BIOINFORMATICS LETTER TO THE EDITOR
Vol. 24 no. 18 2008, pages 2124–2125
doi:10.1093/bioinformatics/btn355

Genome analysis

Metagenomics reveals our incomplete knowledge of global diversity

Miguel Pignatelli[1,2], Gabriel Aparicio[3], Ignacio Blanquer[3], Vicente Hernández[3], Andrés Moya[1,2] and Javier Tamames[1,2,]*
[1]Instituto Cavanilles of Biodiversity and Evolutionary Biology, University of Valencia, Apdo 22085, 46071 Valencia, [2]CIBER of Epidemiology and Public Health (CIBERESP) and [3]Grid and High-performance Computing Group, ITACA, Politechnic University of Valencia, Camino de Vera s/n, 46022 Valencia, Spain

Received on May 23, 2008; revised and accepted on July 10, 2008
Advance Access publication July 13, 2008
Associate Editor: Dmitrij Frishman

Contact: Javier.tamames@uv.es

Metagenomic sequencing obtains huge amounts of sequences from environmental and clinical samples, thus providing a glimpse of the global prokaryotic diversity of both species and genes in these sources. The current trend in metagenomic analysis follows the so-called gene-centric approach, focused on describing the environments by the study of the functional roles of the proteins encoded in the sequenced genes. In this way, it is clear that metagenomic analysis relies heavily on the accurate knowledge of the universe of proteins stored in the databases. Nevertheless, it is known that some biases exist in the composition of databases (which are rich in sequences from common, cultivable and easily accessible organisms), but it is uncertain how big this bias is and how it can influence the analysis of real data, that is, how accurately the databases are describing the global diversity.

In addition to functional assignment of proteins, database completion can also influence greatly the taxonomic classification of metagenomic sequences (binning). Having accurate taxonomic assignments would be essential, since it would help greatly in our understanding of the community dynamics, to predict the effect of changes in their composition, and to study key issues in the evolution of the community, such as the extent of horizontal gene transfer (HGT) or the barriers shaping the species. The analysis of metagenomic sequences without taxonomic assignment will always provide a superficial and incomplete view. But binning is difficult for metagenomic sequences, since they are usually very short and lack enough information to be classified by compositional features and/or phylogenetic analysis (Tamames and Moya, 2008). Although some bioinformatics tools have been proposed for binning, they provide good results only for a reduced fraction of the sequences (Krause *et al*., 2008; Mavromatis *et al*., 2007). To deal with the problem, several authors have relied on the assignment of ORFs to the taxon of their closest relatives in a homology search (Tringe *et al*., 2008; Venter *et al*., 2004). This is a potentially conflicting strategy, which may often fail for poorly known taxa (as it is often the case for metagenomic samples), and can be easily confounded by HGT events (Koski and Golding, 2001).

We have performed a simple experiment to explore: (1) our current knowledge of the universe of sequences, and how this knowledge has evolved in the past years, and (2) the possible extent of failures when taxonomically assigning ORFs to their closest relatives. Using the Blast in Grids (BiG) service (Aparicio *et al*., 2007), we have run BLASTX searches for several metagenomes against the realease 159 (April 2007) of GenBank non-redundant protein database, extracted the homologues found for each putative ORF [following the procedure described in Tamames and Moya (2008)] and assigned the ORF to the taxon of its best hit (where different threshold of minimum identity have been used). Next, we have collected the dates of creation of the GenBank entries of these hits. In this way, we can simulate the results that we would have obtained in the past, by restricting the list of homologues to those already present in the database in a particular date, which allowed us to explore how the results change as the database grows. The experiment was repeated for different metagenomes and different taxonomic depths. One of the results is shown in the Figure 1A, for a farm soil metagenome (Tringe *et al*., 2005). Each row corresponds to a single ORF, displaying the colour of the class to which it would have been assigned in different dates. The plot shows clearly that many of the assignments have changed, even in recent times (between 10% and 20% of the assignments at class level have varied in the last 2 years). This can be easily seen in Figure 1B, which shows the accumulated number of changes with respect to previous dates. Several trends are noticeable: (1) The rate of change does not decrease in recent times, instead it clearly increases for most cases. This trend of change can be noticed even for broad taxonomic ranks such as phylum: for instance, *Acidobacteria* phylum is now recognized as one of the most abundant taxa in many soils (Barns *et al*., 1999), but until recently no sequences were assigned to it. This indicates that the full diversity of these communities is still not well described in the current databases, and that best hit approach for taxonomic classification is at least risky. (2) Although most changes consist in the assignment of previously unclassified ORFs, the classification for many ORFs has also changed. (3) Abrupt changes occur in response to the availability of complete genomes, especially from close species to those represented in the metagenomes. Again for *Acidobacteria*,

*To whom correspondence should be addressed.

