*Bioinformatics*, 32(13), 2016, 2053–2055
doi: 10.1093/bioinformatics/btw122
Advance Access Publication Date: 7 March 2016
Applications Note

OXFORD

Sequence analysis

# ORFanFinder: automated identification of taxonomically restricted orphan genes

**Alex Ekstrom[1] and Yanbin Yin[2]***

[1]Department of Computer Science and [2]Department of Biological Sciences, Montgomery Hall 325A, Northern Illinois University, DeKalb, IL, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on January 20, 2016; revised on February 25, 2016; accepted on February 26, 2016

## Abstract

**Motivation:** Orphan genes, also known as ORFans, are newly evolved genes in a genome that enable the organism to adapt to specific living environment. The gene content of every sequenced genome can be classified into different age groups, based on how widely/narrowly a gene's homologs are distributed in the context of species taxonomy. Those having homologs restricted to organisms of particular taxonomic ranks are classified as taxonomically restricted ORFans.
**Results:** Implementing this idea, we have developed an open source program named ORFanFinder and a free web server to allow automated classification of a genome's gene content and identification of ORFans at different taxonomic ranks. ORFanFinder and its web server will contribute to the comparative genomics field by facilitating the study of the origin of new genes and the emergence of lineage-specific traits in both prokaryotes and eukaryotes.
**Availability and implementation:** http://cys.bios.niu.edu/orfanfinder
**Contact:** yyin@niu.edu
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

ORFans stand for orphan Open Reading Frames (Fischer and Eisenberg, 1999). In addition to ORFans, in literature orphan genes are also known as new genes, lineage-specific genes (LSGs) and taxonomically restricted genes (TRGs) (Tautz and Domazet-Loso, 2011), although some terms are used in very specific contexts, e.g. *de novo* originated new genes. Orphan genes (hereafter ORFans) are new inventions of a genome that enable the organism to adapt to its specific living environment. They are fundamentally important for the origin of new species and the emergence of lineage-specific traits, e.g. morphological diversity, metabolic innovation and pathogenicity (Chen *et al.*, 2013; Khalturin *et al.*, 2009).

ORFans are usually identified by using sequence similarity search tools such as BLAST (Fischer and Eisenberg, 1999; Yin and Fischer, 2006). The simple BLAST method tends to identify very narrowly distributed ORFans and cannot classify ORFans into groups of different ages. Therefore the phylogeny-based approach was developed to map BLAST hits onto a species phylogeny, which allows identifying ORFans of different ages. One method using this approach (Daubin and Ochman, 2004) required a well-accepted species phylogeny, often not available for non-model organisms. Another method termed phylostratigraphy approach (Domazet-Loso *et al.*, 2007) also required a phylogeny, which however is derived from the query species' taxonomic lineage. Although the phylostratigraphy approach was recently shown to have some limitations (Moyers and Zhang, 2015), it has been widely used to study ORFans in various organisms (e.g. Carvunis *et al.*, 2012; Guo, 2013). However, the phylogeny-based methods are more difficult to automate and often need human intervention.

Here, we present a new tool ORFanFinder and its web server. ORFanFinder is fully automated and the results are completely reproducible. While the algorithm behind ORFanFinder is essentially based on the phylostratigraphy approach (Domazet-Loso *et al.*, 2007), it is more strictly formulated for the purpose of complete automation. To our knowledge ORFanFinder is the only open source computer software and the first online server for ORFan identification and classification.

## 2 Algorithm

In our algorithm, the query genome and all the sequences in the databases should have their taxonomic ranks known. For example, *Escherichia coli* MG1655 has the following taxonomic lineage: *Bacteria (superkingdom); Proteobacteria (phylum); Gammaproteobacteria (class); Enterobacteriales (order); Enterobacteriaceae (family); Escherichia (genus); Escherichia coli (species)* according to the NCBI's Taxonomy database.

*Step 1*: If the query genome Q is in the databases, exclude it before doing BLAST search. *Step 2*: In the BLAST result, for each gene g of the query genome Q, record the subject genomes ($H_i$, $i = 1, \ldots, k$) that g has hit(s) in. *Step 3*: Suppose Q has the following taxonomy ranks: species S, genus G, family F, order O, class C and phylum P. Our program considers all named ranks (including e.g. subclass and tribe), while 'no ranks' are excluded because they appear at multiple levels in the hierarchy. *Step 4*: Examine the taxonomy ranks of the subject genomes $H_i$: which phyla (let $P_h$ be an array of phyla of the subject genomes) and how many phyla (let $\#P_h$ be the number of elements in $P_h$) do they belong to? If they all belong to a same phylum, which classes ($C_h$) and how many classes ($\#C_h$) do they belong to? This is done recursively until it reaches the bottom species rank. *Step 5*: For each gene g of the query genome Q, determine if it is an ORFan and if yes, classify it using the pseudocodes shown in Fig. 1 (include only 7 ranks for simplicity but more ranks could be possible). This algorithm is very efficient in genome-wide ORFan identification, with the limitation that sporadically distributed genes (often horizontally transferred or rapidly evolved) are treated equally as universally distributed genes in our algorithm.

