**Databases and ontologies**

**COMPARE, a multi-organism system for cross-species data comparison and transfer of information.**

David Salgado†, Gregory Gimenez†, François Coulier2, Christophe Marcelle1 *

1Developmental Biology Institute of Marseille Luminy (IBDML). CNRS UMR 6216. Université de la Méditerranée. Campus de Luminy, case 907. 13288 Marseille
2Laboratory of Molecular Oncology, Marseille Cancer Institute, UMR599, 27 Bd. Le Rioule, 13009 Marseille, France.

Associate Editor: Prof. John Quackenbush

**ABSTRACT**

Motivation: COMPARE is a multi-organism web-based resource system designed to easily retrieve, correlate and interpret data across species. The COMPARE interface provides access to a wide array of information including genomic structure, expression data, annotations, pathways, and literature links for human and three widely studied animal models: zebrafish, Drosophila, and mouse. A consensus ortholog-finding pipeline combining several ortholog prediction methods allows accurate comparisons of data across species and has been utilized to transfer information from well studied organisms to more poorly annotated ones.

Contact: compare@ibdml.univ-mrs.fr

1 INTRODUCTION

Biological functions are regulated by complex and often poorly characterized networks of genes whose dysfunctions can cause severe diseases in human and in animals. The biological mechanisms involved in many biological functions have been amazingly well conserved during evolution. As a consequence, important insights into the cellular and molecular mechanisms regulating normal and aberrant development, function and repair in higher vertebrates have been gained from analyses performed in sometimes distantly related organisms. To understand the mechanisms regulating biological functions, researchers have generated large sets of diverse types of data in various species. However, these data are often stored in heterogeneous formats and scattered over a multitude of species-specific (e.g. ZFIN, FlyBase, for zebrafish and Drosophila, respectively) and specialized databases (Ensembl, PLATCOM (Choi et al., 2005), ZooDDD (Chen et al., 2006) for genomic or EST-based expression data) that are largely non-interoperable. Such situation makes cross-species comparison and analyses difficult and time-consuming, in particular when large data sets need to be evaluated. Here, we present COMPARE, a resource system that integrates diverse data from human and three widely studied animal models: zebrafish, Drosophila, and mouse. By providing access within a single integrated system to the most relevant information for each species, COMPARE represents a unique tool that not only greatly facilitates interspecies comparisons, but more importantly provides novel information generated through orthology predictions.

2 SYSTEM ARCHITECTURE

COMPARE is a modular system comprising three applications organized around a gene-centered database (COMPARE-DB) built on a PostgreSQL database management system. This system is adapted from NISEED, the generic version of the ascidian database Aniseed (http://aniseed-ibdm.univ-mrs.fr; Tassy et al., in preparation). The first application is a Public-web interface, the COMPARE-WebSite, developed in PHP, Javascript, and PERL-CGI. Through a system, inspired from Aniseed, that allows combinations of diverse biological features, users are able to gradually refine their queries. A COMPARE-Manager has been developed as a restricted access website (using PHP, Perl and BioPerl) that allows a manager to add and update information from public data sources through API (Application Programming Interface), Database connection, web services and flatfiles. The architecture on which COMPARE is built allows easy addition of new data sources, or new animal models through the Manager. Finally, experts can evaluate and amend experimental data via a COMPARE-Curator.

3 RETRIEVING DATA FROM PUBLIC SOURCES

Specific databases are either directly included in the COMPARE-DB or linked as remote sources.

The genomic data and protein domains available in COMPARE are derived from ENSEMBL (release 41 for mouse, human and Drosophila; release 46 for zebrafish), and have been retrieved using the ENSEMBL API (Hubbard, et al., 2007).

Gene Ontology (GO) annotations have been included from three distinct sources. The first source is the GO annotations from ENSEMBL core database. Additional GO annotations were obtained from species specific databases (ZFIN for zebrafish, MGI for mouse and FlyBase for Drosophila). Last, we enriched this information with GO annotations from the GOA consortium. For each GO annotation, we display its original source and associated evidence code.
To provide information on molecular networks, we have integrated molecular pathways from the KEGG database and putative interactors from the iHOP literature-derived information server.

