Inferring functional genome classifications from their annotated proteins

K. Palani Kannan, Markus Göker and Hans-Peter Klenk
Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig, Germany.

Background

It is well known that genome sequences, particularly those of microorganisms, are available at steadily decreasing costs and at exponentially increasing speed. This situation offers much promise for microbial taxonomic classification (Klenk and Goeker, 2010), particularly because methods are available to infer phylogenies from whole genomes. However, a standardized way to classify genomes according to the functionality which they encode – for instance, COGs and inferred pathways – seems not to have been established up to now.

Methodology

The annotated genomes (protein names) were processed based on classification rules for protein names (Fundel and Zimmer, 2006). By implementation of a structured querying system in a locally stored relational database comprising information from Biolexicon (Quochi *et al*. 2008), UniprotKB-GOA (Camon *et al*. 2004), KEGG/KO/COG (Ogata *et al*. 1999, Tatusov *et al*. 2001), proper Clusters of Ortholog Groups (COGs; Tatusov *et al*. 1997) and associated pathways, a system for mapping annotated genomes to (i) COGs and (ii) encoded pathways was established. The mapping to COGs was optimized *via* the comparison with COG annotations of genomes available in the "The integrated microbial genomes system" (Markowitz *et al*. 2010), which are based on BLAST. Basically, the gene product names, COGs and pathways collected from a genome can be regarded as sets (binary presence-absence data) or multisets (abundance data), and (dis-) similarity coefficients using sets as input can be calculated for any pair of genomes. For example, asymmetrical similarity coefficients such as Jaccard's, Sørensen's Steinhaus's (Legendre and Legendre, 1998, pp. 253-268) are promising here. However, the mapping introduces some uncertainty. For instance, the name of a gene product might map to distinct COGs, and, of course, the enzyme related to one COG might participate in several pathways. For this reason, we have complemented the similarity coefficients with a weighting system that reflects this uncertainty. Results from (more) ambiguous mappings are (more severely) deemphasized. In the case of abundance data, this results in coefficients based on the expected abundances, in the case of binary data in coefficients that deal with values between 0 (absence) and 1 (presence).

Conclusion and outlook

A mapping system to derive functional characteristics (KEGG/KO/COG/Pathways) from genomes has been established. The resulting similarity coefficients, which consider the uncertainty related to the mapping, can then be used in conjunction with classification techniques such as clustering and ordination. The resulting genome classifications are expected to group physiologically and biochemically similar organisms together. As next steps, the classifications will be compared in detail with the inferred molecular phylogenies of sets of the target organisms. Presence/absence of genes, functions, and pathways will be reconstructed on the phylogenies and hypotheses of correlated change will be statistically tested, linking evolution and function. The role of horizontal gene transfer for proteins, functions and pathways of interest will be examined.

References

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004) "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology". Nucleic Acids Research 32, D262-D266.

Fundel, K. and Zimmer, R. (2006) "Gene and protein nomenclature in public databases". BMC Bioinformatics 7, 372.

Klenk, H.-P. and Goeker, M. (2010) "En route to a genome-based taxonomy of Archaea and Bacteria?". Systematic and Applied Microbiology 33, 175-182.

Legendre, P and Legendre, L. (1998) “Numerical ecology”. Amsterdam: Elsevier Science.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) “KEGG: Kyoto Encyclopedia of Genes and Genomes”. Nucleic Acids Research 27, 29-34.

Quochi, V., Monachini, M., Del Gratta, R., and Calzolari, N. (2008) “A lexicon for biology and bioinformatics: the BOOTStrep experience”. Proceedings of the LREC’08, 28-30.

Tatusov, RL., Koonin, EV and Lipman, DJ. (1997) “A Genomic Perspective on Protein Families”. Science 278, 631-637.

Tatusov, RL., Natale, DA., Garkavtsev, IV., Tatusova, TA., Shankavaram, UT., Rao, BS., Kiryutin, B., Galperin, MY., Fedorova, ND and Koonin, EV. (2001) “The COG database: new developments in phylogenetic classification of proteins from complete genomes”. Nucleic Acids Res 29, 22-28.

Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N.N. and Kyrpides, N.C. (2010) “The integrated microbial genomes system: an expanding comparative analysis resource”. Nucleic Acids Res 38, D382-D390.