RoundUp a repository of orthologs and corresponding evolutionary distances

I-Hsien Wu, Jian Pu, Todd Deluca, Thomas J. Monaghan, Dennis P. Wall
The Computational Biology Initiative, Department of Systems Biology, Harvard Medical School

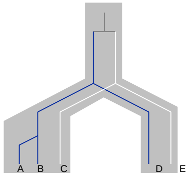
Introduction

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 the rate of protein evolution [1-3], accurate genome annotation [4], and studies of proteins implicated in cancer [5]. In all cases, it is crucial that the orthologous pair be accurate (Figure 1). Moreover, the ability to retrieve orthologs from many genomes at the click of a button would help speed such research projects along by several orders of magnitude.

To meet these challenges, we have pre-computed orthologs for 251 genomes using the reciprocal smallest distance algorithm [6]. This algorithm is an improvement over approaches that rely on BLAST 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 positively misleading when trying to determine functional equivalence [7]. These pre-computed sets of orthologs are stored in a publicly accessible database, RoundUp, the most comprehensive of its kind [8, 9].

Figure 1:

In the case of comparing the evolutionary rates of proteins in the absence of a normalizing molecular clock, such as the rate of synonymous substitutions, estimates of evolutionary rate must be based upon comparisons between sequences that are orthologs (sequences that diverged from each other at the species split), and not paralogs (sequences that diverged at another time). Only if all sequence comparisons share the same time of divergence are protein evolutionary distances expected to be proportional to relative evolutionary rates. For example, in this figure, the orthologous comparisons (A or B with D; C with E) would yield evolutionary distances indicative of the relative rates of protein evolution. By contrast, paralogs comparison of A or B with E would yield an evolutionary distance that would badly overestimate the relative evolutionary rate of these sequences. Procedures that rely on BLAST alone to detect ortholog pairs, such as reciprocal best BLAST hits (rbh) [3,12] are known to be faulty or incomplete [6,7]. To develop RoundUp, we used an improved method for detecting orthologs [6].



Methods

The reciprocal smallest distance algorithm

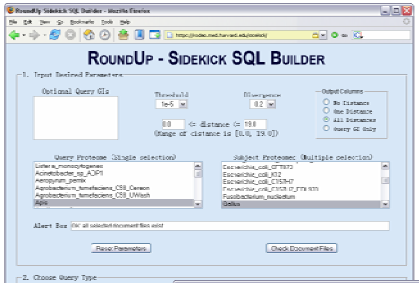
The method (Figure 2) employs BLAST [10] as a first step, starting with a subject genome, *J*, and a protein query sequence, *i*, belonging to genome *I*. A set of hits, *H*, exceeding a predefined significance threshold (we have generally used E-value < 1e-20) is obtained. Then, using Clustalw [13], each protein sequence in *H* is aligned separately with the original query sequence *i*. If the alignable region of the two sequences exceeds a threshold fraction of the alignment's total length (0.8 is our working cutoff), the program PAML [14] is used to obtain a maximum likelihood estimate of the number of amino acid substitutions separating the two protein sequences, given an empirical amino acid substitution rate matrix [11]. The model under which a maximum likelihood estimate is obtained may include variation in evolutionary rate among protein sites, and for more distant comparisons we have generally assumed a gamma distribution of rate across sites, with shape parameter a = 1.53 [15]. Of all sequences in *H* for which an evolutionary distance is estimated, only *j*, the sequence yielding the shortest distance, is retained. This sequence *j* is then used for a reciprocal BLAST against genome *I*, retrieving a set of high scoring hits, *L*. If any hit from *L* is the original query sequence, *i*, the distance between *i* and *j* is retrieved from the set of smallest distances calculated previously. The remaining hits from *L* are then separately aligned with *j* and maximum likelihood distance estimates are calculated for these proteins as described above. If the protein sequence from *L* producing the shortest distance to *j* is the original query sequence, *i*, it is assumed that a true orthologous pair has been found and their evolutionary distance is retained. This algorithm has been tested extensively and is known to produce a more comprehensive and accurate set of orthologs than alternative methods like reciprocal BLAST [6].

The RoundUp data federation and Web Interface

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 in the reciprocal smallest distance algorithm – 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 E-value thresholds, 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.

