Bioinformatics, 35(18), 2019, 3482–3483
doi: 10.1093/bioinformatics/btz099
Advance Access Publication Date: 14 February 2019
Applications Note

Genome analysis

Bacterial Feature Finder (BaFF)—a system for extracting features overrepresented in sets of prokaryotic organisms

Javier López-Ibáñez, Laura T. Martín, Mónica Chagoyen and Florencio Pazos*

Computational Systems Biology Group, Systems Biology Program, Spanish National Centre for Biotechnology (CNB-CSIC), Madrid 28049, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 17, 2018; revised on December 21, 2018; editorial decision on February 7, 2019; accepted on February 12, 2019

Abstract

Motivation: The results of some experimental and computational techniques are given in terms of large sets of organisms, especially prokaryotic. While their distinctive features can provide useful data regarding specific phenomenon, there are no automated tools for extracting them.

Results: We present here the Bacterial Feature Finder web server, a tool to automatically interrogate sets of prokaryotic organisms provided by the user to evaluate their specific biological features. At the core of the system is a searchable database of qualitative and quantitative features compiled for more than 23 000 prokaryotic organisms. Both the input set of organisms and the background set used to calculate the enriched features can be directly provided by the user, or they can be obtained by searching the database. The results are presented via an interactive graphical interface, with links to external resources.

Availability and implementation: The web server is freely available at http://csbg.cnb.csic.es/BaFF. It has been tested in the main web browsers and does not require any especial plug-ins or additional software.

Contact: pazos@cnb.csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the current scenario of massive generation of biological data, it is increasingly common that the results of some experimental and computational techniques are given in terms of large sets of organisms, especially prokaryotic. For example, a metagenomics experiment can deliver a long list of bacteria detected in a given sample (e.g. gut, sea water, etc.), perhaps identified from DNA fragments (Segata *et al.*, 2012). Likewise, the phylogenetic profile of a gene/protein family offers information regarding the set of bacterial species in which that family is present, and also, on the complementary set, those in which it is absent (Pellegrini *et al.*, 1999). The differential/distinctive features of organisms within such datasets can provide useful information on the biological phenomenon under study. In the first of these examples, the genomic/biological features of the microorganisms detected in that sample would be expected to reflect the sample's characteristics (environmental conditions, nutrient availability, etc.), some of which may be unknown. Similarly, in the second example, some characteristic features of the organisms expressing a given gene/system will be related to its biological role and, hence, they could be used to infer that role (e.g. for open reading frame (ORFs) with no known function). By analyzing the distinctive features of a set of organisms it would be possible to detect, e.g. whether such set is particularly enriched in *gram+* or pathogenic organisms, or if they tend to have larger genomes or more genes involved in 'amino acid metabolism' than the average (background).

Since no single tool exists to automatically extract differential features from a generic set of microorganisms, such studies have generally been carried out more or less manually. That is also in part due to the fact that while there is much information available on different microorganisms, it is widely dispersed across different databases and, hence, it is difficult to mine it in a common framework.

Inspired by the widely used approaches to detect the functional annotations enriched in a particular set of biomolecules, e.g. genes/proteins (Huang *et al.*, 2009) or metabolites (Barupal *et al.*, 2018), we applied the same idea to analyze sets of microorganisms. These approaches, globally known as enrichment or over-representation analysis, allow long lists of biomolecules (e.g. those overexpressed in a particular experiment) to be interpreted in biological terms, transforming such lists into a smaller set of meaningful biological keywords. This is done by extracting the annotations that distinguish these molecules from a background (e.g. the whole genome of the organism under study).

We have implemented this kind of analysis for sets of microorganisms in the 'Bacterial Feature Finder' (BaFF) web server. Using a large database of bacterial features (qualitative and quantitative), the system can locate those differentially associated to an input set of organisms relative to a background set, both provided by the user or resulting from a database search. The results are presented through an interactive graphical interface and they can be exported for further processing.

2 Database and analysis

The core database of the system contains information regarding 14 features of prokaryotic organisms, extracted from different online public resources. These include both, qualitative and quantitative characteristics, such as gram staining, pathogenicity, associated diseases, number of genes or the GC content. A detailed list of these features and the resources from where they were obtained is provided in Supplementary Material 1. At present, the latest version of the database contains 23 809 prokaryotic organisms (bacteria and archaea) for which information on at least one of these features is available.

Two different statistical tests are applied to detect the features enriched in a set of organisms, depending on whether they are qualitative or quantitative. The probability of obtaining a given qualitative feature by chance is extracted from the cumulative hypergeometric density function. By contrast, quantitative features are evaluated using a two-sided Kolmogorov–Smirnov test, such that deviations toward low and high values can be detected. More details on these tests are given in Supplementary Material 1. In both cases, a *P*-value is calculated for every feature.

3 Interface

The main interface of the system allows searches of the database to be performed, as well as an enrichment analysis (see Supplementary Material 1 for screenshots of the system). The user can construct searches based on the main features annotated in the database, and the results are lists of microorganisms matching those criteria, where the main identifiers (NCBI TAXIDs) are links to the complete database records of the organisms. The data within those records are hyper-linked to external resources, generally to those from where they were originally obtained. The resulting list of organisms can be downloaded for further processing.

The enrichment analysis requires an input set of organisms and a background set. Both can be provided by the user by uploading files in which organisms are identified by NCBI TAXIDs or Uniprot organism mnemonics, or they can be the result of a previous database search. All combinations are possible and if no background is provided, the whole database is used.

The result of such enrichment analysis is a list of features, grouped into categories and sorted by the *P*-value of the corresponding test (see above). The *P*-values are color-coded based on their significance and, in the case of qualitative features, arrows next to the *P*-value indicate whether the deviation is toward lower or higher values. The list can be sorted by any column and exported to a text file. Moreover, checkboxes are provided to show/hide particular categories of features, so that the list is easier to interpret.

The server also includes a Help/Tutorial and FAQ sections with detailed information on how to use the tool. The Tutorial includes example datasets and explanations on the results obtained for them. The examples cover different scenarios typical of an enrichment analysis, such as the use of input and background sets provided by the user, the use of the search results as input/background for the analysis, etc.

4 Conclusions and future prospects

We have developed a system for interpreting, in biological terms, large repertories of prokaryotic organisms generated by experimental and *in-silico* methodologies. To our knowledge, no similar systems can achieve such goals and hence, BaFF complements the tools currently available to microbiologists and bioinformaticians. Any quantitative or qualitative data available for a sufficiently large number of organisms can be readily incorporated into the database and used for the enrichment analysis. Hence, we plan to expand the system in the future with more of these features. The quality of the bacterial annotations is critical for the performance of the system. While manually curating them is not feasible, the system is open so that the core database can be updated with newer releases of the original resources, and other additional resources can be easily incorporated.

Acknowledgements

We want to thank the members of the Systems Biology Program (CNB-CSIC) for our interesting discussions, especially Juan Nogales and Javier Tamames.

Funding

This work was partially supported by project SAF2016–78041-C2–2-R from the Spanish Ministry for Economy and Competitiveness. J.L.I. is the recipient of a contract from the Spanish Ministry of Science, Innovation and Universities and the European Social Fund [BES-2015–073281].

Conflict of Interest: none declared.

References

Barupal,D.K. *et al.* (2018) Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. *Curr. Opin. Biotechnol.*, **54**, 1–9.

Huang,D.W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein pylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **96**, 4285–4288.

Segata,N. *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811–814.