BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 9(5)

OrthoRBH: A streamlined pipeline for mining large gene family sequences in related species

Mark Ziemann[1]*, Atul Kamboj[2] & Mrinal Bhave[2]

[1]Baker IDI Heart and Diabetes Institute, Melbourne, VIC 3004, Australia; [2]Environment and Biotechnology Centre, Faculty of Life and Social Sciences, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia; Mark Ziemann – Email: mziemann@bakeridi.edu.au; *Corresponding author

Received January 21, 2013; Accepted February 08, 2013; Published March 02, 2013

Abstract:
Plant and animal genomes are replete with large gene families, making the task of ortholog identification difficult and labor intensive. OrthoRBH is an automated reciprocal blast pipeline tool enabling the rapid identification of specific gene families of interest in related species, streamlining the collection of homologs prior to downstream molecular evolutionary analysis. The efficacy of OrthoRBH is demonstrated with the identification of the 13-member *PYR/PYL/RCAR* gene family in *Hordeum vulgare* using *Oryza sativa* query sequences. OrthoRBH runs on the Linux command line and is freely available at SourceForge.

Availability: http://sourceforge.net/projects/orthorbh/

Key words: Reciprocal blast, Gene family, Molecular evolution, Orthology.

Background:
Identification of orthologous genes and gene families is a key task in molecular biology and comparative genomics. Accurate ortholog identification enables the transfer of functional gene information from one species to another and is an important step in the annotation of newly sequenced genomes. There are a multitude of tools to perform this genome-wide annotation such as InParanoid **[1]** and OrthoMCL **[2]**, however from the perspective of the biologist, there is a distinct lack of tools for the fast, targeted identification of specific gene families of interest from newly sequenced transcriptome databases or expressed sequence tag collections.

Methodology:
The reciprocal blast method is widely used as a tool for the identification of orthologous sequences **[3]**. The OrthoRBH method, schematically illustrated in **(Figure 1)** begins with a set of protein sequences from a gene family of interest used as tblastn queries to search a transcript database. Tblastn forward blast is used because protein homology is likely to be greater than the DNA homology, and tblastn also allows for the mining of expressed sequence tags instead of curated protein sequences. Forward blast hits with an e-value better than the threshold are classified as "candidate" orthologs and are return blasted (either blastn or tblastx). If the candidate's return-blast best match is any of the original forward blast queries with an e-value satisfying the threshold, then the sequence is confirmed. The nucleotide transcript sequences corresponding to the original queries and the newly confirmed orthologs are deposited to a fasta file for ClustalW alignment **[4]** which can be used in downstream phylogenetic and molecular evolutionary analysis. If the database to be queried consists of expressed sequence tags and/or next generation sequencing reads, OrthoRBH can identify homologous sequence reads and assemble them into contigs using CAP3 **[5]**. The assembled contigs are then deposited into a fasta file alongside the transcript sequences corresponding to the original queries.

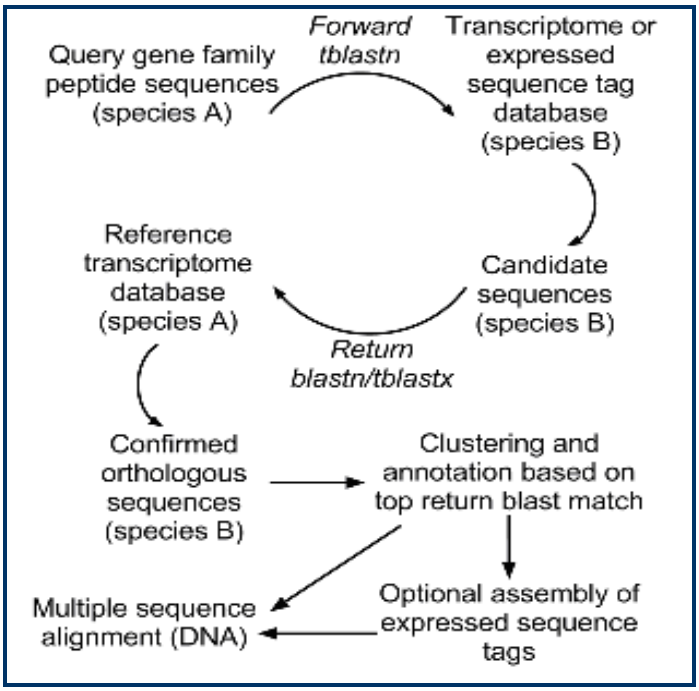

Figure 1: OrthoRBH pipeline schematic diagram. Candidate forward tblastn hits belonging to "species B" are return blasted on the "species A" transcriptome reference sequence. If the return blast best hit is one of the original query sequence, then it is a confirmed homolog and undergoes annotation based upon the name of the best return blast match. If the species B sequence database is comprised of individual expressed sequence tags, these can be assembled into contigs. Newly confirmed orthologs undergo multiple sequence alignment alongside query sequences**.**

Software Requirements:

OrthoRBH functions on the Linux command line and has been developed and tested on the Ubuntu operating system version 12.04. OrthoRBH is dependant on Perl and NCBI BLAST 2.2.25+. Clustal W and CAP3 are optional tools required for multiple sequence alignment and contig assembly steps respectively.

