Better orthology-prediction resources would be beneficial for the whole biological community. A recent meeting discussed how to coordinate and leverage current efforts.

Identifying evolutionarily related genes (that is, homologs) is crucial for understanding the nature of genomic diversity and the routes by which it arises. Further, distinguishing homologs into two types, either ‘orthologs’, genes derived from a common ancestor through speciation, or ‘paralogs’, those derived through a duplication event, has important implications in studying the evolutionary processes that have shaped a given biological system and, since gene duplication is often associated with processes of functional divergence, for inferring the function of related genes. Indeed, many common research processes depend on accurate orthology predictions, such as finding the gene in a model organism corresponding to a human disease gene, inferring the function of a newly sequenced gene using available experimental assays from its orthologs, inferring species phylogenies by tracing the evolution of orthologous groups, or the characterization of newly sequenced genomes in terms of their encoded genes.

The field of orthology prediction clearly requires a new momentum that will help resolve these issues and make better use of available resources. Furthermore, this need becomes more urgent when considering the advent of thousands of new genome sequences, facilitated by next-generation sequencing technologies. Motivated by this prospect, Erik Sonnhammer and Albert Vililla organized the ‘Quest for Orthologs’ meeting at the Wellcome Trust Conference Centre in Hinxton, UK in July 2009, to jointly address these issues by bringing together for the first time key representatives of the major methods and databases in the field of orthology prediction.

The participants gathered for this meeting included experts in gene and genome evolution, developers of orthology-prediction algorithms and databases, and curators of model organism databases. The intimate size of the meeting (approximately 30), the varied perspectives, and the sequestered venue created an ideal environment for intensive and fruitful discussions. All participants were given an opportunity to present their work, while still allowing ample time for informal discussion afterwards. In a thought-provoking talk, Bill Pearson (University of Virginia, Charlottesville, USA) confronted the audience head on by questioning the usefulness of orthology. In his view, homology inference is a more reliable indicator of function conservation, as far as purely sequence-based methods are concerned (though he also stressed the inherent limitations of such methods over long evolutionary distances). He noted that function is often conserved between paralogs; and even if not, the potential benefits of distinguishing orthologs from paralogs are outweighed by the risks of inference errors. Teresa Przytycka (National Center for Biotechnology Information, Bethesda, USA) and Ken Wolffe (Smurfit Institute of Genetics, Dublin, Ireland) gave two examples of how augmenting current methods with additional information (protein-protein interaction and synteny data, respectively) might improve predictions. In contrast, most algorithmic
developments presented at the meeting suggest a general trend toward phylogeny-based orthology inference, a strategy that resembles more closely the original definition of orthology [8]. From all the discussions it was clear that there is a lack of research in the conservation and evolution of protein function. Relevant research will only be possible on the basis of accurately predicted gene histories and functional annotations.

However, the major thrust of the meeting was on identifying points of intersection and the immediate steps that could be jointly undertaken following the workshop to lay the groundwork for the future. As Mike Cherry (Stanford University, Stanford, USA), Pascale Gaudet (Northwestern University, Evanston, USA) and Paul Thomas (SRI International, Menlo Park, USA) explicitly pointed out, and we all clearly recognized, orthologs are essential for more accurately assigning function and the lack of coordination in the field is a major roadblock to progress in this area. Two primary areas for cooperative work were quickly identified: common approaches to benchmark orthology predictions and the formation of standardized protein sets to use as inputs.

**Benchmarking orthology-prediction methods coming from alternative methods**

The assessment of orthology-prediction methods is difficult for at least three reasons: first, orthology is defined from the largely unknown evolutionary history of genes, and thus can only be tested indirectly; second, the optimal trade-off between precision and recall strongly depends on the context; and third, the lack of standardized input datasets and data formats are significant practical hurdles to comparing methods. The meeting saw contributions addressing all three issues. David Roos (University of Pennsylvania, Philadelphia, USA) and Adrian Altenhoff (ETH Zurich, Switzerland) introduced novel comparison strategies - based on latent class analysis and species-tree concordance, respectively [9,10] - that complement existing approaches based on gene order and functional conservation [11]. Yet all of these benchmarks rely on assumptions that do not always hold. For instance, genomic rearrangements, recombination, alternative evolutionary histories, or functional divergence among orthologs all disturb these indicators. More importantly, their relative importance strongly depends on the aim of the user. Thus, we not only recognized the usefulness of multiple and at times contradictory criteria, but also the need for a common understanding on their usage and interpretation. Finally, we agreed that adopting a common dataset would eliminate contradictory criteria, but also the need for a common strategy that resembles more closely the original definition of orthology, OrthoXML [12], which ignited discussions on potential improvements and compatibility issues with existing methods. Such a format would not only facilitate the comparison of methods, but also their combination. For instance, Roos among others suggested a common web interface to the different predictions methods - similar to HPOC [13] and ProGMap [14], the orthology aggregators presented by Michael Lush (European Bioinformatics Institute, Hinxton, UK) and Jack Leunissen (Wageningen University, the Netherlands), respectively. Several working groups have been set up to further develop these ideas and help their realization. Readers interested in participating should contact us.

In retrospect, a remarkable aspect of this meeting is how few of us, despite our strong common interests and goals, had previously met in person. Yet this is an essential first step for building and coordinating collaborative efforts. Given the positive outcomes of this workshop, we are planning to gather again next year to follow up this work and build on the momentum this meeting generated.

**Acknowledgements**

We are grateful to the European Science Foundation and the Wellcome Trust for their generous support, which made this meeting possible, to Pam Garland for coordinating our many requests and making sure we ran on time, to all participants in the meeting for their valuable ideas, and to Michael Ashburner from whom the idea for this valuable workshop also originated.

**References**

1. Gabaldón T: Large-scale assignment of orthology: back to phylogenetics? *Genome Biol* 2008, 9:235.
2. Kuzniar A, van Ham RC, Pongor S, Leunissen JA: The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet* 2008, 24:539-551.
3. Alexeyenko A, Lindberg J, Perez-Bercoff A, Sonnhammer EL: Overview and comparison of ortholog databases. *Drug Disc Today: Technologies* 2006, 3:137-143.
4. Datta RS, Meacham C, Samad B, Neyer C, Sjolander K: Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res* 2009, 37(Web Server issue):W84-W89.
5. Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P: eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 2008, 36(Database issue):D250-D254.
6. Uchiyama I: MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res* 2007, 35(Database issue):D343-D346.
7. Roth AC, Gonnet GH, Dessimoz C: Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 2008, 9:518.

8. Fitch WM: Distinguishing homologous from analogous proteins. *Syst Zool* 1970, 19:99-113.

9. Altenhoff AM, Dessimoz C: Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* 2009, 5:e1000262.

10. Chen F, Mackey AJ, Vermunt JK, Roos DS: Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2007, 2:e383.

11. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7:R31.

12. OrthoXML [http://www.orthoxml.org]

13. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E: The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res* 2008, 36(Database issue):D445-D448.

14. Kuzniar A, Lin K, He Y, Nijveen H, Pongor S, Leunissen JA: ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Res* 2009, 37(Web Server issue): W428-W434.

Published: 29 September 2009
doi:10.1186/gb-2009-10-9-403
© 2009 BioMed Central Ltd