BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 6 2008, pages 866–867
doi:10.1093/bioinformatics/btn031

Sequence analysis

TaxonGap: a visualization tool for intra- and inter-species variation among individual biomarkers

B. Slabbinck[1,*], P. Dawyndt[2], M. Martens[3], P. De Vos[3] and B. De Baets[1]

[1]KERMIT, Research Unit Knowledge-based Systems, Ghent University, Coupure links 653, [2]Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9) and [3]Laboratory of Microbiology, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium

Received on November 6, 2007; revised on January 4, 2008; accepted on January 21, 2008

Advance Access publication January 28, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: Selection of optimal biomarkers for the identification of different operational taxonomic units (OTUs) may be a hard and tedious task, especially when phylogenetic trees for multiple genes need to be compared. With TaxonGap we present a novel and easy-to-handle software tool that allows visual comparison of the discriminative power of multiple biomarkers for a set of OTUs. The compact graphical output allows for easy comparison and selection of individual biomarkers.

Availability: Graphical User Interface; Executable JAVA archive file, source code, supplementary information and sample files can be downloaded from the website: http://www.kermit.ugent.be/taxongap

Contact: Bram.Slabbinck@UGent.be

1 INTRODUCTION

When evaluating multiple genes as candidate biomarkers for the identification of different OTUs, one intuitively is looking for molecular markers that at the same time show the least amount of heterogeneity within OTUs and result in a maximal separation between the different OTUs. The first requirement must guarantee that members of the same OTU have the same (or at least similar) biomarkers, so that they can easily be grouped together based on those markers. The second requirement must establish that members of different OTUs have sufficiently different biomarkers, so that an identification based on those markers cannot erroneously suggest assignment of the members to the same OTU. TaxonGap was especially designed to produce a compact graphical representation of the resolution of individual biomarkers within and between taxonomic units, allowing easy and reliable inspection of the data for evaluations across different OTUs and different biomarkers.

2 IMPLEMENTATION

For a given set of OTUs $\{O_1, O_2, \ldots, O_n\}$, the *s*-heterogeneity within taxon O_i is defined by

$$\max_{x,\, y \in O_i,\, x \neq y} d_s(x, y)\,. \qquad (1)$$

Herein, $d_s(x, y)$ represents the distance between the (different) members x and y of the taxon O_i as measured from the biomarker s. These distances are presented in a separate distance matrix for each biomarker. Likewise, the *s*-separability of taxon O_i is defined by

$$\min_{x \in O_i,\, y \notin O_i} d_s(x,\, y). \qquad (2)$$

The taxon containing that member y for which a minimum distance is reached in the computation of the *s*-separability, is called the closest neighbour of taxon O_i. Note, however, that the closest neighbour relationship is not necessarily symmetric: the fact that O_j is the closest neighbour of O_i does not imply that O_i is also the closest neighbour of O_j.

TaxonGap calculates the matrix of *s*-heterogeneity and *s*-separability values with the different OTUs as matrix rows and the different biomarkers as matrix columns. Headers are placed to the left and on top of the matrix. To improve interpretability of the resulting graphical representation, the OTUs are presented according to their position in a phylogenetic tree, as an alternative to listing them in alphabetic order. With the aim to improve visual inspection and interpretation of the data and to support optimal comparability of the values across the biomarkers, TaxonGap presents the *s*-heterogeneity and the *s*-separability values respectively as light gray and dark gray horizontal bars for the individual biomarkers. The name of the closest neighbour is attached to the right side of the dark gray bar. Light gray bars are printed on top of and are made less thick than dark gray bars. Although not a strict requirement, it is advised that the same OTUs are used for evaluation of the different biomarkers. Missing biomarker data for a given OTU leads to holes in the TaxonGap output matrix (see gene *thrC* for species *Sinorhizobium americanum* in Fig. 1). Also note that there is no necessity to use the same OTU members for evaluating different biomarkers.

TaxonGap is implemented as an executable JAVA archive (JAR) file. The graphical output is formatted as an enhanced postscript (EPS) document. Sequence alignments, pairwise distance matrices and phylogenetic trees can be generated using third-party software packages. To make TaxonGap even more user-friendly, the command line interface allows to use TaxonGap as a backend plugin into this graphical software.

*To whom correspondence should be addressed.

