BioSystems 124 (2014) 86–94

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Reconstruction of phylogenetic trees of prokaryotes using maximal common intervals

Mahdi Heydari[a], Sayed-Amir Marashi[b,c,*], Ruzbeh Tusserkani[d], Mehdi Sadeghi[e]

[a] *Department of Algorithms and Computation, College of Engineering, University of Tehran, Tehran, Iran*
[b] *Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran*
[c] *School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*
[d] *School of Mathematics, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*
[e] *National Institute of Genetic Engineering and Biotechnology, Tehran, Iran*

ARTICLE INFO

Article history:
Received 28 December 2012
Received in revised form 13 August 2014
Accepted 1 September 2014
Available online 4 September 2014

Keywords:
Gene order
Breakpoint distance
Common interval
Genomic rearrangement
Comparative genomics
DCJ

ABSTRACT

One of the fundamental problems in bioinformatics is phylogenetic tree reconstruction, which can be used for classifying living organisms into different taxonomic clades. The classical approach to this problem is based on a marker such as 16S ribosomal RNA. Since evolutionary events like genomic rearrangements are not included in reconstructions of phylogenetic trees based on single genes, much effort has been made to find other characteristics for phylogenetic reconstruction in recent years. With the increasing availability of completely sequenced genomes, gene order can be considered as a new solution for this problem. In the present work, we applied maximal common intervals (MCIs) in two or more genomes to infer their distance and to reconstruct their evolutionary relationship. Additionally, measures based on uncommon segments (UCS's), i.e., those genomic segments which are not detected as part of any of the MCIs, are also used for phylogenetic tree reconstruction. We applied these two types of measures for reconstructing the phylogenetic tree of 63 prokaryotes with known COG (clusters of orthologous groups) families. Similarity between the MCI-based (resp. UCS-based) reconstructed phylogenetic trees and the phylogenetic tree obtained from NCBI taxonomy browser is as high as 93.1% (resp. 94.9%). We show that in the case of this diverse dataset of prokaryotes, tree reconstruction based on MCI and UCS outperforms most of the currently available methods based on gene orders, including breakpoint distance and DCJ. We additionally tested our new measures on a dataset of 13 closely-related bacteria from the genus *Prochlorococcus*. In this case, distances like rearrangement distance, breakpoint distance and DCJ proved to be useful, while our new measures are still appropriate for phylogenetic reconstruction.

1. Introduction

The usual way to reconstruct the phylogenetic tree of prokaryotes is to use the sequences of their 16S rRNA gene (Hao and Gao, 2008). However, it is suggested that tree reconstruction merely based on a single gene is not sufficient to explain many evolutionary events, like insertions, deletions, or horizontal gene transfer (Suyama and Bork, 2001). Different strategies like phylogenomics and supertree reconstruction are proposed to address the same issue in phylogenetic reconstruction (Sanderson et al., 1998; Sicheritz-Pontén and Andersson, 2001; Soltis and Soltis, 2001). As a result, some studies suggested using gene order of the genomes as an alternative source of information for reconstructing phylogenetic trees (Belda et al., 2005; Blin et al., 2005; Luo et al., 2008; Moret et al., 2001). Using these methods, one can obtain phylogenetic trees which take into account the evolutionary history of a genomic sequence. These trees are usually consistent with our knowledge about the phylogenetic relationships of different species (Markov and Zakharov, 2009). Therefore, such trees can be used, in combination with standard methods like 16S rRNA-based trees, to provide a more comprehensive picture of the phylogenetic relations.

Chromosome (genome) rearrangements, as the evolutionary events which shape the genomic structure, were first described more than seventy years ago, where the concept of "breakpoints" (disruption of gene orders) was originally introduced (Dobzhansky and Sturtevant, 1938; Sturtevant and Dobzhansky, 1936). Following the ideas presented in those classical papers, it was suggested that

[*] Corresponding author.
E-mail address: marashi@ut.ac.ir (S.-A. Marashi).

http://dx.doi.org/10.1016/j.biosystems.2014.09.002
0303-2647/© 2014 Elsevier Ireland Ltd. All rights reserved.

the evolutionary distance of two genomes can be estimated by inferring the rearrangement events (Hannenhalli and Pevzner, 1995a; Hannenhalli and Pevzner, 1995b). Typically, in such studies and other similar works, only "common genes" in all genomes are taken into account (Belda et al., 2005; Luo et al., 2008; Markov and Zakharov, 2009). Mathematically speaking, these methods analyze gene permutations (GP) rather than gene order sequences (GOS) (El-Mabrouk and Sankoff, 2012). We will define these terms in the next section. Consequently, these methods neglect gene gain and gene loss events.

In this paper, based on the concept of common intervals (Schmidt and Stoye, 2004; Uno and Yagiura, 2000), we present new measures of pairwise genome distance, which can be used for phylogenetic tree reconstruction. These measures are suitable for the analysis of distant genomes, as they do not require removal of uncommon genes. We show that the phylogenetic trees based on these measures are consistent with reference trees obtained from 16S rRNA or NCBI Taxonomy Browser.