For each gene, the relevant publications are also collected using the gene2pubmed file produced by the NCBI.

We have included in situ hybridization gene expression data, that are available for 3 animal models. These data originated from ZFIN for zebrafish, MGI and EMAGE for mouse and BDGP for Drosophila. Each in situ hybridization experiment is linked to a specific gene through its probe, a particular developmental stage and one or many anatomical parts. To describe the anatomical parts and their relationships, we used the standard species-specific developmental anatomical ontology (OBO http://obofoundry.org).

The probe-to-gene mapping was done using the gene2unigene file in the same way that was done for the literature data. In addition to in situ hybridization data, Unigene data have been integrated. General information on the tissue type in which a gene is expressed is derived from Unigene.

We integrated and correlated these data for 4 species: Drosophila, zebrafish, mouse and human. A complete gene report in the COMPARE-Website presents biological data from the sources described above, together with a GenomeBrowser visualisation tool. Users can track the origin of the information found in COMPARE, since each information is tagged with its source.

4 ORTHOLOGY PREDICTION

Orthology prediction has become a powerful tool widely used in comparative genomics. Currently, there are many different software available but their accuracy varies. We decided to integrate four different orthology prediction methods in COMPARE: OrthoMCL (Li et al., 2003) and InParanoid are locally computed and they have been shown to display the best overall balance between sensitivity and specificity; ENSEMBL Compara (calculated for each Ensembl release) and TreeFam. Thus, users can readily evaluate the reliability of the orthology prediction: a gene identified as a putative ortholog by only one method might be more disputable than one consensually detected by two, three or four methods. It is possible to filter the orthology prediction results accordingly. Importantly, combining ortholog prediction methods has allowed the transfer of information to poorly annotated organisms: for instance, with at least one method predicting a gene as a putative ortholog, we have increased the GO coverage of the zebrafish genome by 40%.

5 QUERY TOOLS

Users can query COMPARE via several tools: i) A Genome Browser (Stein et al., 2002) that displays a transcript in its chromosomal environment; ii) Molecule search: Users can search for genes or ESTs using boolean connectors (AND/OR) either in a “perfect match” or in a fuzzy manner, using identifiers, cross-references or synonyms.; iii) In situ search: In situ hybridizations are retrieved through gene searches or by browsing through the developmental anatomical ontology trees. In the latter case, users can choose to either retrieve all in situ hybridizations at a particular developmental stage or to search for gene expressed in specific anatomical parts; iv) Pathway search: This tool retrieves a list of genes that are involved in a pathway of interest. User can combine and retrieve a list of genes participating in several pathways; v) Gene ontology search: to retrieve genes that have specific GO terms in one or several species; vi) Protein domains: to search in one or several species for genes that have specific protein domain; vii) A BLAST tool is available for sequence similarity search. All these tools are independent but it is possible to chain them to refine the searches; viii) A Batch query tool allows to retrieve data for a list of genes and for their orthologues.

Finally, help sections and tutorial short movies provide the necessary information for newcomers to the COMPARE system.

6 CONCLUSION

COMPARE is a unique resource system that associates several information from public data sources and that aims to facilitate interspecies comparisons and hypotheses building. Moreover, novel information has been generated in COMPARE by orthology prediction. The structure of COMPARE constitutes the core on which specific projects (e.g. tissue- or disease-centered) can be built. Hence, we are developing a muscle-specific interspecies database (MyoBase) that focuses on muscle genes and myopathies by the integration of muscle-specific data into the system. Three additional species will be integrated shortly into COMPARE, chick, Ciona and C. elegans. A data loader will be developed to allow users to upload their own experimental data into the system.

ACKNOWLEDGEMENTS

We thank Drs Pascal Hingamp and Bernard Jacq for helpful comments on the manuscript and Drs Olivier Tassy, Patrick Lemaire and Bernard and Christine Thise for precious advices during the course of this study. We acknowledge the help of Magali Contensin and Jean-Francois Guillemot at the CRFB, hosting the project. This study was funded by a grant from the EU (MYORES Network of Excellence, under the FP6 framework)

REFERENCES


COMPARE, a multi-organism system for cross-species data comparison and transfer of information.


