BIOINFORMATICS LETTER TO THE EDITOR Vol. 28 no. 6 2012, pages 900–904
doi:10.1093/bioinformatics/bts050

Genome analysis

Advance Access publication February 12, 2012

Toward community standards in the quest for orthologs

Christophe Dessimoz[1,*], Toni Gabaldón[2], David S. Roos[3], Erik L. L. Sonnhammer[4], Javier Herrero[5]; and the Quest for Orthologs Consortium[†]

[1]ETH Zurich and Swiss Institute of Bioinformatics, Universitätstrasse 6, 8092 Zürich, Switzerland, [2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), and UPF. Dr. Aiguader, 88, 08003 Barcelona, Spain, [3]Department of Biology and Penn Genome Frontiers Institute, University of Pennsylvania, Philadelphia, PA 19104, USA, [4]Stockholm Bioinformatics Center, Swedish eScience Research Center, Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Box 1031, SE-17121 Solna, Sweden, [5]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Associate Editor: Alfonso Valencia

ABSTRACT

The identification of orthologs—genes pairs descended from a common ancestor through speciation, rather than duplication—has emerged as an essential component of many bioinformatics applications, ranging from the annotation of new genomes to experimental target prioritization. Yet, the development and application of orthology inference methods is hampered by the lack of consensus on source proteomes, file formats and benchmarks. The second 'Quest for Orthologs' meeting brought together stakeholders from various communities to address these challenges. We report on achievements and outcomes of this meeting, focusing on topics of particular relevance to the research community at large. The Quest for Orthologs consortium is an open community that welcomes contributions from all researchers interested in orthology research and applications.

Contact: dessimoz@ebi.ac.uk

Received on September 27, 2011; revised on December 2, 2011; accepted on January 22, 2012

1 INTRODUCTION

The concepts of orthology and paralogy are central to comparative genomics. These terms were coined more than four decades ago (Fitch, 1970) to distinguish between two classes of gene homology: those descended from a common ancestor by virtue of a speciation event (orthologs) versus those that diverged by gene duplication (paralogs). This distinction permits accurate description of the complex evolutionary relationships within gene families including members distributed across multiple species. Detection of orthology and paralogy has become an essential component of diverse applications, including the reconstruction of evolutionary relationships across species (reviewed in Delsuc *et al.*, 2005), inference of functional gene properties (e.g. Chen and Jeong, 2000; Hofmann, 1998; Tatusov *et al.*, 1997), and identification and testing of proposed mechanisms of genome evolution (e.g. Mushegian and Koonin, 1996; Tatusov *et al.*, 1997). In today's context, with the number of fully sequenced genomes growing by the day, accurate and efficient inference of orthology has become an imperative. A plethora of computational methods have been developed for inferring orthologous relationships, many of which provide their predictions in form of web-accessible databases (reviewed in Alexeyenko *et al.*, 2006; Gabaldón, 2008; Koonin, 2005; Kristensen *et al.*, 2011).

In 2009, the first Quest for Orthologs meeting was organized to bring together scientists working in the fields of orthology inference, genome annotation and genome evolution to exchange ideas, tackle common challenges, aiming at removing barriers and redundancy (Gabaldón *et al.*, 2009). The main objectives identified were concerted effort toward standardized formats, datasets and benchmarks, and establishment of continuous communication channels including a mailing list, a website and a regular meeting.

Following the first Quest for Ortholog meeting in 2009, a second meeting was held in June 2011, bringing together 45 participants from 27 different institutions on 3 continents, representing >20 orthology databases (http://questfororthologs.org/orthology_databases). The meeting was structured to include plenary sessions devoted to topics of general interest (reference datasets, orthology detection methodology, practical applications of orthology), and additional discussions focusing on benchmarking, standardized formats, alternative transcripts, ncRNA orthology, etc. In this letter, we summarize the discussions and specific outcomes of the meeting, as well as some of the most important achievements of the Quest for Orthologs community in the past 2 years.

2 DEFINITIONS AND EVOLUTIONARY MODELS

Orthology finds application in multiple, diverse research areas. Depending on the context, the reasons for identifying orthologous genes can vary considerably, sometimes driving the use of subtly differing definitions of orthology and its extension to groups of genes. Brigitte Boeckmann (Swiss Inst Bioinformatics, Geneva, Switzerland) and Christophe Dessimoz (ETH Zürich, Switzerland) reviewed the definitions and objectives of orthologous groups within a unifying framework and discussed the implications of these differences for the interpretation and benchmarking of ortholog databases (Boeckmann *et al.*, 2011). The need for clear evolutionary definitions is particularly acute for multidomain proteins, as their underlying coding sequences often have distinct, and even conflicting, evolutionary histories. In an attempt to salvage

*To whom correspondence should be addressed.

