2011 – 02

Modeling And Comparing the Organization of Circular Genomes

Grace S. Shieh Shurong Zheng, Richard A. Johnson,
Yi-Feng Chang, Kunio Shimizu, Chia-Chang Wang,
Sen-Lin Tang

Jan. 26, 2011

BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–7

Modeling And Comparing the Organization of Circular Genomes

Shurong Zheng [1,2], Richard A. Johnson [3], Yi-Feng Chang [4], Kunio Shimizu [5], Chia-Chang Wang [1], Sen-Lin Tang [6] and Grace S. Shieh [1*]

[1]Institute of Statistical Science and [6]Biodiversity Research Center, Academia Sinica, Taipei 115, TAIWAN, R.O.C., [2]Present address: KLAS and Mathematics & Statistics, Northeastern Normal University, Changchun city, China, [3]Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A., [4]Institute of Biomedical Informatics, National Yang-Ming University, Taipei 112, TAIWAN, R.O.C., [5]Keio University, Tokyo 53706, Japan.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Most prokaryotic genomes are circular with a single chromosome (called circular genomes), which consist of bacteria and archaea. Orthologous genes (abbreviated as orthologs) are genes directly evolved from an ancestor gene, and can be traced through different species in evolution. Shared orthologs between bacterial genomes have been used to measure their genome evolution (Huynen and Bork, 1998). Here, organization of circular genomes is analyzed via distributions of shared orthologs between genomes. However, these distributions are often asymmetric and bimodal; to date there is no joint distribution to model such data. This motivated us to develop a family of bivariate distributions with generalized von Mises marginals (BGVM) and its statistical inference.

Results: A new measure based on circular grade correlation and percentage of shared orthologs is proposed for association between circular genomes, and a visualization tool developed to depict genome structure similarity. The proposed procedures are applied to eight pairs of prokaryotes separated from domain down to species, and 13 mycoplasma bacteria which are mammalian pathogens belonging to the same genus. We close with remarks on further applications to many features of genomic organization, e.g. shared transcription factor binding sites, between any pair of circular genomes. Thus the proposed procedures may be applied to identifying conserved chromosome backbones, among others, for genome construction in synthetic biology.

Availability: All codes of the BGVM procedures are available at http://www.stat.sinica.edu.tw/~gshieh/bgvm.htm.

Contact: gshieh@stat.sinica.edu.tw

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Most of the prokaryotic genomes (1158 out of 1194, NCBI, Aug. 2010) are circular with a single chromosome (called circular genomes henceforth). Orthologous genes (abbreviated as orthologs) are genes directly evolved from an ancestor gene (Tatusov, Koonin and Lipman, 1997), and can be traced through different species in evolution. The fraction of shared orthologs between two circular genomes was found to be more conserved than the order of genes (Huynen and Bork, 1998), in which the fraction of shared orthologs between genomes was employed to measure genome evolution of nine prokaryotes. Here, our emphasis is on the structure of circular genomes, which, for example, plays an important role in synthetic biology. A review paper in synthetic genomics (Carrera et al., 2009) indicates that genome organization may influence gene expression, which is vital for organisms. Further, predicting or modeling the rules of genome organization via comparative genomics may provide valuable information for genome construction.

We reason that genome structure can be studied via distributions of shared orthologs between genomes, e.g. the most or least favored region in which shared orthologs between each pair of bacterial genomes are located. While distributions of shared orthologs are often found to be asymmetric and bimodal, to date there is no joint distribution with closed-formed marginals to model such data. This motivated us to develop a family of joint distribution and its related statistical inferences.

Recent studies show that gene order is extensively conserved between closely related species, but rapidly become less conserved among more distantly related species. This trend is likely to be universal in prokaryotes (Tamames, 2001; Wolf *et al.*, 2001). However, the fraction of shared orthologs between two circular genomes is more conserved than the order of genes (see Figure 6 of Huynen and Bork, 1998). In addition to the ratio of shared orthologs between bacterial genomes, we further incorporate the distributions of shared orthologs of paired circular genomes to infer similarity of their genome organization. By converting the locations of shared orthologs in any paired circular genomes into angles, these pairs of angles can be viewed as bivariate circular vectors.

Most of the marginal distributions of shared orthologs in circular genomes are asymmetric and/or multi-modal, which can be modeled by the generalized von Mises distribution (GVM) (Maksimov, 1967; Yfantis and Borgman, 1982). Therefore, we propose a bivariate circular distribution with each marginal assuming a GVM distribution, and call this distribution the bivariate generalized

*To whom correspondence should be addressed.