BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 8 2005, pages 1713–1714
doi:10.1093/bioinformatics/bti208

*Phylogenetics*

# GeneContent: software for whole-genome phylogenetic analysis

Xun Gu[1,2,*], Wei Huang[3], Dongping Xu[1] and Hongmei Zhang[4]

[1]Department of Genetics, Development and Cell Biology, [2]Center for Bioinformatics and Biological Statistics and [3]Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011, USA and [4]Department of Mathematics and Statistics, University of West Florida, Pensacola, FL 32514, USA

Received on September 3, 2004; revised on November 30, 2004; accepted on December 2, 2004
Advance Access publication December 14, 2004

**ABSTRACT**

**Summary:** GeneContent is a software system to infer the genome phylogeny based on an additive genome distance that can be estimated from the extended gene content data, which contains the genome-wide information (absence of a gene family, presence as single copy or presence as duplicates) across multiple species. GeneContent can also be used to explore the genome-wide evolutionary pattern of gene loss and proliferation.

**Availability:** Distribution packages of GeneContent for both Microsoft Windows and Linux operating systems are available at http://xgu.zool.iastate.edu

**Contact:** xgu@iastate.edu

Since phylogenetic trees inferred from individual genes may be inconsistent, the whole-genome approach, such as the gene content, becomes an attractive approach to extract bulk phylogenetic signals. For instance, some authors (e.g. Snel *et al*., 1999; Huynen *et al*., 1999; Lin and Gerstein, 2000; Korbel *et al*., 2002) estimated the fraction of shared genes for genome pairs, and transformed it to the genome distance matrix by some *ad hoc* distance measures. Other methods include the coefficient of co-occurrence of genomics (Natale *et al*., 2000) and the ratio of orthologs to the number of genes in the smaller genome (Clarke *et al*., 2002). In addition, various parsimony algorithms have also been used (e.g. Fitz-Gibbon and House, 1999; House and Fitz-Gibbon, 2002).

However, the statistical model of genome evolution should be addressed appropriately for having a reliable phylogenetic inference rather than the best phenotypical clustering. To this end, Gu and Zhang (2004) proposed a statistical framework for the phylogenetic gene-content analysis, which has been successfully applied for the tree of life. We have subsequently developed a user-friendly GUI-based software system, GeneContent, to facilitate the further study in comparative genomics.

The software GeneContent deals with two types of gene-content data: the conventional gene content (Snel *et al*., 1999; Huynen *et al*., 1999; Lin and Gerstein, 2000; Korbel *et al*., 2002) contains the genome-wide information for the presence/absence of gene families across multiple species, while the extended gene content (Gu and Zhang, 2004) contains the genome-wide information as follows: absence of a gene family, presence as single copy or presence as duplicates. The advantage of extended gene content for phylogenomics is demonstrated below.

*To whom correspondence should be addressed.

Based on the birth–death stochastic model (Gu and Zhang, 2004), an additive genome distance measure between two species can be defined as $G = 2(\lambda + \mu)t$, where $\lambda$ is the proliferation (duplicate) rate of a gene family, $\mu$ is the loss rate of genes and $t$ is the evolutionary time units. It has been shown that for two genomes, it is difficult to utilize the conventional gene-content data to estimate the genome distance $G$, except for the special case, where $\lambda = 0$. Gu and Zhang (2004) have solved this problem by introducing the concept of extended gene content, and proposed an efficient algorithm for genome-wide phylogenetic analysis since it does not require much computational time.

The interface of the software GeneContent (Fig. 1) is straightforward and easy to use. The input of the data is in the text file, in which the rows correspond to different genomes and the columns to gene families. The values for each entry of the data matrix could represent the size of gene family in the genome, gene content or extended gene content. Our program will trim the input matrix to fit the type of input as specified by the user. GeneContent provides three options to calculate genome distance: the Poisson distance, the gene content (under the special case where $\lambda = 0$) and the extended gene content. By default, both gene content and extended gene content methods will be provided, except that the input matrix only contains two types of values (i.e. 0 for absence and 1 for presence); in this case, the extended gene content method will be disabled. The Poisson distance is available for comparison purpose. Note that the gene-content distance between species ($A$ and $B$) is calculated $D_{AB} = 1 - J_{AB}$, where $J_{AB}$ is the Jaccard coefficient, which reflects the similarity of gene content between $A$ and $B$ (Wolf *et al*., 2002).

After obtaining the genome distance matrix, the software is able to infer the genome phylogeny using the neighbor-joining method (Saitou and Nei, 1987). The statistical reliability of the inferred genome phylogeny is examined by the conventional bootstrapping approach. Since the inferred phylogeny is un-rooted, the option for changing the root under the tree-view is available, as well as other options for visualization editing. The inferred genome tree can be saved as a text file in the Phylip format, which is useful in some cases.

