Genomics Proteomics Bioinformatics xxx (2017) xxx–xxx

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

OrthoGNC: A Software for Accurate Identification of Orthologs Based on Gene Neighborhood Conservation

Soheil Jahangiri-Tazehkand [1,a], Limsoon Wong [2,b], Changiz Eslahchi [1,*,c]

[1] *Department of Computer Science, Shahid Beheshti University, Tehran 1983969411, Iran*
[2] *School of Computing, National University of Singapore, Singapore 117417, Singapore*

Received 22 April 2017; revised 17 July 2017; accepted 28 July 2017
Available online xxxx

Handled by James J. Cai

KEYWORDS

Orthology;
Homology;
Gene neighborhood conservation;
Genomic context;
Comparative genomics

Abstract **Orthology** relations can be used to transfer annotations from one gene (or protein) to another. Hence, detecting orthology relations has become an important task in the post-genomic era. Various genomic events, such as duplication and horizontal gene transfer, can cause erroneous assignment of orthology relations. In closely-related species, gene neighborhood information can be used to resolve many ambiguities in orthology inference. Here we present OrthoGNC, a software for accurately predicting pairwise orthology relations based on **gene neighborhood conservation**. Analyses on simulated and real data reveal the high accuracy of OrthoGNC. In addition to orthology detection, OrthoGNC can be employed to investigate the conservation of **genomic context** among potential orthologs detected by other methods. OrthoGNC is freely available online at http://bs.ipm.ir/softwares/orthognc and http://tinyurl.com/orthoGNC.

Introduction

Currently, sequencing facilities are able to produce large amounts of gene and protein sequences in a short period of time. Hence, many complete genomes of organisms are available today for more in-depth comparative studies. A first step in comparative genomics is the identification of homologous and more specifically orthologous genes. Homologous genes (homologs) are originated from a gene in the last common ancestor. In 1970, Fitch classified homologs into orthologous and paralogous genes [1]. Orthologous genes (orthologs) are homologs that have evolved by speciation event in their last common ancestor. In contrast, paralogous genes (paralogs) are homologs that have evolved by gene duplication in their last common ancestor.

Identification of orthologs is more important and of great interest, since orthologs typically tend to share a similar function [2]. Thus orthology relations can be used to transfer functional annotations (including protein–protein interactions) to newly-sequenced genomes [3,4]. Moreover, by definition, only

* Corresponding author.
E-mail: ch-eslahchi@sbu.ac.ir (Eslahchi C).
[a] ORCID: 0000-0002-3772-792X.
[b] ORCID: 0000-0003-1241-5441.
[c] ORCID: 0000-0002-8913-3904.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2017.07.002

phylogeny of orthologs can reflect the true evolutionary history of the corresponding species correctly [5]. Therefore, only orthologs can be used to infer species phylogenies [6].

Despite the straightforward definition of orthology, the problem of assigning orthology is not trivial. Evolutionary events such as horizontal gene transfer (HGT) and gene loss often complicate the evolutionary history of genes. Hence, many methods and databases have been introduced to tackle the problem of orthology assignment. More than 40 methods and databases are listed in the "Quest for Orthologs" website (http://questfororthologs.org/orthology_databases), which can mostly be classified into two major classes according to the approaches employed.

Methods such as OrthoStrapper [7], HOGENOM [8], LOFT [9], PhylomeDB [10], and OrthoReD [11] are based on phylogenetic analysis. Phylogeny-based methods seem to be more precise, with high specificity reported [7,12,13]. However, ambiguities in the inferred gene trees and species trees, as well as wrong placement of the root can lead to incorrect assignment of orthologs. In addition, these methods require large computational cost, making their usage impractical for large datasets [14–16].

The second class of methods usually employs a clustering algorithm on a weighted graph that is built from pairwise sequence similarities. Examples in this class include OMA [17], OrthoMCL [18], InParanoid [19], Proteinortho [20], and OrthoDB [21]. These methods, because of their tractability, have gained more popularity, particularly when used for large datasets. However, these methods are based on the molecular clock hypothesis by assuming that orthologous sequences are more similar and would fail to detect orthologs when the molecular clock hypothesis is violated [22]. Furthermore, HGT and convergent evolution as well as linage-specific gene loss (**Figure 1**) can introduce false positive relations. Note that

