BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 1. 2–5
Advance Access publication September 2, 2006 doi:10.1093/bib/bbl030

HCOP: a searchable database of human orthology predictions

Tina A. Eyre, Mathew W. Wright, Michael J. Lush and Elspeth A. Bruford

Accepted: 6th July 2006; Received (in revised form): 10th April 2006

Abstract

The HUGO Gene Nomenclature Committee (HGNC) Comparison of Orthology Predictions (HCOP) search tool combines the human, mouse, rat and chicken orthology assertions made by PhIGs, HomoloGene, Ensembl, Inparanoid, Mouse Genome Informatics (MGI) and HGNC, enabling users to identify predicted ortholog pairs for a specified gene or genes. The HCOP resource provides a useful method to integrate, compare and access a variety of disparate sources of human orthology data.
The HCOP search tool, data and documentation are available at http://www.gene.ucl.ac.uk/hcop.

Keywords: *orthology; nomenclature; database; comparative; genes*

INTRODUCTION

Orthologs are genes in two or more species that share significant homology, and are thought to derive from a common ancestral gene without duplication. Now that a number of genome sequences are available, defining orthology relationships is becoming a priority. The HUGO Gene Nomenclature Committee (HGNC) Comparison of Orthology Predictions database (HCOP, http://www.gene.ucl.ac.uk/hcop) integrates the human, mouse, rat and chicken orthology assertions made by a number of different groups. Until now these data have only been available from disparate sources, and a single tool for comparison of these data to identify a consensus orthology prediction has been lacking.

QUERYING THE HCOP DATABASE

The HCOP database can be searched using one or more approved symbols, Entrez Gene Ids [1], HGNC IDs [2], MGI IDs [3] or RefSeq IDs [4]. The wildcard '_' can be used to substitute a single character and '*' or '%' for zero or more characters, to search for a number of related symbols or identifiers. A file containing a list of identifiers can also be uploaded.

Corresponding author. Elspeth A. Bruford, HUGO Gene Nomenclature Committee (HGNC), Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London, NW1 2HE. Tel: 0207 679 7410; Fax: +44 20 7387 3496; E-mail: nome@galton.ucl.ac.uk.

Tina A. Eyre obtained her PhD in bioinformatics with Janet Thornton at UCL in 2005, before working as a bioinformatician for the HUGO Gene Nomenclature Committee for 18 months. In May 2006 she moved to the Sanger Institute, where she currently works as a Senior Computer Biologist.

Mathew W. Wright received his PhD in Neuroscience from University College London; in 2000 he joined the HUGO Gene Nomenclature Committee (HGNC) and in 2004 was appointed as the HGNC's HUMOT (Human and Mouse Orthologous Annotation) editor.

Michael J. Lush received his PhD at Leicester University entitled 'Molecular cloning of Neuropathy Target Esterase'. Subsequently taking a position cloning T-cell antigens and autoantigens in psoriasis (Leicester University). In 2000, he joined the HUGO Gene Nomenclature Committee providing vital Bioinformatics support.

Elspeth A. Bruford received her PhD mapping retinal diseases at the MRC Human Genetics Unit, Edinburgh. After working in publishing, she moved to University College London in 1998 to join the HUGO Gene Nomenclature Committee, of which she is now the Project Co-ordinator.

DATA SOURCES

The data used in HCOP are derived from Phylogenetically Inferred Groups (PhIGs) [5], Homologene [6], EnsEmbl [7], InParanoid [8, 9], Mouse Genome Informatics (MGI) [3] and HGNC [2]. These sources have used a variety of computational methods [10] to generate orthology assertions: PhIGs looks at phylogeny based on the Ensembl dataset [5]; Homologene begins with BLASTP followed by phylogenetic analysis (with synteny where possible); and EnsEmbl includes both BLASTP and synteny evidence. Paralogs are genes related by duplication from a common ancestor and are not restricted within a genome; they are either inparalogs that arise through a gene duplication event after speciation, or outparalogs that arise following a gene duplication before speciation. InParanoid differentiates paralogs as either inparalogs or outparalogs by using the best hits from a BLAST analysis between sequences from two different species. MGI uses both computational and manual approaches [11], and only the HGNC manually curate all of their orthology predictions. HCOP combines the results of these sources into a useful consensus prediction. Previously, HCOP was restricted to human and mouse data [12] but it has recently been expanded to include orthology data for rat and chicken. We are now also considering the inclusion of further orthology data sources.

