Evolutionary rate of orthologs and paralogs

Pavlos Pavlidis

January 12, 2023

Abstract

abstract

Introduction

Intro

Materials and Methods

Data and filtering

Data: We downloaded the whole proteome of a set of organisms S from the Ensembl database [Cunningham et al., 2022], using custom scripts (Supplementary File XXXX). The list of organisms is shown in Table ??. Ensembl proteomes are stored in the `https://ftp.ensembl.org/pub/current_fasta/` directory of the Ensembl ftp server and they are organized in separate folders based on the scientific name of the organism (in a folder called 'pep'). They are represented in FASTA format with information-rich headers (i.e., the protein ID, gene ID, transcript ID as well as the location of the protein in the genome is provided). This information allowed us to filter sequences according to some predefined criteria.

Filtering: Prior to the analysis, we applied to filtering procedures on the protein datasets. The first filter refers to *(i) Keep longer protein isoforms.*

For each distinct Ensembl gene ID, we kept only the Protein ID that corresponds to the longest polypeptide sequence. The second filtering procedure refers to *(ii) keep proteins with a minimum length.* As shown in Table 1, a protein length (after applying filter (i)), ranges between less than ten and several thousands of amino acids. We kept only proteins comprise a minimum length of 100 amino acids since this value corresponds to approximate the 5% of protein lengths (Table 1).

Table 1: The percentiles of protein lengths for the organisms used in the study

	0	5	10	50	90	95	100
Canis lupus familiaris	15	100	134	410	1077	1440	27097
Equus caballus	13	110	154	425	1105	1452	34311
Felis catus	13	105	147	425	1096	1461	27108
Homo sapiens	2	107	137	410	1066	1455	35991
Macaca mulatta	17	106	126	409	1084	1419	35478
Mus musculus	3	112	143	384	1033	1401	35390
Pan troglodytes	18	90	120	384	1035	1399	34270
Pongo abelii	4	102	136	411	1068	1430	34347
Sciurus vulgaris	18	89	119	359	983	1315	34292

Results

References

Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, et al. Ensembl 2022. *Nucleic acids research*, 50(D1):D988–D995, 2022.