

An Ontology-Based Approach to Linking Model Organisms and Resources to Human Diseases

Christopher J. Mungall¹, David Anderson², Anita Bandrowski³, Brian Canada⁴, Andrew Chatyr-Aryamontri⁵, Keith Cheng⁶, P. Michael Conn⁷, Kara Dolinski⁸, Mark Ellisman³, Janan Eppig⁹, Jeffrey S. Grethe³, Joseph Kemnitz¹⁰, Shawn Iadonato¹¹, Stephen D. Larson³, Charles Magness¹¹, Maryann E. Martone³, Mike Tyers¹², Carlo Torniai⁷, Olga Troyanskaya⁸, Judith Turner¹³, Monte Westerfield¹⁴, Melissa A. Haendel⁷

¹Lawrence Berkeley National Laboratory, Berkeley, CA, USA; ²University of Washington, Seattle, WA, USA;

³University of California, San Diego, CA USA; ⁴University of South Carolina, Beaufort, SC, USA;

⁵University of Edinburgh, UK; ⁶Penn State College of Medicine, Hershey, PA, USA;

⁷Oregon Health & Science University, Portland, OR, USA; ⁸Princeton University, Princeton, NJ, USA;

⁹The Jackson Laboratory, Bar Harbor, ME, USA; ¹⁰University of Wisconsin-Madison, WI, USA;

¹¹Kineta, Inc., Seattle, WA, USA; ¹²Samuel Lunenfeld Research Institute, Toronto, ON, Canada;

¹³TCG, Washinton, DC, USA; ¹⁴University of Oregon, Eugene, OR, USA

Keywords: model organism, phenotype, gene orthology, similarity algorithm

Abstract

The scientific community has invested heavily in the creation of genetically modified organisms, other model systems, and large genetic screens because they greatly inform our understanding of human disease. However, it remains difficult to identify organisms suitable for one's research because information about them is not readily accessible. The initiative to Link Animal Models to Human Disease (LAMHDI; <http://lamhdi.org>) was developed to allow users to search for a diverse set of models of disease using both curated disease-model links and inferred paths based on gene orthology and pathway membership. These inferences are made by traversing connections between records in publicly available data from resources such as the Online Mendelian Inheritance in Man (OMIM), Medical Subject Headings (MeSH), EntrezGene, Homologene, and WikiPathways. This allows researchers to rapidly explore and identify a wide range of model systems, visualizing the multi-step genetic relationship between disease and model. However, if LAMHDI were able to semantically link an organism's phenotypic attributes to diseases, genes, expression profiles, etc, their relevance and utility to a given line of research would be much more greatly illuminated and new novel insights between disease, genetics and phenotype discovered.

The relationship between model systems and disease phenotypes [1] is not straightforward, and bioinformatics tools based on phenotypes have been lacking. Constructed model systems typically only replicate subsets of disease phenotypes, and phenotypes may map to more than one human disease [2]. Classification systems where a model system is listed as a "model of disease X" do not solve these difficulties because the model may only recapitulate one aspect of the disease, but not specifically indicate. Further, different vocabularies are used to describe the phenotypic consequences of mutation in different organisms, and these vocabularies are usually tied to the particular anatomies or physiologies of the organism. The different vocabularies also come from different starting points: clinicians and researchers have different vocabularies. Phenotypes may also occur at different scales, and the relationships among them may not be apparent without additional knowledge. Bioinformatics tools, or even researchers, may not relate the statement 'CA1 dendrites are degenerative' to 'degenerative hippocampus' despite potential scientific correlation. Another challenge is to traverse anatomical structures across species. For example, computers do not know that the human cornea is related to the zebrafish retina because they do not know that in both species these structures are part of the eye, nor do they know that the zebrafish eye is

related to the human eye. When phenotype descriptions are captured using an ontology, algorithms can be written to compare phenotypes computationally across species and scale.

Our previous work has shown that ontological annotation of diseases and phenotypes allows computational comparison of phenotypes across species [3]. To describe phenotypes we composed Description Logic (DL) expressions using a phenotype model and the Web Ontology Language (OWL) [4]. We created bridging ontologies that enhance external ontologies such as the Mammalian Phenotype Ontology (MPO) [5], allowing them to be integrated with other phenotype data [6-8]. We applied these methods to the construction of PKB [9], a neurodegenerative disease phenotype knowledgebase called that utilizes the NIF Standard (NIFSTD) modular collection of ontologies [10] to represent a range of human diseases and animal models spanning multiple anatomical scales, from the molecular and subcellular up to the organismal. We also created an integrative ontology called Uberon [11] to allow cross-species inference. Using these tools and methods, we demonstrated that we could identify organisms with similar phenotypes across anatomical scale, mutations in other alleles of the same gene, other members of a signaling pathway, and orthologous genes and pathway members across species based on the similarity of the phenotypes alone.