[†]The complete list of members of the Quest for Orthologs Consortium is provided in the Acknowledgement section.

the gene as the fundamental evolutionary unit, Dannie Durand (Carnegie Mellon University, Pittsburgh, USA) proposed a model of gene homology based on the genomic locus, not the constitutive nucleotides of the gene (Song *et al.*, 2008).

3 DEBATING THE 'ORTHOLOG CONJECTURE'

The 'ortholog conjecture'—that at a similar degree of sequence divergence, orthologs are generally more conserved in function than paralogs—has been a prevailing paradigm, originally supported by theory rather than empirical studies. At the previous Quest for Orthologs meeting, Bill Pearson (University Virginia, Charlottesville, USA) questioned the ortholog conjecture and contended that the sequence similarity be the primary determinant of functional conservation (Gabaldón *et al.*, 2009). Several studies have now been undertaken to compare the properties of orthologs versus paralogs, and generally appear to support the importance of distinguishing orthologs from paralogs.

Erik Sonnhammer (Stockholm University, Sweden) reported significant support for the ortholog conjecture based on conserved domain architecture (Forslund *et al.*, 2011) and intron positions (Henricson *et al.*, 2010). David Roos (University of Pennsylvania, Philadelphia, USA) showed that protein structure is significantly more conserved for orthologs than for paralogs, particularly within protein active sites. Indeed, it is even possible to quantify the importance of orthology, in terms of sequence conservation or RMSD, for structural modeling (Peterson *et al.*, 2009). Toni Gabaldón (Center for Genomic Regulation, Barcelona, Spain) and colleagues found that human–mouse orthologs exhibit more conserved tissue expression than paralogs of a similar age (Huerta-Cepas *et al.*, 2011). Similarly, Klaas Vandepoele (Ghent University, Belgium) reported that for 77% of orthologs between *Arabidopsis* and rice, the expression patterns were more highly conserved than the background distribution, and that expression patterns can also be used to tease out functional similarity even among in-paralogs (Movahedi *et al.*, 2011).

In other tests, however, orthologs were not found to be functionally more conserved than paralogs. Just days before the meeting, Nehrt *et al.* (2011) reported that Gene Ontology (GO) functional annotations (du Plessis *et al.*, 2011) may be less similar among orthologs than among paralogs, and that human–mouse co-expression data across tissues argues against the ortholog conjecture. Discussion at the meeting noted an inherent bias favoring conservation between homologs in the same species, which may inflate the scores of paralogs. Furthermore, using correlation coefficients as a measure of gene expression conservation may also cause problems (Pereira *et al.*, 2009). Overall, this discussion suggests that the debate remains far from being settled.

4 INNOVATIONS IN ORTHOLOGY INFERENCE: INCREMENTAL METHODS AND META-METHODS

Much of the meeting focused on innovations in orthology inference. One trend involves the application of incremental methods, minimizing the need to recompute results as new datasets are added. Ikuo Uchiyama (National Institute for Basic Biology, Okazaki, Japan) described how the Microbial Genome Database (MBGD) uses such an approach to cope with new genomes, and also to identify orthologs in metagenomic samples (Uchiyama *et al.*, 2010). Likewise, the most recent release of the OrthoMCL database permits new genes (and even entire genomes) to be assigned to putative ortholog groups (Chen *et al.*, 2006). Ingo Ebersberger (CIBIV, Vienna, Austria) showed how an incremental approach based on hidden Markov models can be used to identify orthologs in EST libraries, which typically only cover a fraction of all genes (Ebersberger *et al.*, 2009), and Radek Szklarczyk (2012) introduced a new profile-based iterative procedures that pushes the boundaries of reliable homology detection and helps identify disease genes in human.

Another trend involves the application of meta-methods to integrate predictions from multiple datasets, combining their strengths so as to outperform any single underlying method. Michiel Van Bel (Ghent University, Belgium) presented an ensemble method intended to detect orthologs in plant species combining different orthology inference methods—a notorious challenge due to extensive whole genome duplication and paleopolyploidy. This concept lies at the heart of the PLAZA database (Proost *et al.*, 2009). Michael S. Livstone (Princeton University, USA) described how the P-POD database (Heinicke *et al.*, 2007) enables users to compare orthology and paralogy predictions from multiple homology inference methods on 12 reference genomes from the Gene Ontology Consortium (Reference Genome Group of the Gene Ontology Consortium, 2009). With MetaPhOrs, Gabaldón showed that combining the orthologs inferred from several large-scale phylogenetic resources is not only meaningful to increase the total number of predictions, but also to assess the accuracy based on the consistency across different sources (Pryszcz *et al.*, 2011).

