Darwin: an amino acid sequence collection of complete proteomes from eukaryotes with different phylogenetic affinities (v. 03_2020_137)

Joe Win and Sophien Kamoun
The Sainsbury Laboratory, University of East Anglia, Norwich Research Park, Norwich, UK

Background

Every time we find an interesting gene in an organism of interest, the first question is often “how widely is this gene distributed in the eukaryotic kingdom?”. Naturally, one could use NCBI BLAST search against the non-redundant sequence database provided by GenBank to answer this question. However, it can be cumbersome to parse the results and assign them to taxonomic units. It is also not straightforward to get an overview of which eukaryotic groups are represented in the results. Top BLAST hits can be crowded with sequences from closely-related organisms making it difficult gain an overview of the overall distribution across eukaryotes. To streamline this process, we developed an in-house database of complete eukaryotic proteomes. We tagged each sequence with a eukaryotic group handle (two-character symbol) and combined them into a single data set searchable by standalone BLAST on one’s own computer. We named this data set “Darwin” to reflect the diverse nature of the sequences it contains.

Methods

We downloaded predicted proteomes in FASTA format from different sources such as GenBank, Joint Genome Institute (Depart of Energy, USA), Broad Institute (Massachusetts Institute of Technology, USA), Phytozome and a number of other specialized websites catering for a specific organism such as the Arabidopsis Information Resource (TAIR), or the Saccharomyces Genome Database (SGD). All the organisms we included in Darwin are listed in Table 1. To reduce redundancy, we took care not to include the same species more than once unless subspecies were known to show wide diversity. Each sequence header was tagged with a eukaryotic group handle composed of two-character symbols (based on Keeling *et al*., 2005). These handles clearly appear in BLAST output and can be parsed easily. We combined sequences from all proteomes into a single data set and named it “Darwin”.

Results

The current version of Darwin (v. 03_2020_137) contains 2,601,132 amino acid sequences from 137 eukaryotes (Table 1, Data file 1). The sizes of the proteomes were diverse, ranging from ~4000 sequences in some alveolates to 60,000-76,000 in plants. Darwin represents most of the supergroups of eukaryotic kingdom described in Keeling *et al.,* (2005) except those in Rhizaria whose genomes were not available at the time of data set construction. The data set contains larger numbers of proteomes from fungi and plants reflecting areas of interest in our group.

Conclusions

Darwin is provided as a text fatsa file that can be formatted for BLAST searches on standalone computers. The results from the BLAST searches can be parsed to determine how widely a gene of interest is distributed among different eukaryotes. Simple counting of the eukaryotic group handles would also yield an overview of the distribution across taxa. Darwin is also useful for rapidly finding out whether a gene is missing in particular taxa.

Reference

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW (2005) The tree of eukaryotes. *Trends Ecol. Evol.* **20:** 670-676