BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 12 2013, pages 1574–1576
doi:10.1093/bioinformatics/btt189

Databases and ontologies

Advance Access publication April 24, 2013

HOMECAT: consensus homologs mapping for interspecific knowledge transfer and functional genomic data integration

Simone Zorzan[1,2,*], Erika Lorenzetto[1], Michele Ettorre[1,2], Valeria Pontelli[1], Carlo Laudanna[2,3] and Mario Buffelli[1,2,4]

[1]Department of Neurological, Neuropsychological, Morphological and Motor Sciences, Section of Physiology, University of Verona, Strada le Grazie 8, 37134 Verona-Italy, [2]Centre for Biomedical Computing, University of Verona, Strada le Grazie 8, 37134 Verona-Italy, [3]Department of Pathology, University of Verona, Strada le Grazie, 8 37134 Verona, Italy and [4]National Institute of Neuroscience-Italy, Verona, Italy

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Comparative studies are encouraged by the fast increase of data availability from the latest high-throughput techniques, in particular from functional genomic studies. Yet, the size of datasets, the challenge of complete orthologs findings and not last, the variety of identification formats, make information integration challenging. With HOMECAT, we aim to facilitate cross-species relationship identification and data mapping, by combining orthology predictions from several publicly available sources, a convenient interface for high-throughput data download and automatic identifier conversion into a Cytoscape plug-in, that provides both an integration with a large set of bioinformatics tools, as well as a user-friendly interface.

Availability: HOMECAT and the Supplementary Materials are freely available at http://www.cbmc.it/homecat/.

Contact: simone.zorzan@univr.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 20, 2012; revised on April 11, 2013; accepted on April 18, 2013

1 INTRODUCTION

The interpretation of large data sets, in particular when obtained from experiments using model systems, greatly benefits from the transfer of knowledge pertaining to phylogenetically related species. The studies on widely used model organisms, such as mouse, fly or rat, yielded a large wealth of information, essential for the understanding of life complexity (Aitman *et al.*, 2011; Loman *et al.*, 2012; Schuster, 2008). The integration of functional genomic and proteomic expression profiles in the modeling of regulatory networks is a valuable approach in biology (Romero *et al.*, 2012).

HOMECAT (homology mapper for enrichment and comparative analysis with translation) is a plug-in for Cytoscape (Shannon *et al.*, 2003) that allows cross-species data comparison and integration of high-throughput data with automatic identifier conversion. Orthology relationships can be difficult to identify, and several approaches exhibit different sensitivity and specificity (Altenhoff and Dessimoz, 2009; Chen *et al.*, 2007; Hulsen *et al.*, 2006). HOMECAT, at present, can combine data from four homology data sources, to attain better specificity and increased sensitivity. BridgeDB (Van Iersel *et al.*, 2010) usage allows to support 30 species and nearly 100 identifiers formats (71 from microarrays platforms). HOMECAT also interfaces Array Express ATLAS (Kapushesky *et al.*, 2012) to download and integrate high-throughput curated data.

2 COMPARATIVE ANALYSES AND INTEGRATION WITH HIGH-TROUGHPUT DATA

In HOMECAT, input species identifiers can be chosen between nodes attributes and are then converted to query the available homology data sources for the identification of orthologs in one or more destination species. Finally, the identifiers of the orthologs are converted to the selected output format. By default Homologene (Wheeler *et al.*, 2007), OMA (Roth *et al.*, 2008), Compara (Vilella *et al.*, 2009) and OrthoMCL (Li *et al.*, 2003) are supported; the former two are more specific and the other two are more sensitive (Altenhoff and Dessimoz, 2009). OMA and Compara servers are queried directly, whereas Homologene (rel. 67) and OrthoMCL (ver. 5) data are accessed through our server in CBMC. When possible, a direct access to the data was used, to always guarantee the most updated results. All necessary conversions are performed by default through a remote BridgeDB server. BridgeDB database can be installed on any machine supporting Java (see BridgeDB website), and can be used by HOMECAT to reduce conversion times, particularly when hundreds of identifiers are processed. After the search phase, the user can decide whether to use only the orthologs confirmed by all the sources or those indicated by any of them, hence, increasing the reliability or the coverage of the results. Orthologs can be used to enrich the input network, to assign their identifiers to specific attributes or to create a novel network of metanodes, preserving original network connectivity.

Each metanode will contain an input species node, along with its orthologs, and can be expanded or collapsed. In the networks of metanodes, the color of the border of each node always represents the input species data, whereas the internal color of the node is always related to the ortholog data. A pie chart summarizes the orthologs data representation on collapsed metanodes (see Fig. 1A for details). When more than one network are

*To whom correspondence should be addressed.