The performance of the above algorithm has been examined by the universal genome tree of 36 complete genomes (Gu and Zhang, 2004). In the current version, we have implemented some options to explore the pattern of genome evolution. For instance, the proliferation/loss rate ratio can be mapped onto the phylogenetic tree, and the bootstrapping test can be performed to examine whether it remains a constant among lineages. We will upgrade our software

© The Author 2004. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

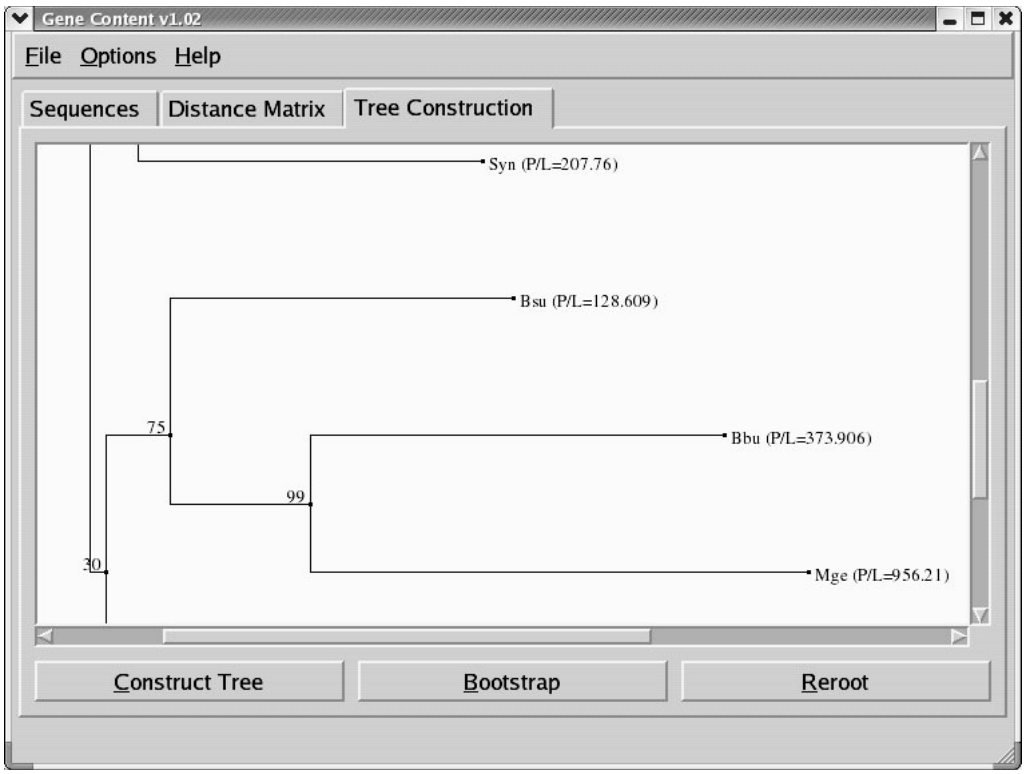


**Fig. 1.** The main interface of GeneContent includes three tabs: sequences, distance matrix and tree construction.

in two directions. The first one is to improve the evolutionary model by considering more factors such as lateral gene transfer and co-evolution among gene families. The second direction is to implement more sophisticated tree-making algorithms, e.g. a fast algorithm for the maximum-likelihood inference of genome phylogeny.

## ACKNOWLEDGEMENT

This work was supported by the NIH grant RO1 GM62118 to X.G.

## REFERENCES

Clarke,G.D.P., Beiko,R.G., Ragan,M.A. and Charlebois,R.L. (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J. Bacteriol.*, **184**, 2072–2080.

Fitz-Gibbon,S.T. and House,C.H. (1999) Whole genome–based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.*, **27**, 4218–4222.

Gu,X. and Zhang,H.M. (2004) Genome phylogenetic analysis based on extended gene contents. *Mol. Biol. Evol.*, **21**, 1401–1408.

House,C.H. and Fitz-Gibbon,S.T. (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J. Mol. Evol.*, **54**, 539–547.

Huynen,M.A., Snel,B. and Bork,P. (1999) Technical comments on Doolittle [1999a]. *Science*, **286**, 1443a.

Korbel,J.O., Snel,B., Huynene,M.A. and Bork,P. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet.* **18**, 158–162.

Lin,J. and Gerstein,M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implication for comparing genomes on different levels. *Genome Res.*, **10**, 808–818.

Natale,D.A., Shankavaram,U.T., Galperin,M.Y., Wolf,Y.I., Aravind,L. and Koonin,E.V. (2000) Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol.*, **1**, RESEARCH0009.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Bol. Evol.*, **4**, 406–425.

Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110.

Wolf,Y.I., Rogozin,I.B., Grishin,N.V. and Koonin,E.V. (2002) Genome trees and the tree of life. *Trends Genet.*, **18**, 472–479.