Two approaches to integrating phenotype and clinical information

Anita Burgun, MD, PhD[1], Fleur Mougin, PhD[2], Olivier Bodenreider, MD, PhD[3]

[1] INSERM U936, School of Medicine, University of Rennes 1, IFR 140, Rennes, France;
[2] LESIM, INSERM U897, ISPED, University Victor Segalen Bordeaux 2, France;
[3] US National Library of Medicine, NIH, Bethesda, USA
anita.burgun@univ-rennes1.fr

Abstract

Linkages between animal models of diseases and human data enable the development of translational research hypotheses. The objective of this study is to investigate two approaches to integrating phenotype and clinical information. On the one hand, we develop a terminology mapping between phenotypes from the Mammalian Phenotype Ontology (MPO) and Online Mendelian Inheritance in Man (OMIM) through the Unified Medical Language System (UMLS). On the other, we associate MPO phenotypes with OMIM manifestations through annotations made to orthologous genes. 1,469 MPO concepts (22%) were mapped successfully to some disease concept in the UMLS, of which 869 were present in OMIM. Among the 16,764 distinct MGI genes associated with human orthologs, 1,968 distinct genes were associated with both MPO and OMIM annotations. The UMLS is a valuable resource for linking phenotype terms to clinical terminologies, and these mappings between terminologies can help enrich gene annotation databases and unify phenotype representation.

Introduction

Animal models of diseases provide valuable insights into the pathogenesis of human diseases. The mouse is an excellent animal surrogate for studying normal development and disease processes in humans. Linkages between mouse and human data enable the generation of translational research hypotheses based on comparative genotype, phenotype and functional analyses. In addition, rather than conducting research in isolation, researchers are now asked to take a broader approach and to consider not only the primary focus of their own research, but also its link to patient health care. Therefore, data integration between basic research data and clinical data is a prerequisite to many translational research activities (e.g., [1-3]) Heterogeneity in phenotype coding is a key issue in integrating translational research data, which are typically generated independently by experimental researchers and bedside physicians. For example, in the Mouse Genome Database, which collects genetic and genomic information about the laboratory mouse, descriptions of mouse phenotypes are coded using the Mouse Phenotype Ontology (MPO). On the other hand, different clinical terminologies are used for the annotation of phenotypes in various datasets, including the Medical Subject Headings (MeSH) for the literature, the NCI Thesaurus (NCIt) for clinical research in oncology, SNOMED CT for patient records, the International Classification of Diseases (ICD) for registries, and Online Mendelian Inheritance in Man (OMIM) for human genetics records.

The objective of this study is to investigate two approaches to integrating phenotype and clinical information. On the one hand, we develop a terminology mapping between phenotypes from the Mammalian Phenotype Ontology (MPO) and Online Mendelian Inheritance in Man (OMIM) through the Unified Medical Language System (UMLS). On the other, we associate MPO phenotypes with OMIM manifestations through annotations made to orthologous genes.

Background

The **Mammalian Phenotype Ontology** (MPO) is developed and used in phenotypic data annotation in the Mouse genome Informatics (MGI) system at the Jackson Laboratory [4]. Other users or collaborators include Rat Genome Database, Mouse Mutagenesis Centers, and Online Mendelian Inheritance in Animals. The February2009 release is available in OBO file format from the MGI website [5],It contains 6,548 concepts that represent phenotypes, corresponding to 15,888 terms. For example four synonymous terms are listed for the concept MP:0000410: *waved hair*, *curly hair*, *waved fur* and *wavy hair*. Phenotypes from MPO include diseases (e.g., MP:0003561 *Rheumatoid arthritis*), normal characteristics (e.g., MP:0000410 *waved hair*) and clinical features (e.g. MP:0001261 *obese*), as well as biological characteristics and test results (e.g., MP:0000218 *increased leukocyte cell number*,).

Online Mendelian Inheritance in Man (OMIM) is a knowledge base of human genes and related phenotypes created at Johns Hopkins University and available electronically through the Entrez system from the National Center for Biotechnology Information (NCBI) [6]. OMIM contains about 19,000 entries, of which 6,500 correspond to descriptions of clinical phenotypes. Each entry is identified by a unique number (e.g., *113705 for *Breast cancer 1 gene* and #277900 for *Wilson disease*). In addition to textual descriptions for genes and phenotypes, OMIM provides two additional resources summarizing the relationship between genes, diseases and manifestations. On the one hand, the Morbid Map presents the cytogenetic map location of 3,800 disease genes described in OMIM (*Wilson disease* is caused by mutation in the ATP7B gene located at locus 13q14.3-q21.1). On the other, the clinical synopses relate some 4,500 disorders described in OMIM to their clinical manifestations (e.g., *High urinary copper* is a manifestation of *Wilson disease*).