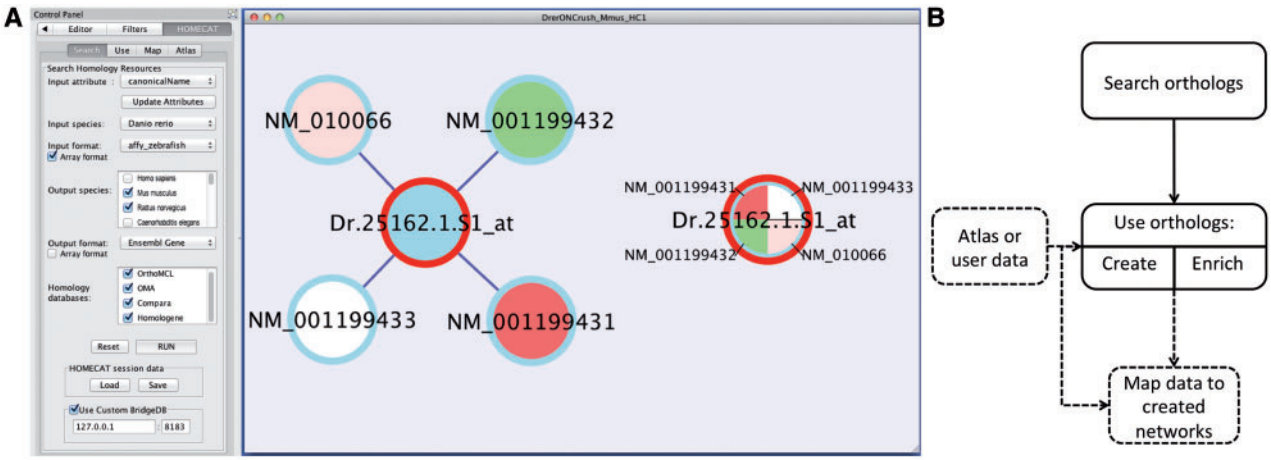

Fig. 1. (**A**) HOMECAT search tab interface and a created network window. Left metanode is expanded with orthologs connected to the input gene. The right metanode represents the same metanode collapsed. Border color always refers to input organism data, whereas node colors are always referred to data for the organisms where orthologs were identified. Data in this figure are from zebrafish and mouse optic nerve crush experiments; see the example in the Supplementary Materials for details on data sources. Red for high, green for low and cyan for no values. Input nodes in expanded metanodes are always cyan in the center, as no other species data can be mapped to them. (**B**) HOMECAT workflow: user searches orthologs and uses results to enrich the input network or create a new network of metanodes. Dashed steps are optional and allow mapping and getting data. Atlas data can be acquired for any Cytoscape network

created, node selections can be extended between networks, facilitating comparisons.

For each ortholog, an attribute summarizes the sources that support its identification. This attribute is useful to adjust the sensitivity attained by combining the results from different orthology sources. The user can filter less strongly supported orthologs and eventually remove them from the network.

External data can be loaded to the resulting network, either from a local file or from ATLAS, by selecting the experiment code and an experimental factor. When data from different orthologs are present, the average is added as an attribute to each metanode. A general scheme of this workflow is depicted in Figure 1B.

As an example, HOMECAT was used to compare microarray data after optic nerve crush in three publicly available datasets from zebrafish, mouse and rat samples. The optic nerve crush is a common model to study the regeneration in the central nervous system. After crush, the optic nerve regenerates in zebrafish, whereas the axonal recovery is absent in mammals. The results highlighted differently regulated genes in fish and mammals, consistent with literature, and showed how the combination of functional genomic and regulatory data analysis can contribute to the identification of putative key factors in comparative biological studies (see the manual in the Supplementary Materials).

3 FEATURES AND EXTENSIBILITY

Orthologs resulting from HOMECAT multiple queries can be saved and loaded in '.hcd' files. Data mapped to attributes can be saved along with Cytoscape networks. A programming interface allows the development of Java classes called *components*, to combine additional homology data sources that are automatically loaded on HOMECAT launch (Supplementary Materials).

ACKNOWLEDGEMENTS

The authors thank Giovanni Scardoni for the suggestions in the plug-in development; Claudio Pascale and Alessio Azzoni for the early development of the plug-in prototypes; Adrian Altenhoff for the support in the development of OMA component.

Funding: Fondazione Cariverona (to CBMC); Fondazione Cariverona and Associazione Italiana per la Ricerca sul Cancro (to C.L.).

Conflict of Interest: none declared.

REFERENCES

Aitman,T.J. *et al.* (2011) The future of model organisms in human disease research. *Nat. Rev. Genet.*, **12**, 575–582.

Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.

Chen,F. *et al.* (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**, e383.

Hulsen,T. *et al.* (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.*, **7**, R31.

Kapushesky, *et al.* (2012) Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **40**, D1077–D1081.

Li,L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

Loman,N.J. *et al.* (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.*, **10**, 599–606.

Romero,I.G. *et al.* (2012) Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.*, **13**, 505–516.

Roth,A.C.J. *et al.* (2008) Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, **9**, 518.

Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.
Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
Van Iersel,M.P. *et al.* (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, **11**, 5.
Vilella,A.J. *et al.* (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
Wheeler,D.L. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.