The pre-computed results are accessible through a web portal to a database federation. The federation, based on WSII technology [16] and our RoundUp data adaptor, allows SQL-based access to the orthology data. The federation is robust: updates and changes to the database repository as new proteomes become available and as older proteomes are updated will not interrupt access. NCBI and other standard warehouses of whole-genomic data are checked weekly for updates and additions, and the RoundUp pre-calculation procedure is run accordingly. Developed using Java technology, the web portal to this database federation is accessible from the Internet using any standard web browser. The web portal allows users without any knowledge of SQL to create arbitrarily complex SQL queries. Presently the portal allows a user to:

1. Discover all orthologs between a specified query genome and any number of other genomes in the database federation.
2. Discover any orthologs between a specified query genome and any number of other genomes in the database federation.
3. Discover all of the transitively closed groups of orthologs among a query genomes and any number of other genomes.
4. Focus results on genes of interest in a query genome by entering GI numbers for those genes.
5. Explore the effects of different parameter settings by choosing one of 12 possible parameter combinations, as well as the evolutionary distance threshold.
6. View the evolutionary distance or orthologs by choosing to view none, one, or all evolutionary distances.


RoundUp Sidekick SQL Builder - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
https://rodeo.med.harvard.edu/sidekick/
Go

RoundUp - Sidekick SQL Builder

1. Input Desired Parameters

Optional Query GIs

Threshold: 1e-5

Divergence: 0.2

0.0 <= distance <= 19.0
(Range of distance is [0.0, 19.0])

Output Columns
○ No Distance
○ One Distance
● All Distances
○ Query GI Only

Query Proteome (Single selection): Listeria_monocytogenes, Acinetobacter_sp_ADP1, Aeropyrum_pernix, Agrobacterium_tumefaciens_C58_Cereon, Agrobacterium_tumefaciens_C58_UWash, Apis

Subject Proteomes (Multiple selection): Escherichia_coli_CFT073, Escherichia_coli_K12, Escherichia_coli_O157H7, Escherichia_coli_O157H7_EDL933, Fusobacterium_nucleatum, Gallus

Alert Box: OK: all selected document files exist

Reset Parameters | Check Document Files

2. Choose Query Type

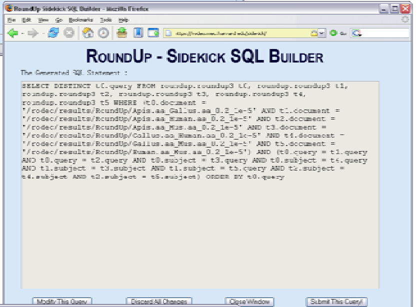
Transitive Closure

Done

Access the RoundUp repository via the Sidekick web application.

Input your query GI numbers, select your parameters, and choose your query type.

Sidekick automatically builds complex SQL statements.

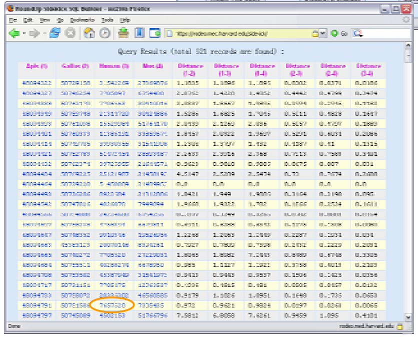

RoundUp Sidekick SQL Builder - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
https://rodeo.med.harvard.edu/sidekick/
Go
rodeo.med.harvard.edu

RoundUp - Sidekick SQL Builder

The Generated SQL Statement :

```
SELECT DISTINCT t0.query FROM roundup.roundup3 t0, roundup.roundup3 t1,
roundup.roundup3 t2, roundup.roundup3 t3, roundup.roundup3 t4,
roundup.roundup3 t5 WHERE (t0.document =
'/rodec/results/RoundUp/Apis.aa_Gallus.aa_0.2_1e-5' AND t1.document =
'/rodec/results/RoundUp/Apis.aa_Human.aa_0.2_1e-5' AND t2.document =
'/rodec/results/RoundUp/Apis.aa_Mus.aa_0.2_1e-5' AND t3.document =
'/rodec/results/RoundUp/Gallus.aa_Human.aa_0.2_1e-5' AND t4.document =
'/rodec/results/RoundUp/Gallus.aa_Mus.aa_0.2_1e-5' AND t5.document =
'/rodec/results/RoundUp/Human.aa_Mus.aa_0.2_1e-5') AND (t0.query = t1.query
AND t0.query = t2.query AND t0.subject = t3.query AND t0.subject = t4.query
AND t1.subject = t3.subject AND t1.subject = t5.query AND t2.subject =
t4.subject AND t2.subject = t5.subject) ORDER BY t0.query
```

