ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Waiting for a Robust Disease Ontology: A Merger of OMIM and MeSH as a Practical Interim Solution

Susan M Bello[1], Allan Peter Davis[2], Thomas C Wiegers[2], Mary E Dolan[1], Cynthia Smith[1], Joel Richardson[1], Judith Blake[1], Carolyn Mattingly[2], Janan T Eppig[1]

[1]Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME
[2]Comparative Toxicogenomics Database, The Mount Desert Island Biological Laboratory, Salisbury Cove, ME

Abstract. The curation of human diseases suffers from the lack of a robust, publicly available disease ontology. While waiting for such an ontology, bioinformatics resources associating genetic and genomic data with human diseases need an interim solution. The Comparative Toxicogenomics Database (CTD) produced a practical, structured vocabulary by curating the association of two subsections of the National Library of Medicine's Medical Subject Headings (MeSH) with Online Mendelian Inheritance in Man (OMIM). The MeSH subsections provide hierarchical access to broad disease terms. OMIM provides detailed disease descriptions and links to associated human genes and mutations. Both resources are freely available and familiar to the research community. Mouse Genome Informatics (MGI) reviewed and modified this vocabulary to adapt it for curation of mouse models of human disease. Here we describe the merged vocabulary and discuss the strengths and weaknesses of this approach. CTD's merged MeSH – OMIM vocabulary can be accessed at http://ctd.mdibl.org/voc.go?type=disease.

Introduction

The need to curate disease-related data is a pressing one for many databases. There is a growing pool of available experimental data and increasing pressure from funding agencies to make clear connections between disease related data from model organisms and the human diseases they reference. There are a number of disease vocabularies and ontologies that may be used for the purpose of annotating disease models each with its own advantages and disadvantages [1]. At Mouse Genome Informatics (MGI) [2] disease annotations are currently made to OMIM [3]. OMIM was initially selected based on the presence of detailed disease descriptions, links between disease records and human genes, and familiarity to biomedical researchers. However, the absence of hierarchical structure in OMIM and the absence of generic disease classes, such as Parkinson Disease, have resulted in a growing collection of mouse models of human disease that cannot be annotated using OMIM since these mice are only described in terms of a generic form of the disease.

Accordingly, MGI sought to identify a disease ontology or vocabulary to better annotate mouse models of human disease. Criteria for a selecting a disease ontology have been published before [1,4]; however, which criteria are considered and how much weight is put on the criteria varies depending on the perspective of the end user. The criteria considered here include several of those described by Bodenreider and Burgen [1] (coverage of diseases, regular maintenance, support for reasoning, open availability). Additional criteria included; stability of the vocabulary, percentage of terms with definitions, inclusion of synonyms and familiarity of the vocabulary to the user community. A final and necessary consideration was presence of links to OMIM.

Several of these additional criteria are generally applicable to any ontology selection process for use in annotation. A stable ontology avoids the need for extensive and repeated re-curation of data. Deep synonym coverage allows for easier identification of diseases from the literature and for more effective searching of the data by users. Definitions provide a description of the disease

to aid in understanding of the disease term and provide a basis for comparison to the model. Familiarity of the user community improves the likelihood that users will readily find the disease(s) for which they are searching. However, the inclusion of links to OMIM was important both for the general value of the OMIM text resources to MGI users and to efficiently use existing disease model annotations in MGI.

Of the existing disease ontologies and vocabularies, two were identified as containing at least some links to OMIM; The Disease Ontology (DO) [5] and the Comparative Toxicogenomics Database's (CTD) combined MeSH and OMIM disease vocabulary [6]. While the DO may grow into a better long-term solution, it is, as of now, not nearly mature or robust enough to be useful for curating disease data. The DO is being extensively revised (negatively impacting stability), only 11% of the terms have definitions (as of 6/21/10), and while OMIM ids are being added many are still missing and there is uneven mapping of OMIM diseases within the DO (Drs. Lynn Schriml and Warren Kibbe, personal communication). Therefore we undertook an extensive review of the CTD merged disease vocabulary.

1 OMIM - MeSH Combined Vocabulary

CTD created, implemented and maintains the merged OMIM-MeSH vocabulary (manuscript in preparation). Two subsections of MeSH were used to create the vocabulary: Diseases [C] and Mental Disorders [F03]. OMIM terms were limited to those with a known gene locus. To merge the vocabularies, all selected OMIM terms were mapped manually to MeSH terms based on semantic similarity (i.e. OMIM 101000, NEUROFIBROMATOSIS, TYPE II, was merged with MeSH D016518, Neurofibromatosis 2; OMIM 125850) or symptom matching for OMIM terms lacking a semantic match. The use of symptoms for mappings allows for rapid and consistent mapping but, curation issues can result and this is discussed more fully below.

