

# RoundUp: a repository of orthologs and corresponding evolutionary distances.

I-Hsien Wu, Todd Deluca, Thomas J. Monaghan, Jian Pu, Dennis P. Wall

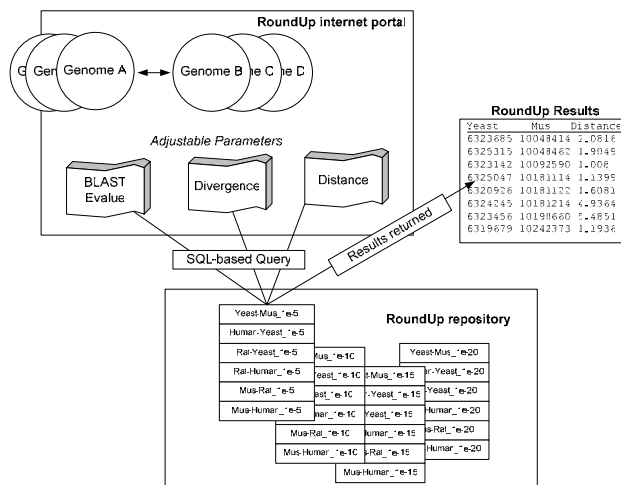
**Keywords:** ortholog detection, orthology database, evolutionary divergence, protein evolution, reciprocal blast.

## 1 Introduction

The ability to accurately detect orthologous or functionally equivalent proteins in different organisms is important to numerous biological research questions, including studies of variables influencing rate protein evolution [1-3], accurate genome annotation [4], and studies of proteins implicated in cancer [5]. The ability to retrieve orthologs from many genomes at the click of a button would help speed such research projects along tremendously.

To meet this challenge, we have pre-computed orthologs for 213 genomes using the reciprocal smallest distance algorithm [6]. This method represents an improvement over approaches that rely on blast hits alone, since it uses global rather than local sequence alignments and evolutionary estimates of distance between sequences rather than blast probability scores, an approach that can often be misleading when trying to determine functional equivalence [7]. These pre-computed sets of orthologs are stored in a publicly accessible database, RoundUp (Figure 1), the most comprehensive of its kind [8, 9].

**Figure 1:** Schematic of RoundUp Ortholog Database. From the website, a user may select two opposing genomes from a list of 213 and set BLAST Evalue, divergence, and distance thresholds as desired. The results are returned in tab-delimited format.



## 2 Software and Files

The reciprocal smallest distance algorithm first uses BLAST [10] to net a list of several possible orthologs between two genomes and then filters this list using a global alignment threshold and evolutionary distance measure. Explicitly, if the alignable region exceeds a predefined fraction of the total length of the protein, a distance is calculated as a maximum likelihood estimate of the number of amino acid substitutions separating the two putatively orthologous protein sequences, given an empirical amino acid substitution rate matrix [11]. The smallest distance is considered to be consistent with a hypothesis of functional equivalency between the two proteins.

The Computational Biology Initiative, Department of Systems Biology, Harvard Medical School, Boston, MA,  
Email: dpwall@hms.harvard.edu

Recovering the true set of orthologs between two lineages will depend on many parameters, including date of divergence between the lineages, rates of gene duplication in either lineage, intensity of molecular selection, and others. To account for such variables that can have a large impact on the size and content of a list of orthologs, and to provide users of RoundUp a certain degree of exploratory power, we adjusted two parameters – BLAST E-value and global pair-wise sequence divergence – when pre-calculating orthology between genomes. Specifically, in our pre-calculations we used four increasingly stringent BLAST scores, 1e-5, 1e-10, 1e-15, and 1e-20, and three increasingly stringent divergence thresholds, 0.8, 0.5, and 0.2. Therefore, for every pair of genomes, our RoundUp repository contains twelve ortholog lists representing all possible combinations of the two variables.

Presently, the RoundUp repository contains orthologs from all possible pair-wise analyses of 213 genomes, comprising 22791 ortholog files. The results are stored in relational form and accessible through a flexible web interface that allows the user to build arbitrarily complex and exploratory queries without needing to know how to construct SQL statements. For example, a user of the RoundUp system can build queries to find all orthologs in all 213 genomes for a given search sequence that are evolutionarily conserved, according to some *a priori* chosen distance threshold. Alternatively, a user may quickly build clusters of orthologous genes for any number of genomes, i.e., build groups of orthologs that exhibit complete transitive closure. To our knowledge, this is the first database of its kind. We expect it will be of value to many molecular biological fields.

## References

- [1.] Fraser, H.B., et al., *Evolutionary rate in the protein interaction network*. Science, 2002. **296**(5568): p. 750-2.
- [2.] Fraser, H.B., D.P. Wall, and A.E. Hirsh, *A simple dependence between protein evolution rate and the number of protein-protein interactions*. BMC Evol Biol, 2003. **3**(1): p. 11.
- [3.] Hirsh, A.E. and H.B. Fraser, *Protein dispensability and rate of evolution*. Nature, 2001. **411**(6841): p. 1046-9.
- [4.] Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome*. Nature, 2002. **420**(6915): p. 520-62.
- [5.] Brown, J.R., et al., *Evolutionary relationships of Aurora kinases: implications for model organism studies and the development of anti-cancer drugs*. BMC Evol Biol, 2004. **4**(1): p. 39.
- [6.] Wall, D.P., H.B. Fraser, and A.E. Hirsh, *Detecting putative orthologs*. Bioinformatics, 2003. **19**(13): p. 1710-1.
- [7.] Koski, L.B. and G.B. Golding, *The closest BLAST hit is often not the nearest neighbor*. J Mol Evol, 2001. **52**(6): p. 540-2.
- [8.] Remm, M., C.E. Storm, and E.L. Sonnhammer, *Automatic clustering of orthologs and in-paralogs from pairwise species comparisons*. J Mol Biol, 2001. **314**(5): p. 1041-52.
- [9.] Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution*. Nucleic Acids Res, 2000. **28**(1): p. 33-6.
- [10.] Altschul, S.F., et al., *Basic local alignment search tool*. J Mol Biol, 1990. **215**(3): p. 403-10.
- [11.] Jones, D.T., W.R. Taylor, and J.M. Thornton, *The rapid generation of mutation data matrices from protein sequences*. Comput Appl Biosci, 1992. **8**(3): p. 275-82.