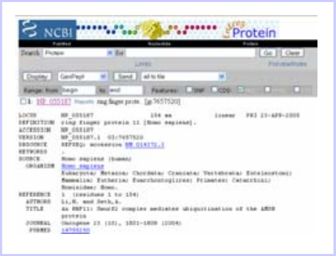
Modify This Query | Discard All Changes | Close Window | Submit This Query!


RoundUp Sidekick SQL Builder - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
https://rodeo.med.harvard.edu/sidekick/
Go
Done
rodeo.med.harvard.edu

Query Results (total 321 records are found) :

Apis (1)	Gallus (2)	Human (3)	Mus (4)	Distance (1-2)	Distance (1-3)	Distance (1-4)	Distance (2-3)	Distance (2-4)	Distance (3-4)
48094322	50729158	31542269	27369876	1.3835	1.3896	1.3895	0.0302	0.0371	0.0186
48094327	50746254	7705897	6754408	2.0762	1.4228	1.4052	0.4452	0.4799	0.3474
48094338	50762170	7706563	30410014	2.0337	1.8667	1.9895	0.2994	0.2945	0.1182
48094349	50759748	21314720	30424886	1.5286	1.6825	1.7045	0.5011	0.4828	0.1647
48094393	50761098	15529984	51764170	2.0439	2.1269	2.036	0.5057	0.4797	0.1889
48094401	50760333	11385191	33359574	1.8457	2.0322	1.9697	0.5291	0.4034	0.2086
48094414	50749785	39930355	31541998	1.2304	1.3797	1.432	0.4387	0.41	0.1315
48094421	50752783	51472454	28593487	2.2633	2.3916	2.368	0.7513	0.7588	0.3483
48094432	50762374	39723065	21655571	0.9623	0.9818	0.9805	0.0675	0.087	0.031
48094434	50769225	25121987	21450193	4.5147	2.5289	2.5474	0.73	0.7674	0.2608
48094464	50729230	51458889	21489953	0.0	0.0	0.0	0.0	0.0	0.0
48094493	50736286	8923304	21312806	1.8421	1.949	1.9085	0.3164	0.3198	0.096
48094542	50747826	4826870	7949094	1.9668	1.9322	1.782	0.1566	0.2534	0.1611
48094566	50748808	24234668	6754256	0.3077	0.3249	0.3265	0.0382	0.0601	0.0164
48094597	50768238	4758854	6670811	0.6311	0.6288	0.6312	0.1275	0.1308	0.0080
48094647	50748352	9910346	19526954	1.2268	1.2063	1.2449	0.2287	0.1934	0.034
48094663	45383123	28070146	8394261	0.7927	0.7809	0.7398	0.2432	0.2229	0.2081
48094665	50740272	7705520	27229031	1.8065	1.8982	7.2443	0.8489	0.6748	0.3305
48094684	50755511	40288274	6678950	0.985	1.1127	1.1922	0.3758	0.4013	0.2103
48094708	50753582	45387949	31541973	0.9418	0.9443	0.9537	0.1506	0.1425	0.0356
48094717	50731151	7705175	12363637	0.4936	0.4815	0.481	0.0805	0.0457	0.0132
48094733	50758072	28835302	46560585	0.9179	1.1026	1.0951	0.1648	0.1735	0.0653
48094791	50751584	7657520	7335435	0.972	0.9621	0.9824	0.0197	0.0263	0.0065
48094797	50745089	4502153	51766796	7.5812	6.8058	7.6261	0.9459	1.095	0.4101

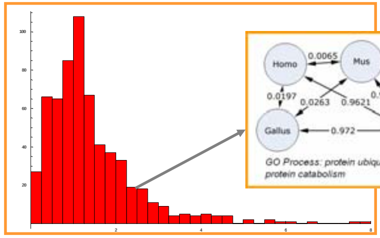
RoundUp then queries our data federation to give you the results.



