Letter to the editor.

Dear editor of Evolutionary Bioinformatics,

In this paper, we introduce a simple Python script that allows automatic generation of orthology tables (known as phylogenetic profiles) from the KEGG database. We believe that using this script will facilitate the work of our colleagues.

Best wishes,

Szymon Kaczanowski.

We suggest following reviewers:

Scientific paper.

**A tribute to Dariusz Izak -Prwlr (profiles crawler) software rapidly generates tables of orthologs of input gene lists Rapid generation of tables of orthologs (A tribute to Dariusz Izak)**

Dariusz Izak and Szymon Kaczanowski

Department of Bioinformatics, Institute of Biochemistry and Biophysics Polish Academy

of Science, Warsaw, Poland

Corresponding author: szymon@ibb.waw.pl

The evolution of functional protein modules, such as complexes and biochemical pathways, is a fundamental question of molecular biology with implications for whole biology. However, studies on the evolution of functional protein modules are best based on an analysis of orthology in different species, as this reveals those species in which the ortholog of a given module subunit is present and if it is duplicated (or multiplicated). Unfortunately, the existing tools are labor demanding and frequently generate erroneous results. To facilitate the generation of orthology tables, we have developed Prwlr (profiles crawler) software. Prwlr uses known KEGG orthology for automatic determination of who is the ortholog of whom. This software saves time and generates a table of orthology for later manual curation. The resulting table can then be used as a starting point for more complex analysis.

Main body of the article

Analysis of molecular evolution is one of the basic methods of current biology. Its prominence is a direct consequence of the fact that biology is based on experiments that are performed in different model organisms. However, comparison of observations performed in different systems is best conducted by analyzing the similarities and differences of these systems. Evolutionary studies provide the key information required for these types of comparisons.

The correct detection of homology and orthology is a starting point for any analysis of evolution. Evolutionarily related proteins have a similar structure and function (1, 2), and this observation is widely used in homology modeling of 3D protein structure. Homology modeling is a computer method of constructing an atomic-resolution model of a “target” protein from its amino acid sequence and then creating an experimental three-dimensional structure of a related homologous protein (the "template")(2).

Orthologous proteins usually have a very similar function, and this is particularly true in the case of orthologs (3). Orthologs are homologous sequences that originate from the same ancestors (homologs; e.g., all globin proteins), but they have been separated from each other after different speciation events.   
As a result, the analysis of homology and orthology serves as a starting point for analysis of the evolution of any functional modules, such as protein complexes and biochemical pathways, as the analysis indicates which proteins play similar roles in a given machinery in different organisms. The presence or absence of an ortholog of a given gene in the genome of an investigated organism suggests that the mechanism and trait encoded by that gene is not present in studied organism (4). “Multiplication” of the orthologs of a given gene, in turn, indicates the occurrence of some change in the analyzed mechanisms (see for example (5)).

For this reason, the analysis of homology and orthology is a basic starting point for the analysis of any biological mechanisms widely used in current biology. For example, our group has applied this approach in studies on the evolution of apoptosis (6) and of SWI/SNF2 complexes(5). Nevertheless, the prediction of orthology remains a complex and labor-demanding task.

A number of useful approaches have been proposed for orthology prediction. For example, databases of orthologs, such as COG (3) and KEGG (7), provide information about which protein gene shows orthology among genes in different known genomes. Well-known algorithms, such as reciprocial blast (8), in-paranoid (9, 10),DODO(11), or Domainoid (12), are used for the prediction of orthology between genes in genomes of different organisms. Analysis of the PFAM (11-13)domain architecture is also frequently applied (for example, in our study(6)). The calculation of phylogenetic trees can then reveal details of evolution of the orthologous genes.

Our own experience indicates that generating orthology tables, known as phylogenetic profiles, using known internet databases is particularly labor intensive. Despite being a seemingly very easy approach, obtaining the information requires a lot of work with internet servers and manual downloading of many www pages. For these reasons, we have developed the Prwlr (profiles crawler) software presented in this paper. Prwlr uses KEGG (7) orthology for automatic determination of “who is the ortholog of whom.”

Below, I describe briefly how to install this software and how to use it.

**Installation:**

The user should have a Unix or Mac system with Python installed.

Use the command **git clone https://github.com/dariusz-izak-doktorat/prwlr3**, which creates the directory containing the imported software. Go to this directory and use the command pip install -r requirements.txt (Python 2) or pip3 install -r requirements.txt (Python 3). This command will install all dependencies.

