BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 16 2007, pages 2180–2182
doi:10.1093/bioinformatics/btm295

Genome analysis

OMA Browser—Exploring orthologous relations across 352 complete genomes

Adrian Schneider*, Christophe Dessimoz and Gaston H. Gonnet

Institute of Computational Science, ETH Zurich, Switzerland

Received on February 8, 2007; revised and accepted on May 25, 2007

Advance Access publication June 1, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Inference of the evolutionary relation between proteins, in particular the identification of orthologs, is a central problem in comparative genomics. Several large-scale efforts with various methodologies and scope tackle this problem, including OMA (the Orthologous MAtrix project).

Results: Based on the results of the OMA project, we introduce here the OMA Browser, a web-based tool allowing the exploration of orthologous relations over 352 complete genomes. Orthologs can be viewed as groups across species, but also at the level of sequence pairs, allowing the distinction among one-to-one, one-to-many and many-to-many orthologs.

Availability: http://omabrowser.org

Contact: schneadr@inf.ethz.ch

1 BACKGROUND

Accurate orthology assignment is a prerequisite for numerous bioinformatics analyses, including the construction of phylogenetic trees, function inference of novel proteins, identification of conserved regions, detection of horizontal gene transfer and transcription binding site prediction. Evidence for the importance of these tasks can be found in the growing number of orthology assignment projects developed in recent years.

The COG method (Tatusov *et al.*, 1997) was the first one to extend the systematic orthology search beyond the relatively simple reciprocally best matches approach. This algorithm, originally applied mostly on bacterial genomes, was later extended and applied to eukaryotic genomes: the KOG database (Tatusov *et al.*, 2003) and the EGO/TOGA project (Lee *et al.*, 2002)'. Following these approaches, more sophisticated orthology determination algorithms were proposed in recent years, most notably InParanoid/MultiParanoid (Alexeyenko *et al.*, 2006; Remm *et al.*, 2001), OrthoMCL (Li *et al.*, 2003), KEGG Orthology (Kanehisa *et al.*, 2004) and Roundup (DeLuca *et al.*, 2006). Other projects, such as IMG (Markowitz *et al.*, 2006) and MicrobesOnline (Alm *et al.*, 2005) employ more basic algorithms, but distinguish themselves by a large number of analyzed sequences.

The OMA project (Dessimoz *et al.*, 2005) is a massive cross-comparison of complete genomes to identify the evolutionary relation between any pair of proteins. The main features of OMA are the large number of genomes from all kingdoms of life, the strict verification of orthology assignments and the determination of the phylogenetic relationship between any two proteins. These major differences to other orthologs projects will be explained in the following subsections. All improvements made to the algorithm since the introductionary paper (Dessimoz *et al.*, 2005) are described on the OMA Browser web page. Most notable are the use of distances instead of scores and statistically sounder measures for establishing the set of potential orthologs in the early stages of the algorithm.

1.1 Scope and automation

The analysis so far has been performed on 352 genomes from all kingdoms of life (44 eukaryotes, 282 bacteria and 26 archaea). Since summer 2004, over 300 CPUs have performed more than 375 years of computation to align 1.6 million proteins, resulting in 1.3 trillion (10^{12}) alignments. In the OMA release from January 2007, 206,326 groups of orthologs are identified, thus making OMA one of the largest orthology projects.

Dealing with such large amounts of data requires a high degree of automation and integrated quality checks. Unlike comparable projects such as the COGs database or KEGG Orthology, OMA does not rely on human intervention. Once a genome database is integrated into the local databases, the process is fully automated.

1.2 Strictness

Orthology inference can be a difficult problem, for instance when gene families went through intense expansion and reduction or when duplication and speciation events occurred at close time intervals. In OMA, we have a strict approach across the entire procedure, such as full dynamic programming alignments instead of Blast or the systematic use of evolutionary distances with confidence intervals. When lacking discriminating information, we favor false negative (missing orthologous relations) over false positive (erroneous orthology assignment). A notable and in this extent unique feature is the systematic verification of every putative pair

*To whom correspondence should be addressed.

of orthologs by an exhaustive search for 'witnesses of non-orthology' in third-party genomes (Dessimoz *et al*., 2006).

1.3 Orthology inference at the level of pairs

Orthology is not necessarily a one-to-one relation and also not always transitive, therefore any clustering approach will have its limits. Although OMA initially focused on groups of orthologs, the OMA Browser also displays information about pairwise orthologous relations, the 'verified stable pairs' in Dessimoz *et al*., (2005), which can be categorized as follows:

- one-to-one orthologs: in both species, there is only one corresponding ortholog.
- one-to-many or many-to-many orthologs: in at least one of the two species, the gene duplicated after speciation.
- paralogs: the two proteins arose through gene duplication, not speciation.