5 STANDARDS AND BENCHMARKING

A primary motivation for this meeting has been to establish standards for efficient data exchange in the orthology community. Until now, virtually every ortholog database has used a different format, posing a major impediment for consumers of orthology data, including annotators and for comparative genomicists. Likewise, the source data for orthology analysis (proteomes) has used a variety of formats (mostly *ad hoc* variations of the Fasta format). To resolve these issues, a working group has developed XML-based formats for both sequence and orthology data (OrthoXML and SeqXML, respectively) (Schmitt *et al.*, 2011). These formats were endorsed by meeting participants, representing many orthology databases, and by the reference proteome project. Documentation and tools are available at http://OrthoXML.org and http://SeqXML.org.

Following on from suggestions at the previous meeting, the Quest for Orthologs 'Reference Proteomes' serves as a common dataset to compare orthology inference methods. Eleanor Stanley (EBI, Hinxton, UK) gave an overview of UniProt's commitment to curate this dataset. Meeting participants suggested that an annual release schedule would be appropriate, and should ensure that most methods are applied to a common and reasonably current dataset. Although driven by the need to benchmark ortholog detection algorithms against a common dataset, we anticipate that the reference proteome project will be useful beyond the orthology prediction community. For example, UniProt curators are eager to test how different ortholog predictions against a consistent dataset can be used to facilitate protein annotation. Complementing the reference proteome project, Raja Mazumder (Georgetown University, Washington,

USA) presented an automated approach to identify representative proteomes—relatively small subsets of all proteomes that capture most of the information available (Chen *et al.*, 2011).

The availability of standardized datasets should significantly ease the challenge of sourcing genomes faced by all providers of ortholog detection, and holds great promise for orthology inference benchmarking. Indeed, previous benchmarking studies have been forced to evaluate orthology predictions based on inconsistent datasets (Altenhoff and Dessimoz, 2009; Boeckmann *et al.*, 2011; Hulsen *et al.*, 2006; Trachana *et al.*, 2011), or have been limited to comparatively small datasets analyzed only by methods available as stand-alone programs (Chen *et al.*, 2007; Salichos and Rokas, 2011). Leveraging the Reference Proteomes, Adrian Altenhoff (ETH, Zürich, Switzerland) presented a web server prototype for orthology benchmarking. The service gathers predictions submitted by ortholog providers and runs a battery of tests, such as an assessment of how well the predictions satisfy a standard definition of orthology (Fitch, 1970), and a test assessing accuracy in predicting GO function annotations (du Plessis *et al.*, 2011).

6 FUNCTIONAL PREDICTIONS

One of the chief benefits of ortholog group assignment is the potential for inferring putative function—particularly as new sequencing methodologies make it increasingly possible to assemble genomes and define genes from species where experimental data is lacking. Such computational inference can be risky, however, as the accuracy of existing annotations is often unknown, particularly for electronically assigned annotations, leading to rampant *in silico* propagation of errors (Gilks *et al.*, 2002). Paul Thomas (USC, Los Angeles, USA) outlined activities of the Gene Ontology (GO) Reference Genomes Project (Reference Genome Group of the Gene Ontology Consortium, 2009), and described a pilot project assigning GO terms to internal nodes of a reference tree (Gaudet *et al.*, 2011). Incorporating a concept of evolutionary breadth (and confidence) into the annotation process would greatly enhance the specificity of orthology-based inference. Nives Škunca (ETH, Zürich, Switzerland) reported an innovative effort to estimate the quality of electronic GO annotations, by tracking changes in stability, coverage and specificity over time. This study suggests a strategy for identifying high confidence electronic annotations that can be relied upon for transitive inference. The availability of a web-based platform for comparing the performance of orthology detection methods (see above) should greatly facilitate the assessment of functional prediction performance. In addition, the development of a curated catalog of ortholog genes with similar function, using experimental data, such as RNAi, expression data or mutant phenotype, would be a useful resource and could improve functional prediction.

7 ADDITIONAL TOPICS

Homology prediction based on similarity is a prerequisite for many orthology prediction methods, and a workshop was held to discuss current approaches and upcoming challenges in assessing sequence similarity. Much discussion was devoted to the need for more realistic models of sequence evolution, which would enable the proper assessment of what level of similarity is expected for two evolutionary related sequences. Tina Koestler (CIBIV, Vienna, Austria) and Jean-Baka Domelevo (LIRMM, Montpellier, France) presented profile-based models of evolution, taking into account particularities of functional or structural regions of protein sequences. Further discussions stressed the necessity of elucidating the mode of evolution of multidomain proteins, particularly in the context of domain rearrangements. In a different take on homology inference, Vincent Miele (LBBE, Lyon, France) reported new methodology to identify robust homologous groups from the structure of similarity networks.

