BIOINFORMATICS APPLICATIONS NOTE

Vol. 29 no. 7 2013, pages 947–949
doi:10.1093/bioinformatics/btt064

Genome analysis

Advance Access publication February 8, 2013

Phylogenomic clustering for selecting non-redundant genomes for comparative genomics

Gabriel Moreno-Hagelsieb[1,*], Zilin Wang[2], Stephanie Walsh[2] and Aisha ElSherbiny[1]

[1]Department of Biology and [2]Department of Mathematics, Wilfrid Laurier University, Waterloo, ON N2L 3C5, Canada

Associate Editor: David Posada

ABSTRACT

Motivation: Analyses in comparative genomics often require non-redundant genome datasets. Eliminating redundancy is not as simple as keeping one strain for each named species because genomes might be redundant at a higher taxonomic level than that of species for some analyses; some strains with different species names can be as similar as most strains sharing a species name, whereas some strains sharing a species name can be so different that they should be put into different groups; and some genomes lack a species name.

Results: We have implemented a method and Web server that clusters a genome dataset into groups of redundant genomes at different thresholds based on a few phylogenomic distance measures.

Availability: The Web interface, similarity and distance data and R-scripts can be accessed at http://microbiome.wlu.ca/research/redundancy/.

Contact: gmoreno@wlu.ca

Received on October 1, 2012; revised on January 31, 2013; accepted on February 5, 2013

1 INTRODUCTION

Genomic databases contain several sequenced strains for species of high interest, whereas other species contain only one genome and no evolutionarily close relatives (Wu *et al.*, 2009). These biases might result in different qualities when predicting such features as functional interactions by comparative genomics. The problem has often been solved by using a measure of phylogenetic distance either directly as a score (e.g. Overbeek *et al.*, 1999) or for normalizing scores (e.g. Zheng *et al.*, 2005). The problem has also been solved by choosing single representatives for particular clades (e.g. Moreno-Hagelsieb and Janga, 2008). However, selecting representatives by species name can be problematic because strains of different species, for example *Escherichia coli* and *Shigella flexneri*, can be so evolutionarily close that they should share their names, whereas strains of single species such as *Prochlorococcus marinus* might be very distant to each other (Moreno-Hagelsieb and Janga, 2008; van Passel *et al.*, 2006). Besides, some sequenced genomes do not have a species name, and the level of acceptable similarity among genomes for some applications might be lower than the most common for species (Moreno-Hagelsieb and Janga, 2008).

The Genomic Similarity Score (*GSS*) and other phylogenomic measures have been shown to concur with more computationally demanding phylogenetic distances (Alcaraz *et al.*, 2010; Kunin *et al.*, 2005). We have used *GSS* to produce datasets of non-redundant genomes, maximizing the number of predicted interactions by phylogenetic profiles (Moreno-Hagelsieb and Janga, 2008), and other practical uses, such as displaying the relationship between conservation of gene order and genomic similarity (Janga and Moreno-Hagelsieb, 2004). The datasets were built by ordering the genomes available by importance (e.g. *Escherichia coli* K12 would take priority because of its annotation quality), and then from largest to smallest. We would then take the first genome in the list, and put any genome with a *GSS* above a chosen threshold into the list of redundant genomes. In the present work, we tested instead two clustering methods to group genomes, besides a couple variants of the *GSS* and DNA signatures (Campbell *et al.*, 1999). As other research groups have asked us for lists of non-redundant genomes, here we make them available to the wider community.

2 DATA AND METHODS

We used 2160 prokaryotic genomes available at the RefSeq database (Pruitt *et al.*, 2007) (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) by November 2012. Orthologs were reciprocal best hits (RBH), as described previously (Moreno-Hagelsieb and Latimer, 2008). We plan to update the data every 6 months.

We calculated the *GSS*s for each genome pair as follows (Janga and Moreno-Hagelsieb, 2004; Moreno-Hagelsieb and Janga, 2008):

$$GSS_a = \sum_{i=1}^{n} \frac{compScore_i}{selfScore_i} \tag{1}$$

where *compScore* is the score of the alignment of protein *i* in genome *A* against its RBH in genome *B*, and *selfScore* is the score of the alignment of protein *i* against itself. Because the *selfScore* for proteins in genome *A* might differ from the *selfScore* for proteins in genome *B*, the final GSS_a is the average of the calculation from the perspective of both genomes.

We also calculated two modified versions of the *GSS*:

$$GSS_b = \sum_{i=1}^{n} \frac{compScore_i/compAlignLength_i}{selfScore_i/selfAlignLength_i} \tag{2}$$

Here the scores were normalized against the length of the alignments. Because *selfScores* normally come from longer alignments than *compScores*, GSS_b is larger than GSS_a. The final GSS_b is the average of this calculation from the perspective of both genomes under analysis.

*To whom correspondence should be addressed.