Here, we bring together the genetic links currently used within LAMHDI with these ontology-based phenotype similarity methodologies. We also combine orthology and gene-phenotype ontology associations to generate “phenolog” hypotheses, non-obvious linkages between human diseases and phenotypes in model organisms such as mice, worms, yeast, and plants [12]. This approach can suggest new models based on the presence of orthologous genes inside a phenolog cluster. We believe that these ontology-based phenotypic and homology-based techniques will be instrumental in enhancing the LAMHDI portal, suggesting new paths from diseases to models, and assisting in the interpretation of existing paths.

At present, LAMHDI is restricted to providing access to organism strains, and does not include *in vitro* model systems and organismally derived resources such as

biospecimens, cell lines, assays, and reagents. Choosing an appropriate *in vitro* model system to test a given hypothesis is currently not straightforward because these resources are often not linked to diseases, genes, gene expression, or the phenotypes of the organisms from which they are derived. Two projects, the Neuroscience Information Framework (NIF) and eagle-i [13], have built ontologies to catalog and link such resources. The eagle-i system has related specific genotypes of organisms to these *in vitro* models, and has expressly represented anatomical, histological, and pathological attributes of biospecimens and cell lines and tied them into the phenotypic representation of the original organism and its genotype. NIF has focused on collating publicly available resources (materials, software tools, data). The two projects are thus highly complementary.

Navigation of organismal resources (*in vivo* and *in vitro*) also benefited from including gene expression data collected using anatomical ontologies from databases aggregated in NIF. Where genes are preferentially expressed, gene expression profiles can be linked between the current LAMHDI data (genes to disease) with cell lines and tissues. For instance, a researcher looking at a set of genes expressed in a particular brain region may investigate a common mechanism for two diseases if tissue from that brain region is available from a bank that focuses on a different disease. By leveraging the organismal resource component of the eagle-i and NIF systems, and the gene expression component of NIF, LAMHDI will be able to offer many new candidate model systems and make these resources easier to navigate. We discuss the issues of integrating these diverse data and our technical approach to this challenge.

Acknowledgement

LAMHDI is supported by contract HHS N268200800014C from NIH/NCRR.

References

1. Houle, D., Govindaraju, D. R., and Omholt, S. Phenomics: the next challenge.: *Nat Rev Genet*, 11 (12), 855-66 (2010)
2. Gupta, A., Ludascher, B., and Martone, M.E. BIRN-M: a semantic mediator for solving real-world neuroscience problems. *Proceedings of the 2003 ACM SIGMOD international conference on*

- Management of data (ACM New York, NY, USA), 678 (2003)
3. Washington*, N.L., Haendel*, M.A. Mungall, C.J., Ashburner, M., Westerfield, M., Lewis, S.E. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7 (11) *Contributed equally (2009)
 4. Mungall, C., Gkoutos, G.V., Washington, N.L., Lewis, S. Representing phenotypes in OWL. In C. Golbreich, A. Kalyanpur, and B. Parsia (eds.): *Workshop on OWL: Experiences and Directions* (Innsbruck, Austria) (2007)
 5. Smith, C. L, Goldsmith, C. W., and Eppig, J.T.: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*, 6 (1), R7 (2005)
 6. Mungall, C.J., Gkoutos, G.V., Smith, C.L., Haendel, M.A., Lewis, S.E., Ashburner, M.: Integrating phenotype ontologies across multiple species. *Genome Biol*, 11 (1), R2 (2010)
 7. Gkoutos, G.V., Mungall, C.J., Doelken, S., Ashburner, M., Lewis, S., Hancock, J.M., Schofield, P.N., Kohler, S., Robinson, P.N. (2009) Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009)*
 8. Hancock, J.M., Mallon, A.M., Beck, T., Gkoutos, G.V., Mungall, C., Schofield, P.N.: Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mamm Genome*, 20 (8), 457-61 (2009)
 9. Maynard, S.M., Mungall, C.J., Lewis, S.E., Martone, M.E.: A knowledge based approach to matching human neurodegenerative disease and associated animal models. *Neuroscience 2010* (230.4; San Diego) (2010)
 10. Bug, W.J., Ascoli, G.A., Grethe, J.S., Gupta, A., Fennema-Notestine, C., Laird, A.R., Larson, S.D., Rubin, D., Shepherd, G.M., Turner, J.A., Martone, M.E.: The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, 6 (3), 175-94 (2008)
 11. Haendel, M.A., Gkoutos, G.V. Lewis, S. Mungall, C.J. Uberon: towards a comprehensive multi-species anatomy ontology.: *International Conference on Biomedical Ontologies*, Buffalo, NY: Nature Proceedings (2009)
 12. McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., Marcotte, E.M.: Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci U S A*, 107 (14), 6544-9 (2010)
 13. Torniai, C., Bashor, T., Bourges-Waldegg, D., Corday, K., Frost, H.R., Johnson, T., Segerdell, E., Shaffer, C.J., Stone, L., Wilson, M.L., Haendel, M.A.: eagle-i: an ontology-driven framework for biomedical resource annotation and discovery. *Bio-Ontologies 2010: Semantic Applications in Life Sciences*, (Boston, MA) ISMB (2010)