Ortholog pairs are represented in the HCOP database as pairs of Entrez Gene (EG) IDs. These are imported directly from MGI, HGNC, Homologene and InParanoid. The other databases provide pairs of Ensembl IDs that were converted to EG IDs using Ensembl data. A non-degenerate list of EG ID pairs is then generated which collates each pair of EG IDs, a list of the databases that made the assertion and a link to the original data. All other data is mapped directly from Entrez Gene. The database is updated automatically once a week, although these updates are reliant on the update frequency of the source databases.

Conserved synteny is a term used to describe genes occurring in a specific chromosomal region in one species and in the equivalent region in a second species. Syntenic chromosomes, predicted by the MGI Mouse and Human Orthology Map (http://www.informatics.jax.org/reports/homologymap/mouse_human.shtml), were used to assess the possible conserved synteny of predicted human/mouse ortholog pairs, based on chromosomal locations provided by Entrez Gene. See http://www.gene.ucl.ac.uk/nomenclature/data/humot_documentation.html#synteny for more details. It is useful to assess the synteny of potential orthologs since syntenic genes are more likely to be true orthologs. Synteny data will be added for other species as it becomes available.

DATABASE IMPLEMENTATION

The HCOP database is a fully indexed PostgreSQL v8.03 database and its search engine is a Perl Common Gateway Interface (CGI) script querying this database. HTML::Template is used to allow rapid generation of complex tables containing multiple ortholog pairs records from simple repeating units.

The data sources described above provided 142 372 pairs of predicted orthologous EG IDs covering 18 131 genes (as of 4 July 2006). These pairs of EG IDs were consolidated into a non-redundant list of 47 177 orthology assertions, with an associated list of databases that support each assertion. These data are stored in the HCOP database, along with additional information about each gene.

HCOP Output

HCOP search results provide the official nomenclatures, sequence accession numbers, database identifiers, aliases and chromosomal locations for each putative ortholog pair. If all databases agree on the ortholog, a list of databases that support that assertion and links to sources of further information are provided. If, however, a consensus ortholog is not agreed upon by all databases, a list of all available predictions is shown and the user is then left to interpret these as they see fit. For instance, users can assess the reliability of the prediction from the number of different sources that identify a particular orthologous gene pair, and from the indicated presence or absence of chromosomal synteny between a gene and its predicted ortholog or vice versa.

The HCOP URL has recently been streamlined to make it more memorable: http://www.gene.ucl.ac.uk/hcop now takes the user directly to the search tool. A particular search result showing a set of genes of interest can be easily returned to, by bookmarking the URL, or linked to from external webpages. Documentation and help is available at http://www.gene.ucl.ac.uk/nomenclature/data/humot_documentation.html.

APPLICATIONS OF HCOP DATA

One of the main goals of the human and mouse Nomenclature Committees is to develop equivalent nomenclature in both species (e.g. *KLF1* in human and *Klf1* in mouse). HCOP was originally developed to generate comparative files listing predicted human/mouse ortholog pairs. These data are available for download at http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/hcop_hum_mus.pl. Already, they have been valuable in identifying human/mouse ortholog pairs with differing approved nomenclature. This resource has subsequently been expanded to identify consensus orthology assertions for rat and chicken, and could in due course include other species, allowing annotation and nomenclature to proliferate out to other mammalian species as their genomic sequence becomes available.

Analysis of HCOP data has also enabled the level of agreement between the orthology assertions made by different databases to be assessed. This analysis has found that there is in general good agreement between the different orthology databases, although most sources provide orthology assertions for only a small number of genes.

The agreement between the assertions made by different databases is very high, typically around 98%. Particularly in cases where the databases involved have used very different methods to generate orthology predictions (e.g. the phylogenetic methods of PhIGs and the BLAST analysis of EnsEmbl), we can be confident that a consensus assertion delivered by HCOP is correct in the majority of cases.