The **Unified Medical Language System**® (UMLS®) Metathesaurus® contains about 1.8 million concepts, which are clusters of synonymous terms coming from almost 150 sources vocabularies. The UMLS concepts are identified by a Concept Unique Identifier or CUI, and more than 36 million relationships between these concepts are present. As such, the UMLS plays a central role in connecting vocabularies, and sharing annotated data [7]. Each Metathesaurus concept is assigned at least one semantic type. Groupings of semantic types, called semantic groups (SGs), represent subdomains of biomedicine such as Anatomy, Chemicals & Drugs, and Disorders [8]. Each semantic type belongs to one and only one SG. Genes and disorders from OMIM are integrated in the UMLS, as well as relations from the clinical synopses.

Methods

The two approaches to integrating phenotype and clinical information under investigation are terminology mapping and mapping through gene annotations.

Terminology mapping

Unlike the vocabulary from OMIM, terms from MPO have not been integrated in the UMLS Metathesaurus. Therefore no direct correspondence between MPO and OMIM can be established through the UMLS. However, MPO terms can be mapped to the UMLS in order to find lexically similar terms and synonyms in clinical vocabularies, including OMIM. Towards this end, we attempted to map all MPO terms to the UMLS, first directly and, in the absence of a direct match, after removing frequent modifiers from the MPO terms. Version 2008AA of the UMLS is used in this study.

Direct mapping. The names of MPO concepts are searched in the UMLS Metathesaurus using exact match and normalized match. In order to prevent the mapping of phenotypes from MPO to semantically irrelevant concepts, the UMLS concepts mapped to are restricted to concepts from SG Disorders.

Mapping after demodification. The MPO concepts that remained unmapped after direct mapping were analyzed. We hypothesized that these terms can be mapped not to pre-coordinated phenotype terms as with the direct mapping, but rather to terms that can be later post-coordinated into a phenotype term. Starting with the list of MPO terms, we performed a syntactic analysis of the MPO terms and identified modifiers that are frequently present in MPO terms. We focused on terms starting with modifiers followed by word phrases. After a modifier is removed from the term, the remaining substring forms the 'demodified term'. For example *abnormal ethmoidal bone* is constituted of the modifier abnormal followed by the demodified term *ethmoidal bone*. A list corresponding to the 100 most frequent modifiers was established. These modifiers were removed from the MPO terms. The demodified MPO terms were mapped to the UMLS Metathesaurus, using exact match and normalized match as before. The SGs were used here, not to restrict the mapping, but to characterize the demodified terms semantically.

Mapping through gene annotations

We integrated phenotypic annotations from two databases, one for the laboratory mouse, MGI, in which phenotypic information is coded with MPO, and one for the Human species, OMIM, in which phenotype coding uses the OMIM list of genetic diseases. Figure 1 gives an overview of the mapping process and shows the resources utilized. The version of these resources used in this study was downloaded in February 2009.

The HMD_HumanPhenotype resource from MGI provides Mouse/Human orthology with phenotype annotation of the mouse genes to MPO. More precisely, it provides the following information: gene symbols (in mouse and human) and MPO IDs corresponding to phenotype annotation. For example, the mouse gene *Brca1* is identified in MGI as 'MGI:104537', it is orthologous to the human gene *BRCA1*, and is associated with sixteen MPO phenotypes, including *tumorigenesis* (MP:0002006).

MGI also provides HMD_OMIM, an information source about Human and Mouse orthologs with OMIM gene IDs. For example the human gene

orthologous to the mouse gene *Brca1* is identified by '113705' in OMIM.

The OMIM Morbid Map is an alphabetic list of diseases described in OMIM and their corresponding genes represented as OMIM gene IDs, gene symbols, and chromosomal locations. For example, in the OMIM Morbid Map, *BRCA1* is associated with the OMIM disease ID '#604370', which corresponds to *Breast-ovarian cancer, familial, susceptibility to,1*.

