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 and the number of proteins remained in the dataset after filtering procedures (i) and (ii)

	0	5	10	50	90	95	100	Proteins
Canis lupus familiaris	15	100	134	410	1077	1440	27097	19543
Equus caballus	13	110	154	425	1105	1452	34311	20149
Felis catus	13	105	147	425	1096	1461	27108	18528
Homo sapiens	2	107	137	410	1066	1455	35991	22492
Macaca mulatta	17	106	126	409	1084	1419	35478	21126
Mus musculus	3	112	143	384	1033	1401	35390	21575
Pan troglodytes	18	90	120	384	1035	1399	34270	20660
Pongo abelii	4	102	136	411	1068	1430	34347	19266
Sciurus vulgaris	18	89	119	359	983	1315	34292	20934

Methods

OrthoFinder: All proteomes were processed with OrthoFinder [Emms and Kelly, 2019] with the default settings. The default settings of OrthoFinder use DIAMOND [Buchfink et al., 2015] instead of BLAST for protein comparisons. DIAMOND uses a similar command line interface as the BLAST and offers a similar functionality but it is orders of magnitude faster than BLAST. The default parameters that OrthoFinder uses when it calls DIAMOND are the following:

```
diamond blastp −d DATABASE −q INPUT −o OUTPUT −−more−
    sensitive −p 1 −−quiet −e 0.001 −−compress 1
```

Processing OrthoFinder results: We implemented a home-made python script to process the Orthologs and Orthogroup results of OrthoFinder. The script assesses the 'orthogroup' similarity of the gene neighborhoods between different orthologs as implemented in the following procedure: The ortholog results of OrthoFinder provide the inferred orthologies for *each pair of organisms* used as input for the analysis. Three types of homologies have been inferred (i.e., one-to-one, one-to-many and many-to-many) depending on the order of speciation and duplication events during evolution (see `http://www.ensembl.org/info/genome/compara/homology_types.html` for a comprehensive description of the aforementioned inferred homologies). We focused on the one-to-many type of homology because it allowed us to infer the speed of evolution between the pair of homologue genes that belong to a genomic regions with either a large or low proportion of homologous genes.

Let $S1$ and $S2$ represent two species for which OrthoFinder results have been obtained and one-to-many types of relationships have been computed. Let X be a gene in $S2$ (the 'one') for which we have identified many orthologs, e.g., A, B, C, D in $S1$ (the 'many'). We set as *neighSize* the size of the neighborhood of each gene (for example 10 genes on each side of the gene). Then, we examine the orthogroups of each of the genes in the neighborhood of the orthologues and we assess the similarity of the orthogroup consitution in the neighborhoods of each gene, one from each species $S1$ and $S2$. Figure 1 illustrates the neighborhood approach.

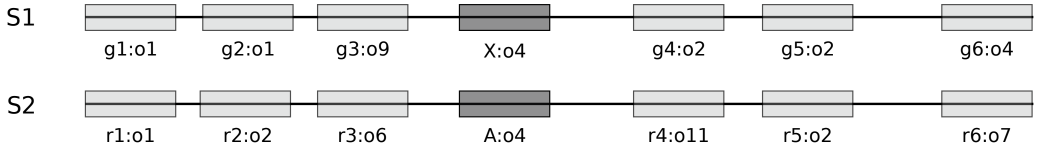

Figure 1: An illustration of the neighborhood similarity in terms of Orthogroups. Two neighborhoods of species $S1$ and $S2$ (top and bottom, respectively) have been drawn. Genes in $S1$ are characterized with the letter **g**, whereas in species $S2$ with the letter **r**. The two focal genes are represented with the letter **X** and **A**, respectively. Orthogroups are denoted with the letter **o**. Thus, **g1:o1** represents the gene 1 in species $S1$ which belongs to orthogroup 1. The two species, together, consist of 7 (**o1**, **o2**, **o4**, **o6**, **o7**, **o9** and **o11**) orthogroups. Out of them, 3 orthogroups (**o1**, **o2**, **o4**) belong to both neighborhoods. Thus, the percent of similarity is $3/7 \approx 42.8\%$

Results

A correlation example between human and cat

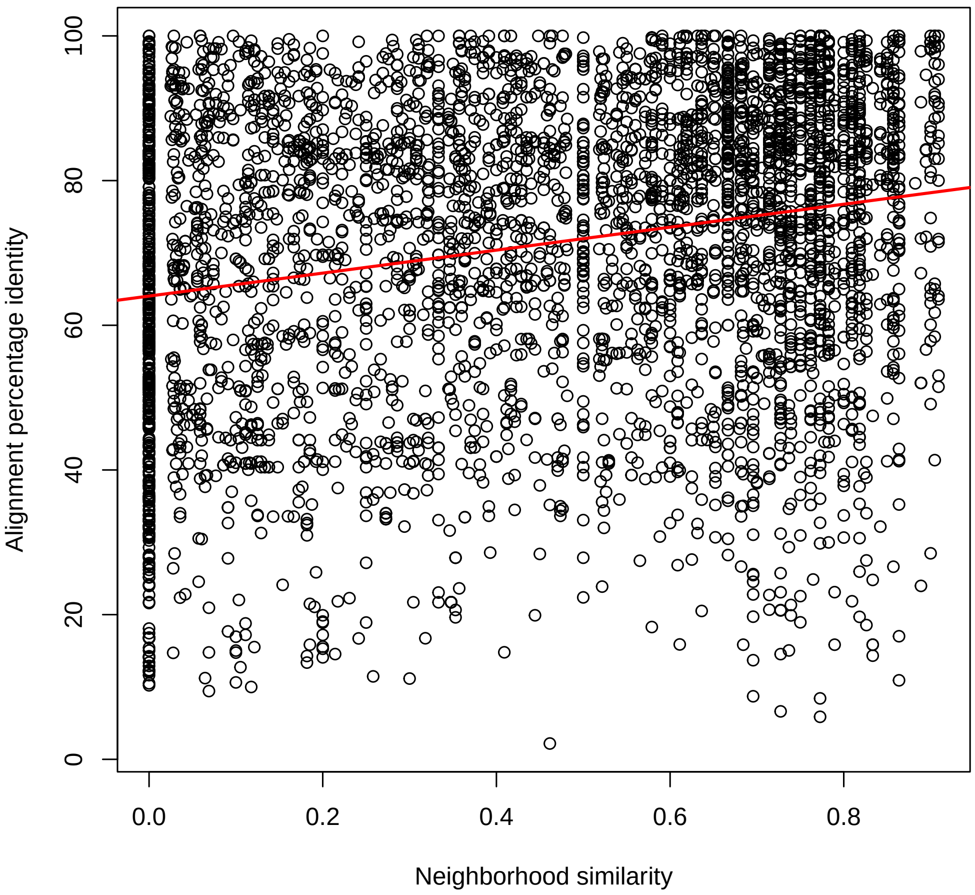

Figure 2: The relation between *Homo sapiens* and **Felis catus** regarding the orthogroup similarity of neighborhoods and percentage of alignment identity between two orthologous genes. The regression coefficient value is positive (15.873) and highly significant (pvalue $< 10^{-16}$) illustrating that orthologous genes that are in similar neighborhoods (in terms of orthogroups) show less differences between them.

References

Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.

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.

David M Emms and Steven Kelly. Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20(1):1–14, 2019.