**Running the test function:**

sh test.sh python (python)

sh test.sh python3 (python3)

sh test.sh (python3.7)

**Preparing the input files:**

1. **Names of species**

First, input files containing the list of species of interest (one taxonomic name per line). Use the names according to the KEGG (www page <https://www.genome.jp/kegg/catalog/org_list.html>), using only the two Latin words (e.g., Homo sapiens, not Homo sapiens (human)).

1. **Names of genes**

Next, input the file containing the list of genes. Again, insert each gene one per line. **The names of the applied genes should be those used in the KEGG database.**

**Running the prwlr script:**

The syntax is inspired by the classical Unix command line.

**python prwlr.py [name file1] [name file 2] output\_name**

e.g. **python prwlr.py species\_list list\_apoptotic\_genes apoptosis** (python)

or **python3 prwlr.py species\_list list\_apoptotic\_genes apoptosis** (python3)

The software then produces the following files of phylogenetic profiles, containing + and – symbols, in Excel and tabulate format.

We also wrote a second script named **prwlr\_ko.py**, which facilitates obtaining the lists of orthologs from different species. The applied syntax is identical to that used in the case of the first script.

**python prwlr\_ko.py [name file1] [name file 2] output\_name** (python)

e.g . **python3 prwlr.py species\_list list\_apoptotic\_genes apoptosis** (python3)

This script produces the following files containing orthologous gene lists from different species in Excel and tabulate format.

An example is shown in the prwlr directory (test and test1 files). The example presents the orthology table of the following yeast apoptotic genes. A similar analysis was performed by our team and published previously.

The prwlr software only enhances the speed of obtaining information from the KEGG server. The user should keep in mind that the results obtained are the starting point for the analysis. For instance, in the example shown, the human gene CPP32B (caspase 3) has no orthologs in invertebrates. This observation indicates that the evolutionary history of the caspase family is complex and could represent a starting point for a longer analysis. I believe that this software may save time for the users and enhance the speed of evolutionary studies.

# Acknowledgments

This work was supported by grants (2017/27/B/NZ8/02502) from the Polish National Science Centre. This work is a tribute to Dariusz Izak, a former graduate student of our group, whose great passion to evolutionary bioinformatics inspired me and other senior collegues. The presented software is a modest attempt to finish his excellent software pwrl developed by him during his undergraduate studies. We hope that he will return to his great passion in future.

**References**

1. J. O. Korbel *et al.*, Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* **3**, e134 (2005).

2. S. Kaczanowski, P. Zielenkiewicz, Why similar protein sequences encode similar three-dimensional structures? *Theoretical Chemistry Accounts* **125**, 643-650 (2010).

3. R. L. Tatusov *et al.*, The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

4. J. O. Korbel *et al.*, Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* **3**, e134 (2005).

5. T. J. Sarnowski, S. Swiezewski, K. Pawlikowska, S. Kaczanowski, A. Jerzmanowski, AtSWI3B, an Arabidopsis homolog of SWI3, a core subunit of yeast Swi/Snf chromatin remodeling complex, interacts with FCA, a regulator of flowering time. *Nucleic Acids Res* **30**, 3412-3421 (2002).

6. J. Klim, A. Gładki, R. Kucharczyk, U. Zielenkiewicz, S. Kaczanowski, Ancestral State Reconstruction of the Apoptosis Machinery in the Common Ancestor of Eukaryotes. *G3 (Bethesda)* (2018).

7. M. Kanehisa *et al.*, KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480-484 (2008).

8. D. P. Wall, H. B. Fraser, A. E. Hirsh, Detecting putative orthologs. *Bioinformatics* **19**, 1710-1711 (2003).

9. K. P. O'Brien, M. Remm, E. L. Sonnhammer, Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**, D476-480 (2005).

10. G. Ostlund *et al.*, InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**, D196-203 (2010).

11. T. W. Chen, T. H. Wu, W. V. Ng, W. C. Lin, DODO: an efficient orthologous genes assignment tool based on domain architectures. Domain based ortholog detection. *BMC Bioinformatics* **11 Suppl 7**, S6 (2010).

12. E. Persson, M. Kaduk, S. K. Forslund, E. L. L. Sonnhammer, Domainoid: domain-oriented orthology inference. *BMC Bioinformatics* **20**, 523 (2019).

13. R. D. Finn *et al.*, The Pfam protein families database. *Nucleic Acids Res* **38**, D211-222 (2010).