In addition to similarity information, other information like syntenic (gene order) information has also been used to assist ORFan identification in closely related species (Zhang *et al.*, 2010). Incorporating this information in the ORFanFinder algorithm is currently not possible as determining gene order itself is a highly challenging problem: gene order is far less conserved than gene sequences.

```
if #P_h == 0
  It's a strict ORFan
elseif #P_h > 1 and P ∈ P_h
  It's a native gene
elseif #P_h == 1 and P ∈ P_h
  if #C_h > 1 and C ∈ C_h
    It's a phylum ORFan
  elseif #C_h == 1 and C ∈ C_h
    if #O_h > 1 and O ∈ O_h
      It's a class ORFan
    elseif #O_h == 1 and O ∈ O_h
      If #F_h > 1 and F ∈ F_h
        It's an order ORFan
      elseif #F_h == 1 and F ∈ F_h
        If #G_h > 1 and G ∈ G_h
          It's a family ORFan
        elseif #G_h == 1 and G ∈ G_h
          If #S_h > 1 and S ∈ S_h
            It's a genus ORFan
          elseif #S_h == 1 and S ∈ S_h
            It's a species ORFan
```

**Fig. 1.** Pseudocodes of the algorithm

## 3 Implementation

*Standalone program*: The ORFanFinder program was written in C language. It expects a BLAST search result file (tabular format) as the input. For example, to identify ORFans in *E. coli* MG1655 genome one has BLASTed all its 4141 proteins against the NCBI nr database. ORFanFinder program will require the following as the inputs: (i) the taxonomy ID of *E. coli* MG1655, which is 511145; (ii) a query ID file with all the 4141 protein IDs of MG1655; (iii) a taxonomy node file parsed from NCBI taxonomy database's nodes.dmp file; (iv) a NCBI nr ID—taxonomy ID mapping file. In fact, along with the ORFanFinder program, the software package has included pre-computed files to be used in (iii) and (iv), if one has searched against the NCBI nr database or the UniProt database. If one intends to use a different BLAST database, we have detailed instructions in a readme file on how to make a customized file for (iii) and (iv). The output of ORFanFinder is a tab-delimited file containing two columns: protein IDs of the query genome and ORFan group (e.g. species ORFan).

*Web server*: In addition to the standalone program, a web server was developed for users who do not have programming experience. Two types of data are allowed to submit to our web server: (i) FASTA format sequences, which will be used to run BLAST against NCBI-nr and then run ORFanFinder on a Linux cluster; (ii) tabular format (-outfmt 6) BLAST result file, which is pre-computed by the user elsewhere, e.g. on their own computers. It should be noted that running BLAST search against a large database such as NCBI-nr is the most time-consuming, while the ORFanFinder program itself runs relatively much faster.

## 4 Evaluation

To assess ORFanFinder's performance, we have run it on the proteomes of *E. coli* MG1655 (Taxonomy ID: 511145) and *Arabidopsis thaliana* (Taxonomy ID: 3702). ORFans of the two species have been extensively studied in previous papers (e.g. Daubin and Ochman, 2004 and Yu and Stoltzfus, 2012 for *E. coli* and Donoghue *et al.*, 2011 and Lin *et al.*, 2010 for *A. thaliana*). Comparing to these previous ORFan sets shows that ORFanFinder performs fairly well (sensitivity = 91.9% and specificity = 71.5% for *E. coli* and sensitivity = 91.6% and specificity = 72.2% for *A. thaliana*). The detailed results are available in Supplementary data.

## Acknowledgements

We acknowledge the Department of Computer Science of NIU for providing free access to the Linux computing cluster Gaea and the Yin lab members for helpful discussions.

## Funding

This work has been supported by the National Institutes of Health (1R15GM114706), the Research & Artistry Award and the startup package from Northern Illinois University to YY.

*Conflict of Interest*: none declared.

## References

Carvunis,A.R. *et al.* (2012) Proto-genes and de novo gene birth. *Nature*, **487**, 370–374.

Chen,S.D. *et al.* (2013) New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.*, **14**, 645–660.

Daubin,V. and Ochman,H. (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res*., **14**, 1036–1042.

Domazet-Loso,T. *et al*. (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet*., **23**, 533–539.

Donoghue,M.T.A. *et al*. (2011) Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evol. Biol*., **11**, 47.

Fischer,D. and Eisenberg,D. (1999) Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.

Guo,Y.L. (2013) Gene family evolution in green plants with emphasis on the origination and evolution of Arabidopsis thaliana genes. *Plant J*., **73**, 941–951.

Khalturin,K. *et al*. (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*., **25**, 404–413.

Lin,H.N. *et al*. (2010) Comparative analyses reveal distinct sets of lineage-specific genes within Arabidopsis thaliana. *BMC Evol. Biol*., **10**.

Moyers,B.A. and Zhang,J. (2015) Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol*., **32**, 258–267.

Tautz,D. and Domazet-Loso,T. (2011) The evolutionary origin of orphan genes. *Nat. Rev. Genet*., **12**, 692–702.

Yin,Y. and Fischer,D. (2006) On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. *BMC Evol. Biol*., **6**, 63.

Yu,G. and Stoltzfus,A. (2012) Population diversity of ORFan genes in Escherichia coli. *Genome Biol. Evol*., **4**, 1176–1187.

Zhang,Y.E. *et al*. (2010) Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Res*., **20**, 1526–1533.