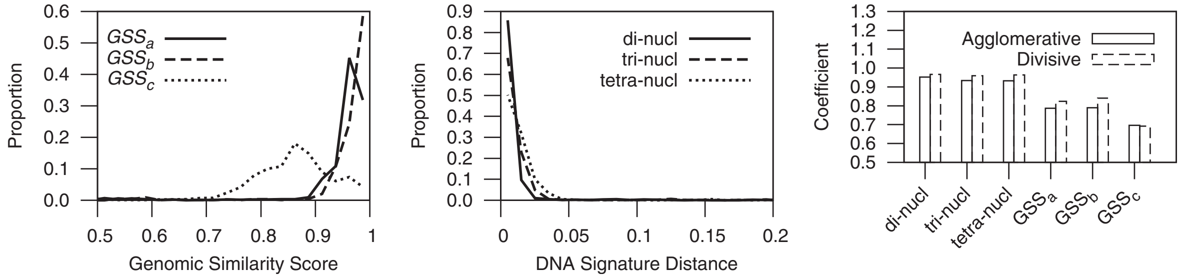

Fig. 1. Same-species clustering. Most of the same-species strains have *GSS*s >0.90 and DNA signature distances <0.25. GSS_b concentrates most of the same-species strains above the 0.9 score, and GSS_c spreads them the most. DNA signatures produce the best clusters, as shown by agglomerative/divisive coefficients, with divisive clustering producing slightly better clusters

$$GSS_c = \frac{\sum_{i=1}^{n} compScore_i}{\sum_{j=1}^{m} selfScore_j} \qquad (3)$$

Instead of the *selfScores* of proteins finding RBH, we calculated the total sum of *selfScores* of all proteins in the genome of reference. This has the effect of producing a smaller *GSS*. Again, the final GSS_c was the average from the perspective of both genomes under analysis.

We also clustered our genomes by di-, tri- and tetra-nucleotide DNA signatures with Manhattan distances (Campbell *et al.*, 1999).

To build groups of redundant genomes, we used a hierarchical clustering method (Hastie *et al.*, 2009). We used $1 - GSS$ to construct a dissimilarity matrix to run the clustering algorithms.

All computations were carried out using *R* (R Development Core Team, 2012). The commands *agnes* and *diana* in the *cluster* package were used to run the two main strategies in hierarchical clustering: agglomerative or bottoms-up and divisive or top-down, respectively.

To assess the clustering structures, we used the agglomerative/divisive coefficients defined as 1 minus the average ratio of dissimilarity of one unit to the cluster it first merges with, to the dissimilarity of its merging in the final step of the algorithm. A coefficient close to 0 means that the algorithm did not find a natural cluster structure (the data consist of a single cluster), whereas a coefficient close to 1 means that clear clusters have been identified.

3 RESULTS

GSS_a and GSS_b concentrate strains sharing species names (1103 strains in 224 species) into the 0.9–1.0 range of the *GSS* distribution, whereas GSS_c spreads several of them between the 0.7 and 0.9 range (Fig. 1). Thus, GSS_c might be the least adequate for clustering redundant genomes. Because DNA signatures are faster to calculate than *GSSs*, we are starting to experiment with DNA signatures for our work in comparative genomics. Thus, we have also clustered our genome dataset based on di-, tri- and tetra-nucleotide signature distances. DNA signature distances among same-species strains concentrate at <0.03.

The divisive coefficients were higher, and thus better, than the agglomerative coefficients using all three *GSS*s and DNA signature distances (Fig. 1). The coefficients using GSS_c were the lowest among all three *GSS*s for both algorithms, thus confirming that GSS_c is the least adequate for clustering genomes. However, DNA signature distances produced the best overall clusters, with divisive coefficients >0.95.

The appropriate threshold for defining groups of redundant genomes depends on the application (e.g. Moreno-Hagelsieb and Janga, 2008). Thus, we recommend experimenting with different thresholds. The Web site provides 10 thresholds per measure that should suffice for such experimentation. Should other users prefer to experiment at thresholds other than those provided through the Web site, we also provide the R scripts used to derive these groups, and the pertinent distance/similarity data. With some effort, inexperienced users would be able to modify these scripts to obtain groups at other thresholds. Finally, we also provide cluster files that can be read by software for drawing and manipulating phylogenetic trees (Newick format).

ACKNOWLEDGEMENTS

We thank Duncan Murdoch for help with R issues. The Shared Hierarchical Academic Research Computing Network (SHARCNET).

Funding: Work supported with WLU funds and Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Conflict of Interest: none declared.

REFERENCES

Alcaraz,L.D. *et al.* (2010) Understanding the evolutionary relationships and major traits of bacillus through comparative genomics. *BMC Genomics*, **11**, 332.

Campbell,A. *et al.* (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl Acad. Sci. USA*, **96**, 9184–9189.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer series in statistics, Springer, New York, NY.

Janga,S.C. and Moreno-Hagelsieb,G. (2004) Conservation of adjacency as evidence of paralogous operons. *Nucleic Acids Res.*, **32**, 5392–5397.

Kunin,V. *et al.* (2005) Measuring genome conservation across taxa: divided strains and united kingdoms. *Nucleic Acids Res.*, **33**, 616–621.

Moreno-Hagelsieb,G. and Janga,S.C. (2008) Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. *Proteins*, **70**, 344–352.

Moreno-Hagelsieb,G. and Latimer,K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, **24**, 319–324.

Overbeek,R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

R Development Core Team*R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

van Passel,M.W. *et al.* (2006) The reach of the genome signature in prokaryotes. *BMC Evol. Biol.*, **6**, 84.

Wu,D. *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, **462**, 1056–1060.

Zheng,Y. *et al.* (2005) Phylogenetic detection of conserved gene clusters in microbial genomes. *BMC Bioinformatics*, **6**, 243.