Associations between OMIM diseases and their manifestations are extracted from the OMIM Clinical Synopses. No clinical synopsis is provided for *BRCA1*.

The method exploiting the four resources listed above can be summarized as follows. For each pair of Mouse/Human orthologs, we extract (1) the MPO phenotypes annotations associated with the mouse gene in MGI (from HMD_HumanPhenotype), (2) the OMIM gene ID associated with the human ortholog in MGI (from HMD_OMIM). The OMIM gene is associated with OMIM diseases through the Morbid Map and to manifestations through the Clinical Synopses. Finally, the terminology mapping is utilized to assess the equivalence of phenotypes between MPO and OMIM.

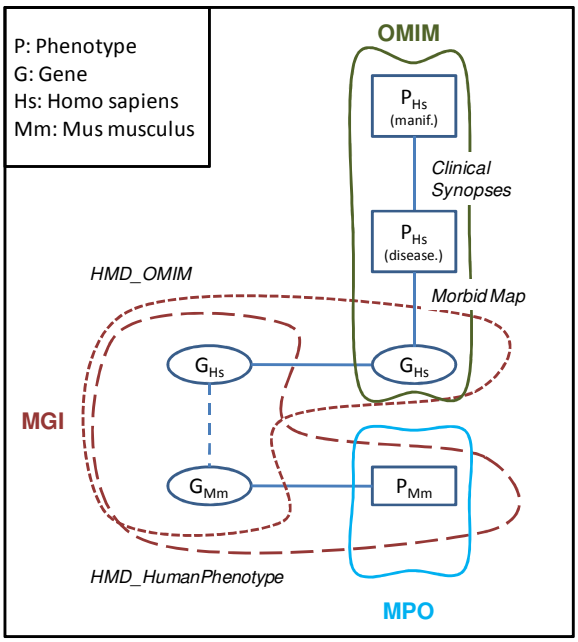

Figure 1. Mapping through gene annotations: Overview of the methods

Results

Terminology mapping

Direct mapping. 2,167 MPO terms (14%), and 1,535 MPO concepts (24%) were mapped to UMLS concepts. Among them, 1,469 MPO concepts were mapped to concepts categorized as Disorders in the UMLS. Examples of MPO concepts mapped successfully include:

- MP:0000062 *increased bone density* mapped to C1141880 *Bone density* increased (normalized match);
- MP:0000081 *craniostosis* mapped to C0010278 *Craniosynostosis* using synonyms in MPO (exact match);
- MP:0000061 *brittle bones* mapped to C0029434 *Osteogenesis Imperfecta*, using synonyms in UMLS (exact match).

Among the 1,469 concepts mapped successfully to some UMLS concept in Disorders, 1,176 were present in SNOMED CT, 1,001 in MedDRA, 869 in OMIM, 317 in ICD10, 654 in MeSH, and 524 in the NCI Thesaurus.

As expected, many of the 13,721 unmapped terms (5,013 concepts) include a modifier. For example, MP:0000100 *abnormal ethmoidal bone*, MP:0000101 *absent ethmoidal bone*, MP:0000687 *small lymphoid organs*, and MP:0000689 *abnormal spleen structure*.

Mapping after demodification. The 100 most frequent modifiers present in MPO terms include *abnormal*, *absent*, and *small*. After removing the most frequent modifiers from the 10,779 terms, we obtained a list of 8,453 unique demodified terms for which a direct mapping to the UMLS was attempted.

2,708 (20%) of 13,721 previously unmapped MPO terms were successfully mapped to some UMLS concept after demodification. They correspond to 1,812 unique terms (21%), and 1,754 MPO concepts (out of 5,013, 35%). MPO terms remaining unmapped after demodification include phenotypes specific to animals e.g., *snout shape abnormalities*, and terms related to biology such as *inability to present cytosolic antigens to Class-I restricted cytotoxic T cells*.

According to their categorization with SGs, demodified terms correspond mostly to anatomical structures (SG: Anatomy) and physiology processes (SG: Physiology). About 1,500 terms correspond to the pattern "<modifier> followed by <anatomical structure>", including MP:0000005 *increased brown fat* mapped to the UMLS Metathesaurus concept C0006298 *Brown Fat* (SG: Anatomy). More than 500 terms correspond to the pattern "<modifier> followed by <physiological process>", including MP:0000057 *abnormal osteogenesis* mapped to C0029433 *Osteogenesis* (SG: Physiology).

