. Bork, and M.A. Huynen. STRING: a web-server to retrieve and display the repeatedly occuring neighbourhood of a gene. *Nucleic Acids Research*, 28:3442–3444, 2000.
22. M. Suyama and P. Bork. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends in Genetics*, 17(1):10–13, January 2001.
23. J. Tamames. Evolution of gene order conservation in prokaryotes. *Genome Biology*, 2(6), 2001.
24. J. Tamames, G. Casari, C. Ouzounis, and A. Valencia. Conserved clusters of functionally related genes in two bacterial genomes. *Journal of Molecular Evolution*, 44:66–73, 1997.
25. Y. Tateno, N. Takezaki, and M. Neiji. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evolution*, 11(2):261–277, 1994.
26. L. Wang, R. Jansen, B. Moret, L.Raubeson, and T. Warnow. Fast phylogenetic methods for genome rearrangement evolution: an empirical study. In *proceedings of fifth pacific symp. of Biocomputing (PSB'02)*, pages 524–535, Hawaii, USA, 2002.
27. D.W. Wood, J.C. Setubal, and N.F. Almeida et al. The genome of *agrobacterium tumefaciens*: insights into the evolution and evolution of a natural genetic enginner. *Science*, 294:2317–2323, December 2001.