

## STRUCTURED VOCABULARIES IN MOUSE GENOME INFORMATICS

J.A. Kadin, J.A. Blake, J.E. Richardson, M. Ringwald, C.J. Bult, J.T. Eppig, and the Mouse Genome Informatics Group.

The Jackson Laboratory, Bar Harbor, ME 04609.

The Mouse Genome Informatics (MGI) group at The Jackson Laboratory is working on several informatics projects involving large structured controlled vocabularies.

The first is the Anatomical Dictionary of Mouse Development which is being developed in collaboration with MRC Human Genetics Unit, Edinburgh and the University of Edinburgh<sup>1</sup> as part of the Gene Expression Database (GXD) project. This dictionary represents each of the 28 Theiler stages of mouse development as a hierarchy of anatomical structures. Currently stages 1-22 are completed along with an abridged adult mouse dictionary. Since June 1998, gene expression results have been annotated against this dictionary with over 1253 genes annotated in more than 51,360 results involving more than 1967 structures. The dictionary itself has over 7357 structures defined.

A second large structured vocabulary is the Gene Ontology (GO). This project is a collaboration among the *Saccharomyces* Genome Database (SGD), FlyBase, and MGI to develop three independently structured vocabularies (gene ontologies) describing molecular functions, biological processes, and cellular locations of gene products. Genes are annotated in the three respective databases against these vocabularies allowing powerful queries within each database and providing a large set of standard terms to support queries across the three species' databases. Each of the three vocabularies is structured as a directed acyclic graph; all three together comprise over 4704 nodes (terms). MGI currently has over 4382 genes annotated with these vocabularies.

A third effort, which is still in the very early stages, is to develop controlled vocabularies describing phenotypic traits and mouse models for human diseases in consultation with mutagenesis centers and other phenotyping groups. These vocabularies will provide a rich set of terms for annotating phenotypes and will be critical to our ability to query, compare, and analyze phenotypic data in the future.

Through the Mouse Genome Database (MGD), the Gene Expression Database (GXD) project, and the Mouse Genome Sequence (MGS) project, the Mouse Genome Informatics group provides integrated resources of mouse genetic, genomic, and biological information. Areas of data coverage include maps and mapping data, gene nomenclature and gene descriptions, phenotype descriptions, strain polymorphisms and characteristics, molecular reagents, gene expression data, cDNA to gene associations, ortholog relationships to human and other model organisms, and a MouseBLAST server.

MGI is accessible to the public at <http://www.informatics.jax.org>. MGD is supported by NIH grant HG00330. GXD is supported by NIH grant HD33745. MGS is supported by DOE grant DE-FG02-99ER62850 and NIH grant HG01559.

---

1. Edinburgh collaborators: J. Bard, R. Baldock, D. Davidson, M. Kaufman