Orthology inference has been traditionally focused on the study of protein coding genes, but there is increasing interest in applying similar analyses to non-coding RNAs (ncRNAs). For example, both Ensembl (Flicek *et al.*, 2011) and miROrtho (Gerlach *et al.*, 2009) have started to provide orthology predictions for a subset of ncRNAs, largely based on synteny. Most of the discussion centered on the difficulties in use of phylogenetic methods for the analysis of ncRNAs: phylogenetic models used for protein coding genes usually assume that sites evolve independently, but ncRNAs often violate this assumption, owing to the importance of secondary structure conservation. Several models specifically developed for RNA sequences have been implemented in phylogenetic packages [e.g. PHASE (Gowri-Shankar and Rattray, 2007) or RAxML (Stamatakis, 2006)], but these models are not widely known. Other limitations hindering phylogenetic study of ncRNAs, include the difficulty in reliably detecting these genes. The RFam database (Gardner *et al.*, 2011) contains a high-quality set of ncRNA families, but its scope is limited to families for which an expert multiple alignment is available. A central repository for RNA sequences has been recently proposed (Bateman *et al.*, 2011) and we see this as important for boosting interest and helping to drive evolutionary studies on RNA sequences.

8 ACHIEVEMENTS AND OUTLOOK

The disparate but interconnected communities represented at this meeting have taken an important step toward better understanding one another. Inferring orthology is a non-trivial task, for many reasons. There are certainly significant computational and algorithmic challenges, but at a more basic level, differing applications driving the quest for orthologs has led to differing definitions of orthology (particularly with respect to subcategories, such as in-paralogs or co-orthologs), the use of different source datasets and different metrics for evaluating performance. The most important achievement to emerge from the Quest for Orthologs effort thus far is a series of consensus agreements, on:

- reference proteome datasets, including a minimal set suggested for benchmarking ortholog detection algorithms, and a larger set, greatly facilitating data sourcing;
- data exchange formats, including OrthoXML and SeqXML; and
- an analysis platform providing for comparison of developer-supplied ortholog calls using diverse metrics (include metrics supplied by users and developers).

The many different uses of orthology detection ensure that there will continue to be a multitude of useful algorithms. Some will be optimized for computational efficiency and/or scalability. Some

will focus on specific phylogenetic groups, which may be highly homogenous or relatively diverse, may or may not exhibit synteny and may include introns or operons, etc. Still other methods will be tailored to handle multidomain proteins, alternative transcription units, metagenomics data, etc. (Dessimoz, 2011).

The availability of reference datasets permits all groups to use the same proteomes, while also minimizing the effort to source the raw data. The OrthoXML format allows predictions to be exchanged efficiently, and the benchmarking platform permits consistent assessment of the results. One of the highlights of the June 2011 meeting was the discussion of orthology prediction methods—a discussion that could only take place because different algorithms were applied to the same source data. Proposed benchmarks are publicly accessible from the Quest for Orthologs portal (http://questfororthologs.org), in order to encourage other researchers to use this platform.

It will be exciting to see the progress of Quest for Orthologs initiatives over the coming years—the next meeting is tentatively scheduled for 2013. In the meantime, the reference proteomes will be updated and enlarged to sample taxonomic space, and the benchmarking service will be made publicly available. We invite all interested parties to join the orthology community, using the contacts available at the aforementioned Quest for Orthologs portal.

ACKNOWLEDGEMENTS

We are grateful to the European Science Foundation (Program on Frontiers of Functional Genomics) for their financial support, which made this meeting possible. We also thank the EBI, and Alison Barker in particular, for the organizational support. DSR and the OrthoMCL database are funded, in part, by a Bioinformatics Resource Center contract from the US NIH (HHSN266200400037C). Members of the Quest for Orthologs Consortium: Adrian Altenhoff, Rolf Apweiler, Michael Ashburner, Judith Blake, Brigitte Boeckmann, Alan Bridge, Elspeth Bruford, Mike Cherry, Matthieu Conte, Durand Dannie, Ruchira Datta, Christophe Dessimoz, Jean-Baka Domelevo Entfellner, Ingo Ebersberger, Toni Gabaldón, Michael Galperin, Javier Herrero, Jacob Joseph, Tina Koestler, Evgenia Kriventseva, Odile Lecompte, Jack Leunissen, Suzanna Lewis, Benjamin Linard, Michael S. Livstone, Hui-Chun Lu, Maria Martin, Raja Mazumder, David Messina, Vincent Miele, Matthieu Muffato, Guy Perrière, Marco Punta, David Roos, Mathieu Rouard, Thomas Schmitt, Fabian Schreiber, Alan Silva, Kimmen Sjölander, Nives Škunca, Erik Sonnhammer, Eleanor Stanley, Radek Szklarczyk, Paul Thomas, Ikuo Uchiyama, Michiel Van Bel, Klaas Vandepoele, Albert J. Vilella, Andrew Yates and Evgeny Zdobnov.

