# News & Views from around the ISB Community

## Check if your database or repository is missing

Elsevier is keen to work with researchers and data repositories to ensure that data that is relevant for scientific, technical and medical research can be easily discovered and accessed. One of the ways in which we are doing this is by creating bidirectional links between data repositories and online articles on ScienceDirect. This provides ScienceDirect's readers with one-click access to relevant, trusted data that may help to validate research or drive further investigations. Linking helps to make articles and data better discoverable, attracting more usage. Sharing the data that underpins conclusions is not only good scientific practice, but also increasingly required by funding bodies.


Specific journal instructions for authors depend on the data repository: in some cases data is extracted from the article by curators, while in other cases authors need to upload their data manually.

We need to hear from you with your detailed instructions. Detailed information for already established partnerships is available in the (recently revised and updated) listing of supported databases. Alternatively, copy & paste this link: http://goo.gl/8Jh8k

If you are a data repository manager interested in setting up bidirectional linking with Elsevier publications, please contact us at articleofthefuture@elsevier.com

Submitted by Adriaan Klinkenberg.

## The M:N Project at MGD: Beyond 1:1 Orthology Assertions

The Mouse Genome Database (MGD) curates, integrates, and provides comprehensive genetics, genomic, and phenotypic information for the laboratory mouse, a primary model organism for experimental investigation of human biology and disease. MGD is found at http://www.informatics.jax.org.

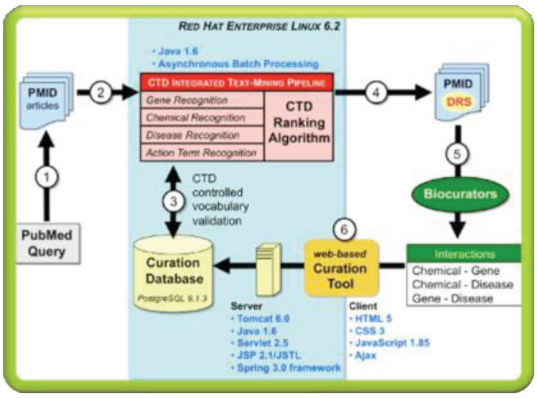
A core component of MGD data for over 20 years has been the curated assertion of 1:1 orthology between mouse, human, and rat protein-coding genes. Now, with completely sequenced genomes available for comparative analysis, phylogenetic analysis clearly identifies cases where descent from common ancestor does not always define a 1:1 relationship, but rather that gene duplication following an ancestral speciation event more correctly results in M:N relationship between genes in different species.

This has implications for the study of human biology in the mouse system and for the presentation of inferential functional and disease associated assertions based on comparative analysis. MGD has recently restructured its database to accommodate such homology classes with concurrent changes in presentation of data related to homology classes and in the representation of human diseases associated with mouse genes by curation of comparative or experimental data. We load data from all mammalian species with completed genome sequences, and will next extend our data to include chicken and Zebrafish protein-coding gene classes. While 1:1 assertions predominate (~80%), we now more clearly represent cases such as the Serpina1 gene class (1 human, 5 mouse, 1 rat), and provide better cross-referencing among related genes, the diseases that have been studied in respect to those genes, and the relationship between genomic features in related genomes. This work is supported by NIH NHGRI grant HG-000330.

Submitted by Judith A Blake, Richard Baldarelli, Mary Dolan, Mark Airey, Jon Beal, Sharon Giannatto, David Miers, Jill Lewis, Carol J. Bult, Janan T. Eppig and James Kadin.

## The Comparative Toxicogenomics Database Text-Mining Pipeline

The Comparative Toxicogenomics Database (CTD; http://ctdbase.org/) confirmed the effectiveness of its text-mining pipeline in evaluating and prioritizing scientific literature for the manual curation of chemical-gene-disease information. The results were selected for a special issue of PLoS Text Mining Collection in April (http://goo.gl/f1ubr).




For the study, CTD tested their sophisticated text-mining algorithm by using it to evaluate the text from 15,000 articles and assign a relevancy score to each document. A representative sample of the corpus was sent to their team of biocurators to manually read and evaluate on their own, blind to the computer's score. The biocurators concurred with the algorithm 85% of the time with respect to the highest-scored papers, and there was a clear step-wise progression, wherein the likelihood of an article's true relevancy decreased linearly as the text-mining scores declined.

Ranking papers by text mining allowed biocurators to focus on the most relevant papers and avoid the extraneous ones, increasing productivity by 27% and novel data content by 2-fold. The curated articles were also broad and encompassing with respect to data coverage, finding both shared as well as unique biological processes, pathways, and toxicological end-points, confirming that the ranking system could help identify articles that contribute to a mechanistic understanding of toxicity.

By incorporating similar text mining-based scoring, other databases may also be able to enhance their manual curation by prioritizing more relevant articles, thereby increasing data content, productivity, and efficiency.

Submitted by Allan Peter Davis.

ISB Spotlight provides a snapshot of some of the work and activities of members of the International Society for Biocuration (ISB). The Spotlight features brief descriptions of a range of databases and biocuration tools, re-published, with permission, from the 'News & Views' section of ISB's monthly newsletter. The newsletter, with the complete 'News & Views from the ISB Community', is freely available from http://biocurator.org/newsletter.shtml.