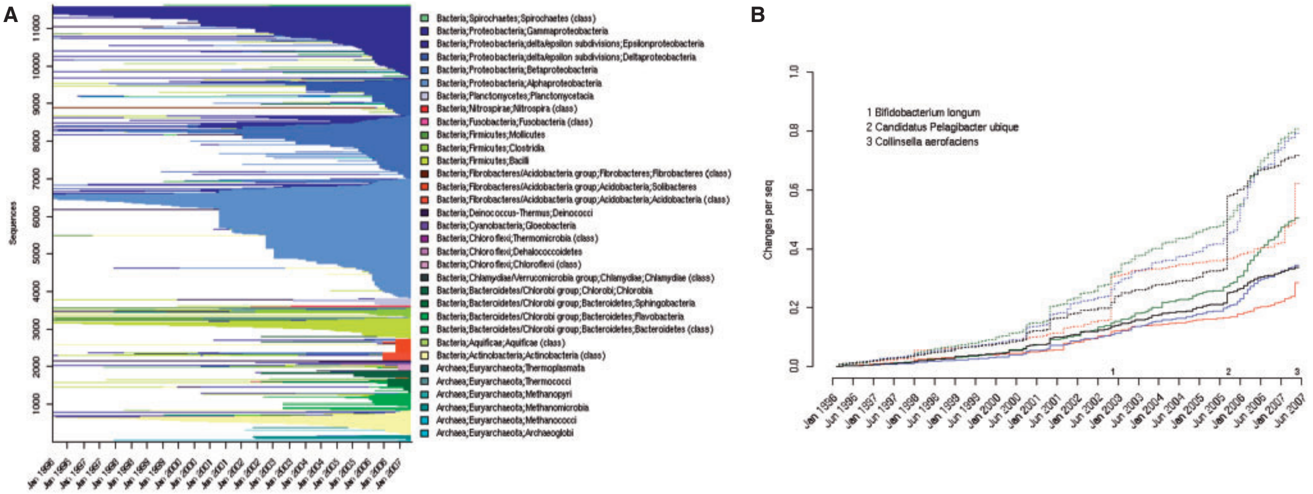

Fig. 1. (**A**) Assignment of a set of ORFs from the farm soil metagenome. We analyzed 60 000 randomly selected sequences from the full metagenome, using blastx searches and a threshold of 60% minimum identity between query and hit proteins. Each row in the plot corresponds to a single ORF, showing with colours how the assignment has varied in time (colours indicate different taxonomic classes). (**B**) This plot shows the accumulated number of ORFs that changed their assignment between consecutive months (expressed as the ratio of number of changes/total number of ORFs), for different metagenomes (Blue, whale fall; Red, human gut; Green, farm soil; Black, Sargasso sea). The changes are divided between new assignments (the ORF was previously unassigned, dashed lines) and assignment changes (assignment to a different taxon, solid lines). The figure also emphasizes three dates in which many assignments change, in accordance to the release of particular complete genomes of importance for the description of these microbial communities.

the few assignments correspond to the release of the two unique completed genomes for this taxon, which were sequenced in 2006. This illustrates how strongly genome sequencing is influencing our knowledge of the universe of proteins, and claims for a sustained effort to sequence more genomes from poorly known taxa.

We wish that these results could help to understand the constraints that the information currently available in the databases imposes to the analysis of metagenomic data, and to improve the current strategies of metagenomic annotation. Additional plots for different metagenomes and taxonomic ranges can be found in our web page http://metagenomics.uv.es/Supp/BI-2008-metagenomics/suppl.html.

ACKNOWLEDGEMENTS

Funding: Financial support was provided by projects BFU2006-06003 from the Spanish Ministry of Education and Science (MEC) and GV/2007/050 from the Generalitat Valenciana, Spain. J.T. is a recipient of a contract in the FIS Program from ISCIII, Spanish Ministry of Health.

Conflict of Interest: none declared.

REFERENCES

Aparicio,G. *et al*. (2007) A grid-enabled software architecture and implementation of parallel and sequential BLAST. In *Proceedings of the Spanish Conference on Science Grid Computing*. Ed CIEMAT, Madrid.

Barns,S.M. *et al*. (1999) Wide distribution and diversity of members of the bacterial kingdom Acidobacterium in the environment, *Appl. Environ. Microbiol.*, **65**, 1731–1737.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.

Krause,L. *et al*. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.*, **36**, 2230–2239.

Mavromatis,K. *et al*. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, **4**, 495–500.

Tamames,J. and Moya,A. (2008) Estimating the extent of horizontal gene transfer in metagenomic sequences. *BMC Genomics*, **9**, 136.

Tringe,S.G. *et al*. (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.

Tringe,S.G. *et al*. (2008) The airborne metagenome in an indoor urban environment. *PLoS ONE*, **3**, e1862.

Venter,J.C. *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.