Mapping through gene annotations

16,764 distinct MGI gene identifiers are associated with human orthologs. Among them, 11,330 are present in OMIM. 5,174 genes present in both MGI and OMIM are associated with at least one MPO phenotype. For example, MGI:99607 corresponds to *ABCA1* and is annotated with sixteen MPO concepts.

We obtained from the Morbid Map a list of 3,754 genes with OMIM disease annotation. 1,968 distinct genes were associated with both MPO annotation and OMIM annotation. However, none of them was associated with both MPO concepts and OMIM codes that mapped exactly the same UMLS concepts. For example, MGI:101875 *TPHO* is associated with two MPO phenotypes, MP:0005387 *immune system phenotype* and MP:0005397 *hematopoietic system phenotype*, and one OMIM code '#187950' which corresponds to C0040028 *Thrombocythemia, Essential*, but no mapping exists between the MPO terms and the OMIM one.

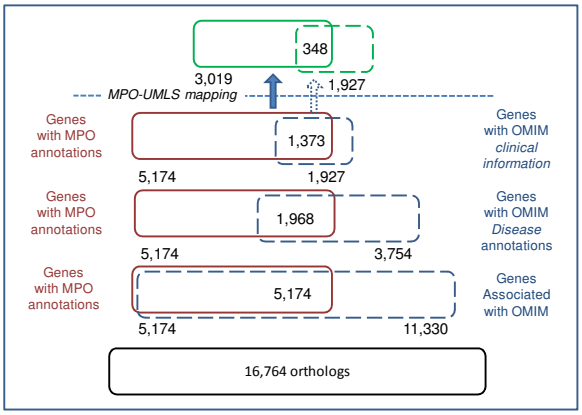

Figure 2. Mapping through gene annotations: Results

Among the 3,754 genes associated with OMIM disease annotation, clinical information could be found for 1,927 genes in the Morbid Map or from the Clinical Synopses. 1,373 genes have both clinical manifestations in OMIM and MPO annotations.

At least one MPO annotation could be mapped to the UMLS for 3,019 distinct MGI gene IDs. For example, *ABCA1* is associated with MP:0005381 *digestive/alimentary phenotype* which was mapped successfully to the UMLS concept C1559265 *Adverse Event Associated with the Gastrointestinal System*. Because OMIM is integrated in the UMLS, all clinical annotations from OMIM were found in the UMLS. Of the 1,373 genes with both clinical annotations from OMIM and MPO, 348 had annotations mapped to UMLS concepts from both MPO and OMIM, although not necessarily the same annotations. One such gene is Hfe discussed in the following section.

Discussion

Example of application

The mappings we generated can be used to enrich databases. Consider the mouse gene Hfe (MGI:109191). It is annotated by 9 MPO phenotypes, including *digestive/alimentary phenotype* (MP:0005381) and *liver/biliary system phenotype* (MP:0005370). For the orthologous human gene, HFE, OMIM provides clinical annotations such as *Esophagitis*, which maps to C0014868 in the UMLS and to *esophageal inflammation* (MP:0003278) in MPO. Similarly, *liver cirrhosis* in OMIM maps to C0023890 in UMLS and to *liver cirrhosis* (MP:0003046) in MPO. It is therefore possible to enrich phenotype annotations in MGI by adding more precise clinical features from OMIM. More generally, as soon as a phenotype from a source database maps to some phenotype in the terminology target database, it is possible to not only add the annotation from the target database to the source database, but also to "translate" it automatically into the own vocabulary of the source database.

Compositionality of phenotype terms

The Phenotype And Trait Ontology (PATO) [9] is an ongoing effort to provide terms for annotating phenotype information across species. PATO relies on post-composition: phenotype coding combines one entity (E), which can be an anatomical structure, a biological process, or a physiological function, and the qualities (Q), qualitative or quantitative, ascribed to this entity. In contrast, MPO terms are pre-coordinated terms. For example, MP: 0005141 *Liver Hyperplasia* in MPO corresponds to the combination of two terms in the EQ model, one for the quality PATO:0000644 *hyperplastic*, and the other one for the entity MA:0000358 *liver*. Not surprisingly, our results of mapping MPO terms to UMLS after de-modification confirm that many pre-coordinated terms from MPO combine one modifier (equivalent to qualities in PATO) and one Anatomy and Physiology concept (equivalent to entities in PATO). The method used in this study illustrates the feasibility of extracting the entity component from the MPO term. Further work is needed to take into account the semantics carried by the modifiers in the MPO terms and to create post-coordinated expressions for phenotypes from these entities.

