OrFin: A web tool for detection of putative orthologs

Mohit Midha[1], Raja Polavarapu[1], Potshangbam Angamba Meetei[1], Hari Krishnan[1], Krishnaveni Mohareer[2] & Vaibhav Vindal[1]*

[1]Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad 500046, India; [2]Institute of Life Sciences, Hyderabad; Vaibhav Vindal–Email: vvls@uohyd.ernet.in; Phone: +91-40-23134583, Fax: +91-40-23010120; *Corresponding author

Abstract:
Identification of ortholog is one of the important tasks to understand a novel genome. It helps to assign functional annotations, from one organism to another organism. To identify the putative ortholog, Reciprocal Best BLAST hit (RBBH) method is known to be an efficient approach. OrFin makes use of the same approach to identify pair of orthologous proteins for a given set of sequences of two species. It is a user-friendly web tool which works with user defined parameters to search RBBHs. Results are produced in both html and text format.

Availabilty: This web tool is freely available at http://bifl.uohyd.ac.in/orfin

Keywords: Bioinformatics Tool, Ortholog prediction

Background:
In the post-genomic era, genome-sequencing projects are progressing at a fast pace. Hence the availability of the genome sequences has generated vast sequence data that is available via public domain databases. Complete genome sequences pose challenges and opportunities to a computational biologist to understand the genome function and its complexity **[1-3]**. To explore the hidden genomic information, building the relationship between the genomes of various species is an important step to begin. Ortholog identification is used efficiently to compare and to understand the functional aspect of un-annotated genes present in the genome.

Reciprocal Best BLAST Hits (RBBHs) is known to be an efficient approach to identify the orthologs **[4-6]**. Our web tool, OrFin also makes use of this approach to identify pair of orthologous proteins for a given set of two proteomes. It allows user to alter the criteria to retrieve the RBBHs. It is user friendly with a web interface that will have potential implications to assist features associated with orthologous proteins.

Algorithm:
The web server takes the input sequences for two organisms and returns the orthologous pair of proteins. Firstly, it filters the identical proteins, if any, followed by Reciprocal Best Blast Hits methods to retrieve the orthologs. Ortholog search for these multiple identical proteins is carried out with one of the protein sequence as a representative. Later in the results, representative protein is replaced with actual protein/ORFs along with their identified orthologous proteins. User defined parameters are incorporated in this web-tool. Flow chart of the methodology followed is depicted in **Figure 1.**
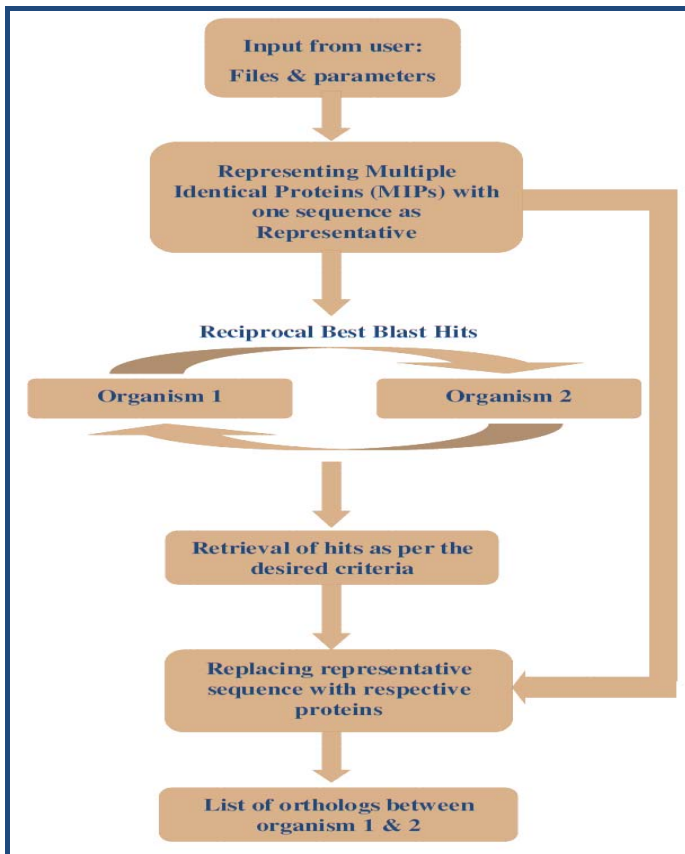
BIOINFORMATION

Figure 1: Flowchart illustrating the methodology

Figure 2: Screenshot of OrFin with input parameters and results as an example

Web interface:

The web interface is provided through PHP. In client tier, we have allowed the user to go through different options, e.g. selection of listed organisms or to provide organism details. In later case files (.faa) are to be uploaded. After successful submission of the data our application will retrieve the orthologous pairs of proteins. It also allows the user to choose E-Value for BLAST and alignment length criteria to retrieve the RBBHs. Additionally, user can also upload genome coordinate files (.ptt) for each of the genomes submitted in order to retrieve a well formatted output which can be Protein IDs, ORF IDs and their combination with gene names. Apache HTTP Web Server is used for the management of communication between different tiers or layers of the application. Programs running at the backend of OrFin have been written using Perl-CGI scripts. At the backend, mysql database server is used for storing all the relevant data and user outputs. An example of executing a job using OrFin server has been shown in **Figure 2.**
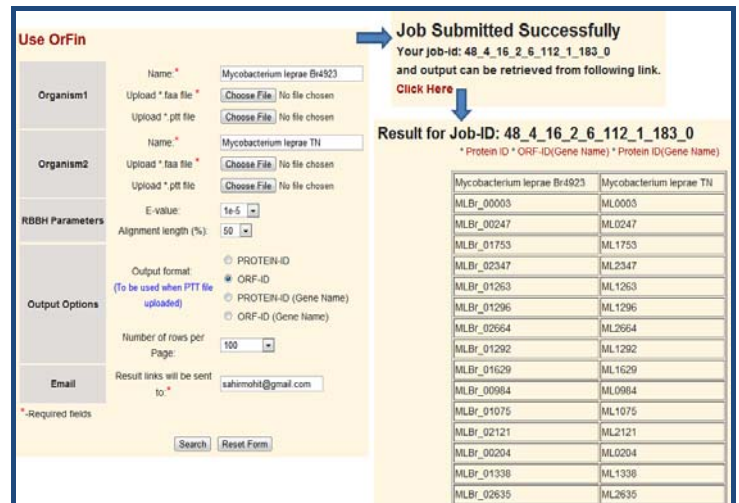
Applications:

OrFin is built to identify the putative orthologous proteins between two proteomes. It will have potential implications to employ comparative genomics to assist not only in functional annotation but also in phylogenetic footprinting.

References:
[1] Tsoka S & Ouzounis CA, *FEBS Lett.* 2000 **480**: 42 [PMID: 10967327]
[2] Young DB, *Nat Med.* 2001 **7**: 11 [PMID: 11135599]
[3] Mirny LA & Gelfand MS, *Genome Biol.* 2002 **3**: PREPRINT0002 [PMID: 11897020]
[4] Altschul SF *et al. J Mol Biol.* 1990 **215:** 403 [PMID: 2231712]
[5] Moreno-Hagelsieb G & Latimer K, *Bioinformatics* 2008 **24**: 319 [PMID: 18042555]
[6] Midha *et al. PLoS ONE.* 2012 **7**: e36094 [PMID: 22563442]