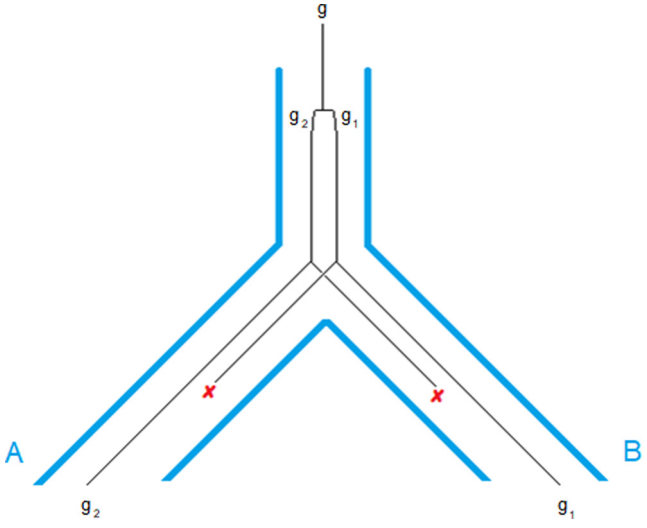

Figure 1 An example for lineage-specific gene loss
Suppose gene g has undergone a duplication event, resulting in two genes, g_1 and g_2, which is followed by a speciation event. Deletion of g_1 in lineage A and deletion of g_2 in lineage B can lead to the wrong assignment of g_2 from lineage A and g_1 from lineage B as orthologs. According to the duplication event in the last common ancestor, g_1 and g_2 are paralogs. Since duplication event occurs prior to speciation, g_1 and g_2 are out-paralogs.

duplications prior to a speciation event (in-paralogs) and duplications after a speciation event (out-paralogs) [23] introduce one-to-many or many-to-many orthology relation, further complicating the process of orthology detection. Most similarity-based methods that employ clustering, present the orthology relations as ortholog groups instead of pairwise relations. As a result, these groups can contain in-paralogs and out-paralogs, making their usage inappropriate for studies such as phylogeny inference where one-to-one orthology relations are needed.

To increase the quality of inferred orthologs, some methods attempt to take genomic context into account [24–29]. Due to genome rearrangements, as well as gene gains/losses, genomic context is less conserved beyond the genus level. Nonetheless, it can be a strong evidence of orthology when found. Conservation of gene neighborhood can assist in distinguishing orthologs from out-paralogs [30,31]. Similarly, it can prevent misinterpretation of HGTs and genes with convergent evolution as orthologs. Another interesting advantage of employing genomic context is to reveal orthologous genes or proteins that share low sequence similarities [32]. These orthologs can be missed in clustering or homology inference steps due to trade-off in favor of specificity. Moreover, with the advent of next-generation sequencing, many closely related genomes are available today. As a result, comparative studies are extended to closely related species and even strains of a same species where genomic context is highly conserved.

In this paper, we extend the method of Jun et al. [29] in both homology detection and orthology detection to increase the accuracy of predictions. We propose OrthoGNC, a similarity-based method and a software that outputs high-quality pairwise orthology relations based on gene neighborhood conservation. Moreover, OrthoGNC can be employed to investigate the conservation of genomic context among potential orthologs detected by other methods.

Method

Jun et al. tested a simple method based on gene neighborhood conservation to extract orthology relations in mammalian proteomes [29]. According to their method, two genes are orthologous if they are homologous and share at least one homologous neighbor in a neighborhood size of three upstream and three downstream genes. Also, homology between two genes are defined as Blastp E-value < 1e−5. We have extended this method in both homology detection and orthology detection. Similar to the aforementioned method, OrthoGNC performs three main steps to infer orthology relations: (i) computing pairwise sequence similarities, (ii) identifying homologous sequences, and (iii) inferring orthologs according to gene neighborhood conservation. However, in each step, parameters can be set to various values to provide desired output. These parameters can be easily adjusted by the user in a configuration file or using a user-friendly GUI. Moreover, in OrthoGNC, inference of orthology relation can be done in an iterative routine to produce more accurate and sensitive results.

OrthoGNC steps

OrthoGNC is implemented in Java and accepts both DNA and protein sequences in FASTA format. In addition, the order of

Please cite this article in press as: Jahangiri-Tazehkand S et al, OrthoGNC: A Software for Accurate Identification of Orthologs Based on Gene Neighborhood Conservation. Genomics Proteomics Bioinformatics (2017), https://doi.org/10.1016/j.gpb.2017.07.002