Overlap with clinical terminologies

The degree of overlap between phenotype terminologies used in basic research and medical terminologies used for coding clinical and epidemiological data is a crucial determinant of translational research**.** The highest coverage was obtained for SNOMED CT (80%), The lowest score was obtained for ICD (21%).

Terminology mapping and data integration

Some characteristics of MPO hinder its mapping to UMLS. For example, spatial or anatomical concepts and disease names are sometimes represented as synonyms, e.g., *heart*, *cardiac*, *cardiovascular*, *cardiovascular system*, and *circulatory* are synonyms of *cardiovascular system phenotype* (MP:0005385) in MPO. Restricting mappings to UMLS concepts from the SG Disorders helped us prevent inaccurate mappings to other categories, such as Spatial Concepts. For example C0521362 *Gastrointestinal* is present in OMIM and related to *gastrointestinal phenotype* in MPO but it is categorized as Spatial Concept in the UMLS. Moreover, ambiguity between gene names and disease names is frequent. For example 235200 in OMIM corresponds to *HFE gene*, as well as *Hemochromatosis, hereditary*, which is the disease associated with HFE. No manual validation was performed in this study. While semantic group restriction prevents many inaccurate mappings, manual validation may be required to confirm whether fine-grained mappings are correct. For example the mapping (through normalized match) between *Absent eyes* (MP:0001293) and C0339054 *Acquired anophthalmos* (rather than C0003119 *Anophthalmos*) is arguable, despite the fact that '*eye; absent*' is listed as a UMLS synonym for *Acquired anophthalmos*, not *Anophthalmos*.

Phenotypes differ from diseases in two aspects: (i) phenotype is the set of observable characteristics of an organism, including physical, developmental, behavioral and biochemical characteristics, while diseases are more complex entities (ii) phenotypes may correspond to the absence of diseases as well as to the presence of a disease. MGI phenotypes mostly correspond to observable characteristics while most OMIM diseases represent complex syndromes related to clinical manifestations. Not surprisingly, we did not capture much redundant information between MGI and OMIM, in terms of ortholog genes associated with the same phenotypes.

The UMLS can be expected to play a crucial role in unifying phenotypic information associated with orthologous genes. From the perspective of a given database, it makes it possible to integrate annotations extracted from another database and to code them using one's own vocabulary. More generally, unification provided by the UMLS enables the integration of phenotypic annotations coming from heterogeneous sources. While pairwise mappings are only applicable to a fixed pair of terminologies, terminology integration in the UMLS supports data integration for a large number of terminologies, including, as we showed with MPO, when these terminologies are not themselves integrated in the UMLS. This approach can be used not only to get equivalent terms in different terminologies but also to address granularity issues. For example, MP:0005397 *hematopoietic system phenotype*, and OMIM '#187950' *Thrombocythemia, Essential* both annotate *TPHO*. Although not equivalent, these two terms share several characteristics. Further work is needed to take into account subsumption relations, and possibly other kinds of relations, such as manifestations of diseases, in mapping phenotype across terminologies.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Mullen SA, Scheffer IE. Translational research in epilepsy genetics: sodium channels in man to interneuronopathy in mouse. Arch Neurol 2009;66(1):21-6
2. Sivamani RK, Pullar CE, Manabat-Hidalgo CG, Rocke DM, Carlsen RC, Greenhalgh DG, et al. Stress-mediated increases in systemic and local epinephrine impair skin wound healing: potential new indication for beta blockers. PLoS Med 2009;6(1):e12
3. Speakman J, Hambly C, Mitchell S, Krol E. The contribution of animal models to the study of obesity. Lab Anim 2008;42(4):413-32
4. Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 2005;6(1):R7
5. Mouse Genome Informatics (MGI) Web: http://informatics.jax.org
6. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 2009;37(Database issue):D793-6
7. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-70
8. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform 2001;84(Pt 1):216-20
9. PATO - Phenotypic Quality Ontology: http://obofoundry.org/wiki/index.php/PATO:Main_Page