Input options:

OrthoRBH requires a configuration file which specifies the path to the query peptide sequences (species A), the target transcriptome sequence or expressed sequence tag library (species B) and the background transcriptome sequence (species A). The config file further enables selection of either tblastx (sensitive and slow) or blastn (insensitive and fast) algorithm for the return blast search, as well as fine-tuning of the e-value threshold for forward and reverse searches. Options to OrthoRBH allow the user to skip the construction of blast index if this has previously been generated; and enable/disable the contig assembly function depending on the type of target transcript sequence database used. There is further detail on the configuration file and other options in the accompanying documentation.

Output options:

OrthoRBH creates a results directory named as specified in the the config file. The forward blast and return blast results are stored in tabular form. The candidate hits are stored in fasta format. The confirmed ortholog sequences (if any) are organised/clustered and renamed based upon their top return blast match. The confirmed ortholog sequences, either full length cDNA or assembled contigs are collated together with the nucleotide sequences corresponding to the original query sequences in a fasta file to facilitate downstream phylogenetic applications.

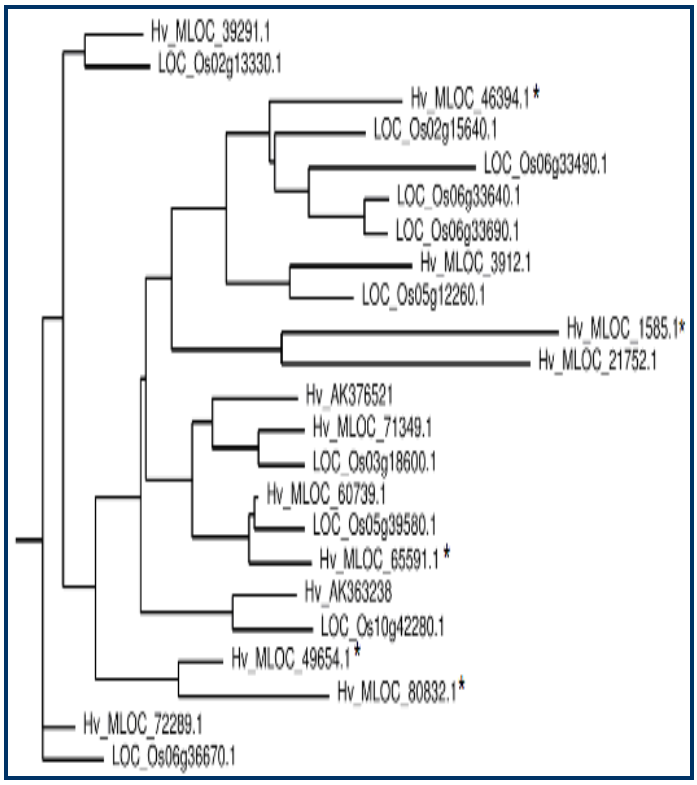

Figure 2: Phylogenetic Neighbour joining tree of PYR genes from O. sativa (labeled "LOC_Os") and newly identified homologs in H. vulgare (labeled "Hv"). Gene tree was generated using inferred peptide sequences submitted to ClustalW. Entries marked with an asterisk are those identified by tblastx return search but not with the less sensitive blastn return search**.**

Case study with a plant gene family:

To demonstrate the efficacy of OrthoRBH, we undertook a search for *PYR/PYL/RCAR* genes (simply termed *PYR* in this work) in the newly-released *Hordeum vulgare* transcriptome database **[6]** using the 10 member *PYR* family of *Oryza sativa* **[7]** as queries **(Figure 2**). When using the more sensitive tblastx return search, 13 *H. vulgare PYR* genes were detected, while 5 of these were not identified when using the less sensitive blastn return search. This demonstrates that tblastx remains the most suitable return search algorithm unless the two species involved are recently diverged. In this instance, OrthoRBH using tblastx took 66 s, while the blastn run took 17 s on a four-core personal computer.

Caveats and future development:

OrthoRBH is not recommended in situations where sequence homology is very low, for instance, identifying homologous sequences between plants and animals. OrthoRBH is not suited for the discovery of non-protein-coding genes. The accuracy of return blast classification is limited by the length of sequences

in the target sequence database. Thus, contig assembly of transcript reads less than 50 bp is not recommended, especially for species pairs with are rich in in-paralogs. Future development of OrthoRBH will be focused upon improving the speed of mining homolog transcripts from large next generation sequencing data sets.

Conclusion:

As the number of fully sequenced genomes grows rapidly, OrthoRBH software will satisfy a need for an automated pipeline to rapidly mine gene families of interest. The software is designed for use by biologists and non-expert bioinformaticians.

References:

[1] Altschul SF *et al. J Mol Biol*. 1990 **215**: 403 [PMID: 2231712]
[2] Ostlund G *et al. Nucleic Acids Res*. 2010 **38**: D196 [PMID: 19892828]
[3] Li L *et al. Genome Res*. 2003 **13**: 2178 [PMID: 12952885]
[4] Larkin MA *et al. Bioinformatics*. 2007 **23**: 2947 [PMID: 17846036]
[5] Huang X & Madan A, *Genome Res*. 1999 **9**: 868 [PMID: 10508846]
[6] The International Barley Genome Sequencing Consortium *et al*. *Nature*. 2012 **491**: 711 [PMID: 23075845]
[7] Umezawa T *et al. Plant Cell Physiol*. 2010 **51:** 1821 [PMID: 20980270]

Edited by P Kangueane
Citation: Ziemann *et al.* Bioinformation 9(5): 267-269 (2013)