The HCOP search identifies a single predicted human ortholog for 96% of mouse and rat genes and 97% of chicken genes, even in cases where three or more databases have predicted an ortholog. In contrast, multiple possible orthologs are predicted in one or more species for approximately 1500 of the 18 131 genes (8%). The majority of these belong to large gene families with many closely related members, such as the olfactory receptor genes. In these cases the predicted multiple 'orthologs' may actually represent paralogs. The maximum number of predicted orthologs for a particular gene in the current dataset is 24 mouse orthologs for human PRAME family member 2 (*PRAMEF2*), which lies in a region of chromosome 1 (1p36.1) that is known to have undergone considerable duplication in both humans and rodents [13]. One important result of this work is the ability to identify genes such as these for which automated orthology prediction has proved unsuccessful, so that they become the focus of manual curation work. However, it should be noted that in some cases the assignment of orthology may not currently be possible.

For several databases coverage is somewhat low. Approximately 83% of the 18 131 genes have orthology assertions made by three or more of the six databases. HGNC provides the smallest number of orthology assertions (approximately 6000), while HomoloGene provides the greatest number (around 56 000). While the coverage of HGNC is small, its data is the only set to have been fully manually curated, providing a good check of the quality of the automatic methods used by other groups.

Given the low coverage of many databases and the problems associated with some large gene families it would be unwise to rely entirely on a single source of orthology predictions. In an attempt to avoid these problems, HCOP provides a valuable method to simultaneously compare data from a number of sources.

CONCLUSIONS

HCOP provides a useful tool to rapidly obtain and compare orthology data from a variety of sources, allowing users to make an informed identification of the most likely correct ortholog. In the future it is likely to be expanded to cover more species, facilitating our understanding of orthology relationships throughout mammalian species.

CITATION

Authors are requested to cite this article and the HCOP resource in the following format: 'The HCOP Search Tool, HUGO Gene Nomenclature Committee (HGNC), Department of Biology, University College London, Wolfson House, 4

Key Points

- Now that a number of genome sequences are available, defining orthology relationships is becoming a priority.
- The HUGO Gene Nomenclature Committee (HGNC) Comparison of Orthology Predictions database (HCOP) integrates the human, mouse, rat and chicken orthology assertions made by PhIGs, HomoloGene, Ensembl, Inparanoid, MGI and HGNC. Until now these data have only been available from disparate sources and a single tool for comparison of these data to identify a consensus orthology prediction has been lacking.
- The HCOP resource, available at http://www.gene.ucl.ac.uk/hcop, provides a useful method to integrate, compare and access these sources of human orthology data.

Stephenson Way, London NW1 2HE, UK (URL: http://www.gene.ucl.ac.uk/hcop)' [Include month and year in which you retrieved the data cited.]

Acknowledgements

Many thanks to the other members of the HGNC team, especially Prof. Sue Povey and Fabrice Ducluzeau. This work was funded by the Wellcome Trust as part of the HUMOT Project. The HGNC is also supported by NHGRI grant P41 HG003345 and the UK Medical Research Council.

References

1. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005; **33**:D54–8.
2. Eyre TA, Ducluzeau F, Sneddon TP, *et al*. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res* 2004;**34**:D319–21.
3. Blake JA, Eppig JT, Bult CJ, *et al*. The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res* 2006;**34**:D562–7.
4. Pruitt KD, Tatusova I, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2005;**33**:D501–4.
5. Dehal PS, Boore JL. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* 2006;**7**:201.
6. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2006;**34**: D173–80.
7. Birney E, Andrews D, Caccamo M, *et al*. Ensembl 2006. *Nucleic Acids Res* 2006;**34**:D556–61.
8. Remm M, Storm CEV, Sonnhammer ELL. Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. *J Mol Biol* 2001;**314**: 1041–52.
9. O'Brien, Remm M, Sonnhammer ELL. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 2005;**33**:D476–80.
10. Wright MW, Bruford EA. Human and orthologous gene nomenclature. *Gene* 2006;**369**:1–6.
11. Eppig JT, Bult CJ, Kadin JA, *et al*. The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res* 2005;**33**: D471–5.
12. Wright MW, Eyre TA, Lush MJ, *et al*. HCOP: the HGNC comparison of orthology predictions search tool. *Mamm Genome* 2005;**16**:827–8.
13. Birtle Z, Goodstadt L, Ponting C. Duplication and positive selection among hominin-specific PRAME genes. *BMC Genomics* 2005;**6**:120–39.