As far as we know, the only other project providing orthology inference at the level of pairs is Ensembl (Hubbard *et al*., 2005).

Nevertheless, for many analyses (particularly for phylogenetics), it is convenient to have groups of orthologs with at most one protein from each species and every pair inside the group being a pair of orthologs. Therefore, the OMA Browser also provides a group-centric view. OMA groups are cliques of orthologs chosen in a way such that the alignment scores are maximized. The clique requirement ensures that the above stated properties of an orthologous group are fulfilled.

2 OMA BROWSER

2.1 Implementation

The OMA browser is a web application using as basis *Darwin* (Gonnet *et al*., 2000), a software package for bioinformatics developed within our group. Benefits include efficient data-structures for biological sequences, and a large library of functions for bioinformatics analyses. While most data is pre-computed, some computations are performed in real time or on user request.

2.2 Protein-centric view

Proteins of interest can be accessed through a search interface. Searches can be conducted on identifiers, accession numbers or descriptions. Furthermore, we provide a fast sequence search on the 1.6 million protein sequences that takes as input any sequence substring.

The protein view provides cross-references, mainly to GenBank, Ensembl or Swiss-Prot/UniProt (for more than 92% of the proteins a link to another databases can be provided), annotations as found in the source database, chromosome/locus information as well as links to the different types of orthologs and the corresponding OMA group.

2.3 OMA group-centric view

The OMA group detail view contains several 'tabs'. The main view is a list of all proteins in the group with an identifier and a description, providing the complete information of the protein family. Since the OMA project itself is automated, no additional annotation by hand is performed. A short description of the OMA group is inferred from the available sequence descriptions.

A multiple sequence alignment of all proteins of a given group can be requested and will be displayed after computation is completed.

Related groups can be explored by two options: 'Close groups' are OMA groups in which at least one protein is orthologous to a protein in the current data-set, while 'Phyletic profile' lists groups having similar patterns of presence/absence across species. This is a possible way of identifying interacting protein families, where either all members must be present in a genome or none of them are required. Whenever available, Gene Ontology (Harris *et al*., 2004) annotations of the different proteins of the group can also be compared and provide additional indication about the functionality of the proteins.

2.4 Data export and integration

All the data can also be downloaded from the browser web page in numerous formats: FASTA, text (list of IDs), *Darwin* database (SGML format) or in a COG-compatible format. These files are available for all OMA groups in one file or for each group individually.

The OMA Browser offers also a SOAP-based application programming interface (API), allowing for the integration of the OMA data into applications or web services. Funding to pay the open access charges was provided by ETH Zurich.

Conflict of Interest: none declared.

REFERENCES

Alexeyenko,A. *et al*. (2006) Automatic clustering of orthologs an inparalogs shared by multiple genomes. *Bioinformatics*, **22**, e9–e15.

Alm,E.J. *et al*. (2005) The MicrobesOnline web site for comparative genomics. *Genome Res*., **15**, 1015–1022.

DeLuca,T.F. *et al*. (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, **22** (16), 2044–2046.

Dessimoz,C. *et al*. (2006) Detecting non-orthology in the COG database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res*., **34**, 3309–3316.

Dessimoz,C. *et al*. (2005) OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In: McLysath,A. and Huson,D.H. (eds) In *RECOMB 2005 Workshop on Comparative Genomics*. Vol. LNBI 3678, of *Lecture Notes in Bioinformatics*. Berlin/Heidelberg: Springer-Verlag, pp. 61–72.

Gonnet,G.H. *et al*. (2000) Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics*, **16**, 101–103.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32** (Database issue), 258–261.

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33** (Suppl.1), D447–D453.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32** (Database issue), 277–280.

Lee,Y. *et al*. (2002) Cross-referencing eukaryotic genomes: TIGR orthologous gene alignments (TOGA). *Genome Res*., **12**, 493–502.
Li,L. *et al*. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*., **13**, 2178–2189.
Markowitz,V.M. *et al*. (2006) The integrated microbial genomes (IMG) system. *Nucl. Acids Res*., **34** (Database issue), D334–338.
Remm,M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol. Biol*., **314**, 1041–1052.
Tatusov,R.L. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, http://www.biomedcentral.com/1471–2105/4/41.
Tatusov,R.L. *et al*. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.