Of the 7052 OMIM phenotype or disease records [3], 4365 were associated with a gene map locus. CTD had mapped 4049 of this set to one or more MeSH terms[6]. The 316 unmapped terms were largely records for phenotypic variation (e.g., Hair Morphology 2, OMIM:139450; ABO Blood Group, OMIM:110300). CTD loads the vocabularies on a monthly basis and any discrepancies are identified and curated. The merged vocabulary can be accessed at: http://ctd.mdibl.org/voc.go?type=disease.

2 MGI Review

The CTD disease vocabulary was reviewed to determine the extent of coverage of OMIM terms in use by MGI. The first review (conducted in June 2010) identified 347 OMIM terms in MGI but absent from the CTD disease vocabulary. 259 of these were in CTD's unmapped set. The remaining 88 terms were either new OMIM terms or OMIM terms without a gene map locus. A second review (conducted in August 2010) identified 212 terms in MGI but absent from the CTD disease vocabulary. 37 were repeats from the first review. 90 were new OMIM terms. 85 were existing OMIM terms without a gene map locus. All unmapped OMIM terms were then examined and either mapped to appropriate MeSH terms or added to the unmapped term set.

The reviews also revealed a difference in the desired level of granularity of the vocabulary between CTD and MGI. CTD merged many gene specific OMIM disease terms with the general disease term. For example, AGAMMAGLOBULINEMIA 1 (OMIM 601495) caused by a mutation in *IGHM* and AGAMMAGLOBULINEMIA 6 (OMIM 612692), caused by a mutation in *CD79B*, were both merged with MeSH D000361, Agammaglobulinemia. MGI would prefer both OMIM terms be made children of the MeSH to allow for distinction between orthologous and non-orthologous mouse models. Therefore the CTD OMIM to MeSH mappings were reviewed to identify all instances where an additional level of granularity was desired. The mappings have been internally annotated to record such cases. With these modifications the CTD vocabulary includes terms for all of MGI's existing mouse models of human diseases, allows for annotation of models that cannot be

annotated using OMIM and will improve user access to disease model annotations.

3 Symptom Based Mapping

Mapping of OMIM terms based on disease symptoms has consequences. That a disease produces a symptom in an organ or tissue does not necessarily mean that the disease is a disease of that organ or tissue. For example, in MeSH, albinism is a child of eye diseases and pigmentation diseases, while experts would agree that albinism is a pigmentation disease, many would not consider it an eye disease. However, symptom based mapping can also explain and illuminate a disease. For example, mapping the OMIM term RIDDLE SYNDROME (OMIM 611943) to the MeSH terms for its symptoms (immune deficiency syndromes, learning disorders, and facies) provides insights into the disease. Unfortunately, some symptom descriptions may lead to erroneous mappings if the mapping is not constructed or reviewed by a clinician. Symptoms described as being "like" some other disease or syndrome, may be semantically, yet erroneously, mapped to that disease. For example, patients with Lujan-Fryns Syndrome are described as having "Marfanoid habitus", a term seemingly related to the term 'Marfan' but whose definition is not related to Marfan Syndrome. The symptom based association results in a mapping of Lujan-Fyrns Syndrome to Marfan Syndrome, which is incorrect. These kinds of situations require experts in disease phenotypes to identify, review and curate. Such clinical experts must be an integral part of any disease ontology development effort.

4 The Future

CTD's merged vocabulary is an interim solution to a pressing curation need. Both MGI and CTD plan to migrate annotations to a comprehensive disease ontology, once one is mature and ready for broad use. The merged vocabulary should inform development of a new disease ontology and incorporation of OMIM and MeSH identifiers into developing disease ontologies will greatly aid in future migration of existing annotations to a new ontology.

Acknowledgments

CTD is supported by ES014065 from NIEHS/NLM, ES014065-04S1 from NIEHS, and P20RR016463 from NCRR. MGI is supported by HG000330 from NHGRI/NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Bodenreider, O., and Burgun, A.: Towards desiderata for an ontology of diseases for the annotation of biological datasets. ICBO 39-42 (2009).
2. Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., Eppig, J.T. and the Mouse Genome Database Group.: The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Res. 39(suppl 1), D842--D848 (2011)
3. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), http://omim.org/ (2/9/2011)
4. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25(11):1251-5 (2007).
5. Osborne, J.D., Flatow, J., Holko, M., Lin, S.M., Kibbe, W.A., Zhu, L.J., Danila, M.I., Feng, G., Chisholm, R.L.: Annotating the human genome with Disease Ontology. B.M.C. Genomics 10, S1:S6 (2009)
6. Davis, A.P., King, B.L., Mockus, S., Murphy, C.G., Saraceni-Richards, C., Rosenstein, M., Wiegers, T., Mattingly, C.J.: The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Res. 39(Database issue), D1067--72 (2011)
7. Medical Subject Headings, http://www.nlm.nih.gov/mesh/