Quality ranking of 16S sequences: an approach based on poset theory

Author(s) Wim De Smet, Bert Verslyppe, Karel De Loof, Paul De Vos, Bernard De Baets, Peter Dawyndt

Institution(s) 1. UGent, Ghent University, K. L. Ledeganckstraat 35, 9000 Gent, Belgium

Abstract:

In the field of microbial taxonomy, the gene sequences coding for the 16S ribosomal RNA (rRNA) are now an integral part of many taxonomic studies. Because of their usefulness in divining the evolutionary past, 16S rRNA sequences of many taxa have proliferated in the International Sequence Database Collaboration (INSDC) databases. Increasingly, several sequences are available for any one species and tools become necessary to support researchers who want to quickly and easily gather available sequences and assess their quality. The StrainInfo project (http://www.straininfo.net/) already provides a useful resource to microbiological researchers, by integrating information about microbiological cultures, available in Biological Resource Centers (BRCs) around the world, on one single strain passport. Included is taxonomic and sequence data that can then be used to automate the sequence selection process. We used a ranking algorithm based on the theory of partially ordered sets to select an appropriate sequence and compare the results with those of the All-Species Living Tree Project. Preliminary work was also done to make this work available to the community through generally available web services. Exploration of the results of this comparison show some limitations in the sequence retrieval process, caused by a lack of usable annotations. Comparison with the results of a manually curated data set reveal several controversial or surprising picks, depending on the quality criteria used. The approach used shows promise as a way to quickly explore available sequences for particular genera and visualize quality differences. Despite some limitations, automated retrieval often finds enough sequences to rank, visualize and build a first approximate phylogenetic tree of any genus with. Leveraging data available within StrainInfo we can make this a single step process, helping researchers to waste less time searching for sequences and more time doing research. Visualization of the poset by the various used criteria is an especially promising way to quickly get an overview of available sequences and their relative merits.

Key words: 16S rDNA, poset, ranking, sequence retrieval, web services