2. Basic definitions

2.1. Gene order sequence (GOS)

Let $\Sigma = \{1, \ldots, n\}$ be a finite set of genes. A gene order sequence (GOS), $G = (\gamma_1, \gamma_2, \ldots, \gamma_t)$, is defined as an ordered list of genes, i.e., $\gamma_1 \in \Sigma$ for all $1 \leq i \leq l$. Position of each gene γ_1 is simply its index i. In general, l can be greater than n since a GOS can contain repetitive elements. Please note that each γ_1 can be labeled as "+" or "−" depending on their orientation on the genome. If this is the case, then the GOS is a signed GOS, otherwise it is unsigned.

Let the interval $[x,y]$ denote the set $\{x, x+1, \ldots, y-1, y\}$, with $x < y$. We define $\Gamma_{[x,y]}$ (G) as the set of genes appearing between the positions x and y of G. In general, it is possible to have two GOS's, G^1 and G^2, with $\Gamma_{[1,n]}$ $(G^1) \neq \Gamma_{[1,n]}$ (G^2).

2.2. GOS common intervals

Informally speaking, two genomic segments are GOS common intervals if, regardless of repeated genes in each segment, they contain the same set of genes.

Suppose that two GOS's A and B of set $\{1, \ldots, n\}$ are given as input. A pair of intervals $([x_A,y_A], [x_B,y_B])$ with $1 \leq x_A < y_A \leq n$ and $1 \leq x_B < y_B \leq n$ is called a common interval if it satisfies $\Gamma_{[x_A,y_A]}(A) = \Gamma_{[x_B,y_B]}(B) = M$. The common interval size, $|M|$, is equal to the number of different genes in each interval.

Let A and B be two GOS's. The ordered pair $([i,j], [i',j'])$ of two gene order sequences A and B is a maximal common interval (MCI) if there exists no different common interval $([x_A,y_A], [x_B,y_B])$ such that $[i,j] \subsetneqq [x_A,y_A]$ or $[i',j'] \subsetneqq [x_B,y_B]$. See Fig. 1 for an illustrative example.

2.3. Gene permutation (GP)

A gene permutation P on n different genes is a rearrangement of these genes into a particular order. Therefore, there is a one-to-one correspondence between the elements of each two GPs P^1 and P^2, i.e., $\Gamma_{[1,n]}$ $(P^1) = \Gamma_{[1,n]}$ (P^2). In other words, if two genomic regions with n different genes have the same gene content then each region can be considered as a permutation of the other. Please note that in this definition gene duplications are not allowed.

3. Measures of evolutionary distance of two genomes

Evolutionary rearrangement events do not occur very often (Fertin et al., 2009). Therefore, the whole genomic structures evolve usually slower than DNA sequences (Morozov et al., 2013). Measures which are introduced in this study will reflect the influence of these evolutionary events including insertion and deletion in similarity or discrepancy score.

A measure of evolutionary distance between a pair of genomes is a measure to estimate how different the two genomes are. Having a measure for evolutionary distance, distance matrix is an $n \times n$ symmetric matrix D in which D_{ij} represents the evolutionary distance of the genomes i and j. In this manuscript, we focus only on those measures in the literature which are based on gene orders in genomes.

3.1. Rearrangement distance measures

The strategies to estimate distance measures can be categorized into two main groups. In the first type of strategies, a certain number of predefined genomic "rearrangement events" are considered. The distance measure can be computed by solving the "rearrangement problem", i.e., the problem of finding a minimum number of rearrangement events necessary to transform original GOS to a target GOS (Delgado et al., 2010). By solving this problem the distance of the two genomes can be determined. Distance measure computing strategies in this category may differ in the allowed "rearrangement events", or in the schemes of rearrangement event penalties.

If all of the genomes can be written as GPs of the same set of genes, then the rearrangement problem is tractable in polynomial time (El-Mabrouk and Sankoff, 2012). However, when other evolutionary events like duplication are taken into account, most strategies to solve this problem become NP-hard (Blin and Rizzi, 2005; Delgado et al., 2010). Some authors have suggested

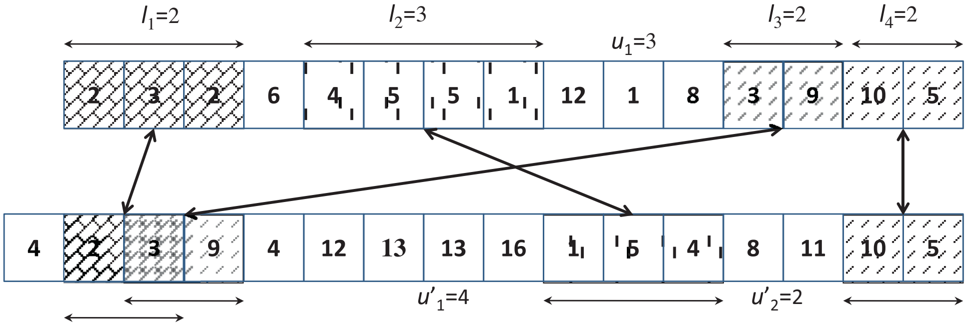

Fig. 1. Schematic representation of maximal common intervals and uncommon segments between two GOS's. In our analysis, the two GOS's are not necessarily of the same length and may not have the same gene content. Minimum size of maximal common intervals and also uncommon segments were assumed to be 2. In the first GOS there exists one uncommon segment with size u_1, while in the second GOS there are two uncommon segments with size u'_1 and u'_2. Additionally, there are four maximal common intervals between the two GOS's with sizes $l_1, \ldots, l_4$. Note that in the second GOS, two common intervals have overlap.