ORFanID: A Search Engine to Discover De Novo Orphan Genes

Richard S. Gunasekera, Ph.D.[1,2], Suresh Hewapathirana, M.Sc.[3], Vinodh Gunasekera, M.E.E.[2,3], and Paul Nelson, Ph.D[4].

[1]Deptartment of Biological Sciences, University of Houston-Victoria, 3007 N. Ben Wilson Drive, Victoria, TX 77901, [2]Rice University, 6100 Main St. Houston, TX 77005, [3]Bioinformatics Consultants, Chesalon LLC, 5527 Theall Road, Houston, TX 77066, [4]Biola University, 13800 Biola Avenue, La Mirada, CA 90639

Introduction

Orphan genes are unique DNA sequences found in species that do not have orthologous DNA sequences in related species or at corresponding taxonomic rank in GenBank or other such databases. The ubiquitous presence of these orphan genes, also known as Taxonomically Restricted genes, in various sequenced genomes is a mystery and a problem to be solved in Genetics. ORFanID is a standalone and web-based software engine that identifies ORFan genes from the genomes of specified species or from a given list of gene sequences. The scope of the search for orphan genes can be defined by the selection of the taxonomy level of interest. Detectable homologous sequences are found for candidate gene in the NCBI databases. From these findings the ORFanID engine identifies and depicts orphan genes. Results may be viewed and analyzed graphically for the purpose of scientific research and inquiry.

Methods and Procedures

ORFanID accepts the amino acid sequence of a single or multiple gene sequence in the FASTA format with their NCBI gene IDs(GI) along with the organism name and the taxonomy level of the selected organism for appropriate restriction of the search. Figure 1 shows the input frame where the gene sequence, the name of the organism, and the taxonomy is specified.

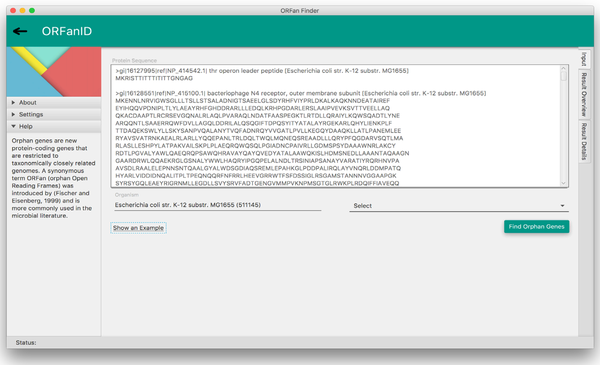

Figure 1: ORFanID Sequence Input Frame

ORFanID engine executes several scripts as a pipeline as follows:

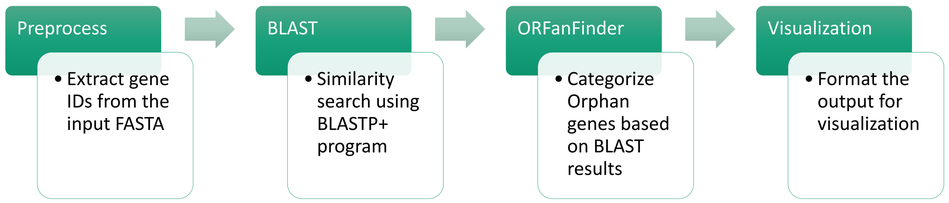

Figure 2: ORFanID Script Pipeline

The execution sequence is as follows:

1. ORFanID extracts the gene ID(s) from the header(s) of the input FASTA file.
2. All the genes are submitted to the online NCBI BLASTP[1] program to find homologous sequence against NCBI non-redundant protein database. Accuracy of the results can be adjusted based on e-value and maximum number of targets sequences and the taxonomy level (to restrict the search), and the results will be reported in a tabular format.
3. Genes will be categorized according to a filtering algorithm by the ORFanFinder command-line tool[2] , based on the homology sequences found from blast results.
4. Output of the ORFanFinder is formatted to produce tables and charts for interactive analysis and visualization.

ORFanID Settings (figure 3):

The settings frame of ORFanID allows the user to specify the database, taxonomy, species files used for input. For the orphan gene search, the maximum e-value and target sequences may also be specified. The Blast database may be selected as local or online and the number of threads used for computation can be defined for software load balancing purposes.