□ 1: NP_055187 Reports ring finger prote...[gi:7657520]

```
LOCUS       NP_055187                154 aa            linear   PRI 23-APR-2005
DEFINITION  ring finger protein 11 [Homo sapiens].
ACCESSION   NP_055187
VERSION     NP_055187.1  GI:7657520
DBSOURCE    REFSEQ: accession NM_014372.3
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;
            Hominidae; Homo.
REFERENCE   1  (residues 1 to 154)
  AUTHORS   Li,H. and Seth,A.
  TITLE     An RNF11: Smurf2 complex mediates ubiquitination of the AMSH
            protein
  JOURNAL   Oncogene 23 (10), 1801-1808 (2004)
   PUBMED   14755250
```

Figure 3: A histogram of the 621 transitively closed results between Apis, Gallus, Homo, and Mus.



Present status and capabilities

RoundUp 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. Presently, the RoundUp repository contains orthologs from all possible pair-wise analyses of 251 genomes, comprising 376,500 ortholog files. This vast repository offers exciting opportunities to explore the conservation of genes, the evolution of genetic pathways, and the occurrences of gene gains and losses over evolutionary time scales and more.

A user of the RoundUp system can build queries to find all orthologs in all 251 genomes for a given search sequence that are either evolutionarily conserved or labile, 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.

As an example we executed the transitive closure query shown to the left. After 32 seconds the query returned the 621 transitively closed orthologs among Apis, Gallus, Homo, and Mus, with an average distance of 1.4. Figure 3a shows a single gene from this query. The gene is a ring finger protein, known to have ubiquitin-protein ligase activity and to be vital for proper functioning of protein ubiquitination during protein catabolism.

It will be interesting to investigate the functional annotations for each transitively closed gene in this and other lists from different combinations of organisms. Though many may turn out to be predictable for their functional importance, others may shed light on interesting layers of gene conservation not revealed by other methods of investigation.

RoundUp is currently capable of executing queries such as:

What are the transitively closed orthologs among a variety of genomes?

What is the average distance of a set of orthologous sequences?

How does the set of orthologous sequences change when you relax the threshold parameters?

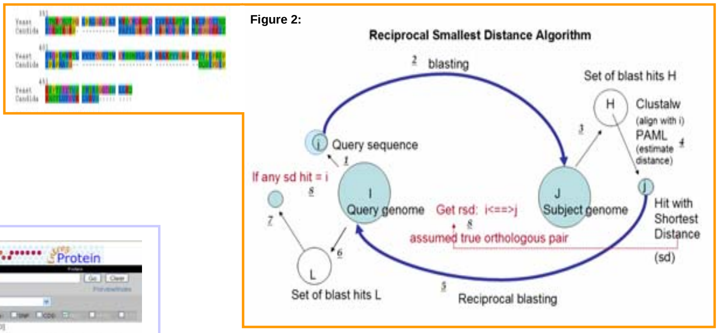


Figure 2:

Future Objectives:

Integrate ortholog data with protein-protein interaction data, expression profile data, and data from Gene Ontology, etc., using the RoundUp database federation.

Extend and update orthology data as genomes are released and updated.

Provide more SQL query templates to answer interesting biological questions.

Extend RoundUp with functions over query result sets. E.g. statistical operations on distances or set operations to compare the effects of different parameters.

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.
[12.] Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, **12**, 962–968.
[13.] Thompson,J.D., Plewniak,F., Thierry,J.C. and Poch,O. DbClustal: Rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.
[14.] Yang,Z. (2000) Phylogenetic Analysis by Maximum Likelihood (PAML). University College London, London.
[15.] Nei,M., Xu,P. and Glazko,G. (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl Acad. Sci. USA*, **98**, 2497–2502.
[16.] http://www.ibm.com/software/data/integration/federation.html

Department of Systems Biology at Harvard Medical School
Computational Biology Initiative