Funding: Open access charges were funded through the meeting registration fees.

Conflict of Interest: none declared.

REFERENCES

Alexeyenko,A. *et al.* (2006) Overview and comparison of ortholog databases. *Drug Discov Today*, **3**, 137–143.

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

Bateman,A. *et al.* (2011) RNAcentral: a vision for an international database of RNA sequences. *RNA*, **17**, 1941–1946.

Boeckmann,B. *et al.* (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief. Bioinform.*, **12**, 423–435.

Chen,R. and Jeong,S. (2000) Functional prediction: identification of protein orthologs and paralogs. *Protein Sci.*, **9**, 2344–2353.

Chen,F. *et al.* (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.

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

Chen,C. *et al.* (2011) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One*, **6**, e18910.

Delsuc,F. *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.*, **6**, 361–375.

Dessimoz,C. (2011) Editorial: orthology and applications. *Brief. Bioinform.*, **12**, 375–376.

du Plessis,L. *et al.* (2011) The what, where, how and why of gene ontology–a primer for bioinformaticians. *Brief. Bioinform.*, **12**, 723–735.

Ebersberger,I. *et al.* (2009) HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.*, **9**, 157.

Fitch,W. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Flicek,P. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Forslund,K. *et al.* (2011) Domain architecture conservation in orthologs. *BMC Bioinformatics*, **12**, 326.

Gabaldón,T. (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.*, **9**, 235.

Gabaldón,T. *et al.* (2009) Joining forces in the quest for orthologs. *Genome Biol.*, **10**, 403.

Gardner,P.P. *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.

Gaudet,P. *et al.* (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.*, **12**, 449–462.

Gerlach,D. *et al.* (2009) miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.*, **37**, D111–117.

Gilks,W.R. *et al.* (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.

Gowri-Shankar,V. and Rattray,M. (2007) A reversible jump method for bayesian phylogenetic inference with a nonhomogeneous substitution model. *Mol. Biol. Evol.*, **24**, 1286–1299.

Heinicke,S. *et al.* (2007) The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One*, **2**, e766.

Henricson,A. *et al.* (2010) Orthology confers intron position conservation. *BMC Genomics*, **11**, 412.

Hofmann,K. (1998) Protein classification and functional assignment. *Trends Guide Bioinformatics*, 18–21.

Huerta-Cepas,J. *et al.* (2011) Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Brief. Bioinform.*, **12**, 442–448.

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

Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.

Kristensen,D.M. *et al.* (2011) Computational methods for Gene Orthology inference. *Brief. Bioinform.*, **12**, 379–391.

Movahedi,S. *et al.* (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in arabidopsis and rice. *Plant Physiol.*, **156**, 1316–1330.

Mushegian,A.R. and Koonin,E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.

Nehrt,N.L. *et al.* (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.*, **7**, e1002073.

Pereira,V. *et al.* (2009) A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics*, **183**, 1597–1600.

Peterson,M.E. *et al.* (2009) Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci.*, **18**, 1306–1315.

Proost,S. *et al.* (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, **21**, 3718–3731.

Pryszcz,L.P. *et al.* (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, **39**, e32.

Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput. Biol.*, **5**, e1000431.

Salichos,L. and Rokas,A. (2011) Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One*, **6**, e18755.

Schmitt,T. *et al.* (2011) SeqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinform*.

Song,N. *et al.* (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.*, **4**, e1000063.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Szklarczyk,R. *et al.* (2012) Iterative orthology prediction uncovers new mitochondrial proteins and identifies C12orf62 as the human ortholog of COX14, a protein involved in the assembly of cytochrome c oxidase. Genome Biol. (in press).

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Trachana,K. *et al.* (2011) Orthology prediction methods: A quality assessment using curated protein families. *BioEssays*, **33**, 769–780.

Uchiyama,I. *et al.* (2010) MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Res.*, **38**, D361–D365.