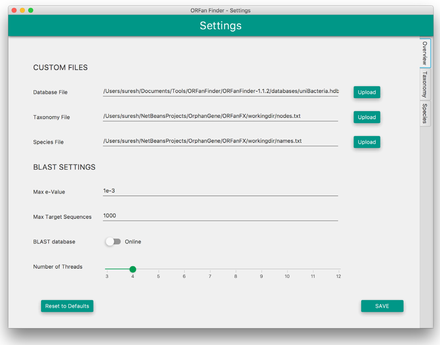

Figure 3: ORFanID Settings

Results

An example of the interactive visualization of the ORFanID is given below (figure 4):

In figure 4, the top left table summarizes the results by showing the number of orphan genes found at each taxonomy level of the selected species (Bacteria, in this example). The graph on the top right displays the same data graphically to improve analysis. The table at the bottom shows the categorization of the orphan genes discovered, along with the taxonomy level for each individual gene. Once the user selects a gene from the bottom table, the homologous sequence found from BLASTP will be displayed in a “Results Detail” tab (fig. 5)

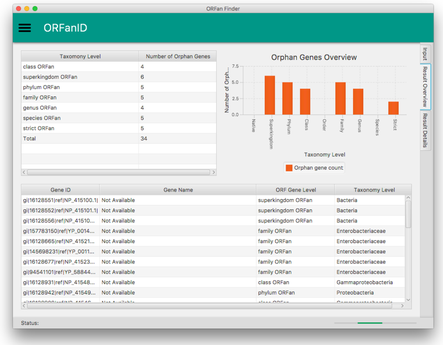

Figure 4: ORFanID Results Page

In figure 5, the table at the bottom of the window to the left shows the taxonomy levels of different homologous BLAST hits for the selected gene. The number of BLAST matches for each taxonomy are interactively visualized in the chart. For example, if there are hits for the Family level, but not in Phylum, Class, Order, then that gene is specifically restricted to Family, therefore it’s a Family level orphan gene. All the tables in the software are sortable by column and results can be easily filtered with a fuzzy search function.

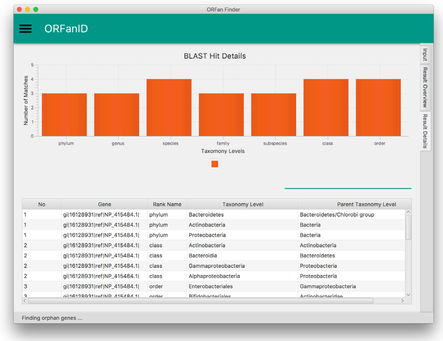

Figure 5: ORFanID Results Detail

The ORFanID Web-based application[3] provides additional functionality such as table pagination, and exporting results in PDF or CVS formats.

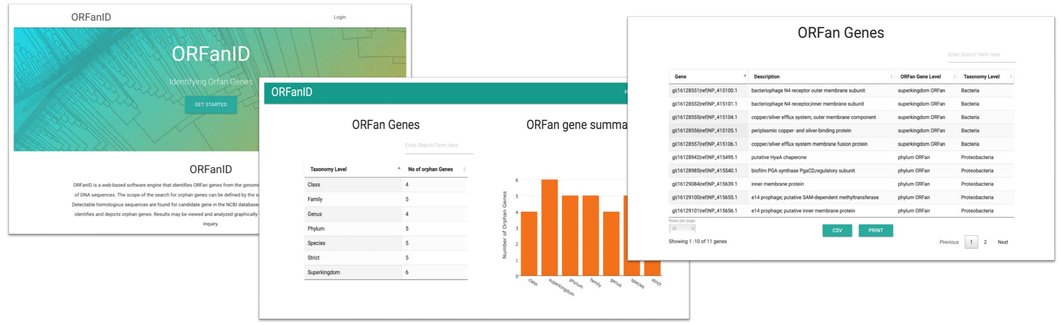

Figure 6: ORFanID Collective

Discussion

By identifying these unique DNA sequences, ORFanID can help discover the origin, function and other significance of orphan genes. The software is able to identify genes unique to genus, family, or species etc. at differing taxonomy levels. Based on the parameters specified, some of orphans (Taxonomy Restricted Genes) may or may not fall under the given classification for strict ORFans. As such, ORFanID can help delineate the actual sequence and function of *de novo* genes discovered in species and at all levels of the taxonomy tree.

References

[1]. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410 [2]. Ekstrom, A. & Yin, Y. (2016) "ORFanFinder: automated identification of taxonomically restricted orphan genes." Bioinformatics; 32 (13): 2053-2055. doi: 10.1093/bioinformatics/btw122 [3]. http://orfangenes.org