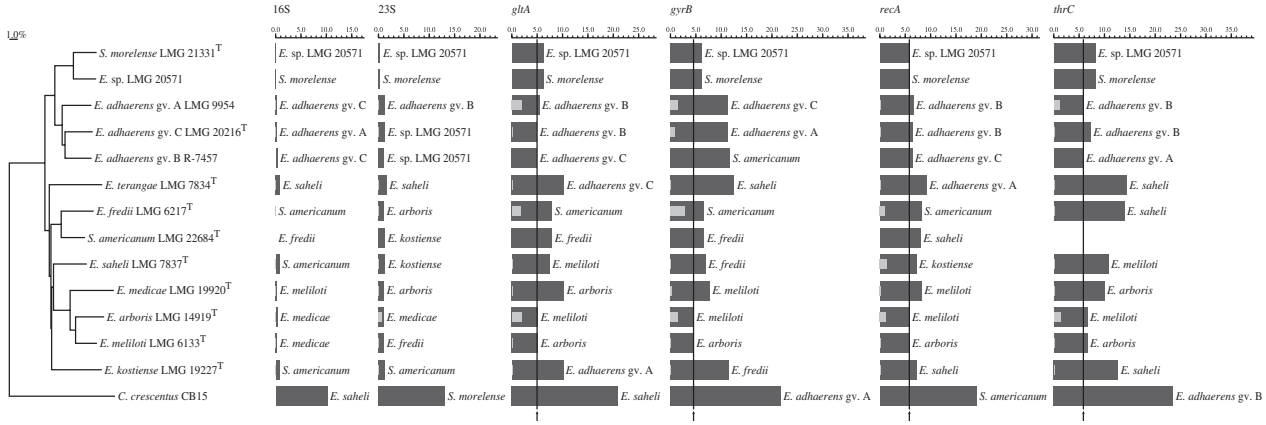

Fig. 1. Matrix of heterogeneity (presented as light gray horizontal bar) and separability (presented as dark grey horizontal bar) values with different OTUs as matrix rows and different biomarker genes as matrix columns. The species/genomovars of the genus Ensifer were ordered according to their phylogenetic position in a maximum likelihood tree calculated from the concatenated sequences of their type strains (when available) or another representative strain (see tree on the left). For each OTU and each biomarker, the closest neighbour (i.e. the taxon with the smallest separability w.r.t. this OTU and this biomarker) is listed at the right side of the dark grey bar. For each biomarker, the vertical black line denotes the smallest separability recorded.

3 DISCUSSION

Figure 1 shows a TaxonGap output example for the multilocus sequence identification schema of the genus *Ensifer* introduced by Martens and co-workers (Martens *et al.*, 2008). The genus *Ensifer*, comprising the former *Sinorhizobium* species and *E. adhaerens*, belongs to the Alphaproteobacteria. The study evaluated 10 housekeeping genes (*atpD*, *dnaK*, *gap*, *glnA*, *gltA*, *gyrB*, *pnp*, *recA*, *rpoB* and *thrC*) on 34 representatives of the genus *Ensifer*. Sequence data were analyzed, and the heterogeneity and separability values were calculated using TaxonGap for all housekeeping genes and the 16S and 23S rRNA genes. Clearer species boundaries and higher discrimination were observed for all housekeeping genes compared to rRNA genes. Housekeeping genes with the best identification of *Ensifer* strains, due to high separability and low heterogeneity values, were *gyrB*, *gltA*, *recA* and *thrC*.

Besides this study, TaxonGap was also successfully applied for evaluation of the genes *pheS* and *rpoA* for species identification of the genus *Lactobacillus* (Naser *et al.*, 2007).

It should be noted that the application of TaxonGap is not restricted to comparison of genetic or molecular markers. In fact, TaxonGap accepts pairwise distance matrices generated from any kind of biomarkers. All biomarkers that enable the calculation of pairwise distance matrices may be compared using TaxonGap.

4 CONCLUSIONS

The graphical representation produced by TaxonGap offers a number of advantages over comparing individual phylogenetic trees for evaluation of different biomarkers in polygenomic identification studies. First of all, a separate row is reserved in the TaxonGap output matrix for heterogeneity and separability values of different biomarkers for a single taxon, which is not the case when comparing phylogenetic trees. Even after a tedious process of swapping branches, it is not always possible to draw phylogenetic trees in a way that enables clear visual comparison. This is especially true when phylogenetic trees for multiple genes need to be compared. In addition, TaxonGap uses the same scaling for depicting distance values based on individual biomarkers. Few software tools for drawing phylogenetic trees allow precise control over scaling. Both placement and scaling improve comparability of the heterogeneity and separability for individual taxa. Secondly, we want to point out the fact that phylogenetic trees present approximations of the underlying distance values instead of using minimum and maximum as aggregation operators. This is important when comparing *s*-heterogeneity and *s*-separability for all species of a given biomarker *s*. To underscore the overall success rate of individual biomarkers to discriminate between the OTUs, TaxonGap depicts the overall separability (dark gray) per OTU as a vertical line for each biomarker. This line is omitted when the overall separability is too small. Finally, the graphical output of TaxonGap remains compact, even for data sets where the number of OTU members or biomarkers grows large. This is because the software has a built-in aggregation based on the individual OTUs and biomarkers.

TaxonGap thus allows for a more straightforward evaluation of the discriminatory power of individual biomarkers in an OTU identification scheme, as opposed to the need of comparing separate gene trees drawn for each of the OTUs in the scheme.

ACKNOWLEDGEMENT

This work is funded by the Belgian Science Policy (C3/00/12).

Conflict of Interest: none declared.

REFERENCES

Martens,M. *et al.* (2008) Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus *Ensifer* (including former *Sinorhizobium*). *Int. J. Syst. Evol. Micr.*, **58**, 200–214.

Naser,M.S. *et al.* (2007) Identification of lactobacilli by *pheS* and *rpoA* gene sequence analysis. *Int. J. Syst. Evol. Micr.*, **57**, 2777–2789.