Proceedings of Medinfo 2010 (in press).

A Framework for Comparing Phenotype Annotations of Orthologous Genes

Olivier Bodenreider[a], Anita Burgun[b]

[a] *US National Library of Medicine, NIH, Bethesda, USA*
[b] *Division INSERM U936, School of Medicine, University of Rennes 1, IFR 140, Rennes, France*

Abstract

Objectives: *Animal models are a key resource for the investigation of human diseases. In contrast to functional annotation, phenotype annotation is less standard, and comparing phenotypes across species remains challenging. The objective of this paper is to propose a framework for comparing phenotype annotations of orthologous genes based on the Medical Subject Headings (MeSH) indexing of biomedical articles in which these genes are discussed.* ***Methods:*** *17,769 pairs of orthologous genes (mouse and human) are downloaded from the Mouse Genome Informatics (MGI) system and linked to biomedical articles through Entrez Gene. MeSH index terms corresponding to diseases are extracted from Medline.* ***Results:*** *11,111 pairs of genes exhibited at least one phenotype annotation for each gene in the pair. Among these, 81% have at least one phenotype annotation in common, 80% have at least one annotation specific to the human gene and 84% have at least one annotation specific to the mouse gene. Four disease categories represent 54% of all phenotype annotations.* ***Conclusions:*** *This framework supports the curation of phenotype annotation and the generation of research hypotheses based on comparative studies.*

Keywords

Phenotype, Comparative Study, Medical Informatics Computing, Medical Subject Headings.

Introduction

Scientific discoveries for improvement of human health typically begin "at the bench" with basic research, then progress to the clinical level. Studies using animal models play an important role in the generation of research hypotheses based on comparative genotype, phenotype and functional analyses, later tested in human clinical studies [1, 2]. The growing importance of animal models in translational research has given rise to the development of large, species-specific knowledge bases containing curated functional and phenotype annotations, in addition to gene sequences [e.g., 3, 4, 5].

Tools for comparing and contrasting annotations of orthologous genes have started to emerge as part of portals developed by model organism communities (see, for example, the graphs representing the functional annotations of genes across species on the Mouse Genome Informatics system[1]). The Gene Ontology has brought standardization to the functional annotation of gene products, making it relatively easy to compare functions across species. In contrast, phenotype annotation is less standard and the comparison of phenotypes across species remains challenging [6].

Data mining approaches to identifying gene-phenotype associations have been used [7] and orthology has been exploited in some studies [8]. In contrast to the word-based approach in [8], we specifically take advantage of the hierarchical structure of MeSH in order to aggregate diseases into high-level categories to facilitate the analysis of the results.

The objective of this paper is to propose a framework for comparing phenotype annotations of orthologous genes based on the MeSH indexing of biomedical articles in which these genes are discussed. The framework is applied to mouse and human orthologs. We outline two possible applications of this work, namely to support the curation of phenotype knowledge bases and to help researchers with hypothesis generation.

Materials

Several publicly-available data sources are used to relate phenotype annotations across species. Mouse to human orthology information is acquired from MGI. The association between genes and biomedical articles is provided by Entrez Gene. Finally, MeSH index terms corresponding to diseases are extracted from Medline.

MGI

The Mouse Genome Informatics (MGI) system is a knowledge base about the laboratory mouse developed at the Jackson Laboratory [4]. Among the information provided is mammalian orthology and, in particular, curated and uncurated lists of human genes corresponding to mouse genes. In addition to gene identifiers in the MGI system, cross-references to Entrez Gene identifiers are provided.

[1] http://www.informatics.jax.org/function.shtml

Entrez Gene

Entrez Gene is the gene-centric knowledge base developed at the National Center for Biotechnology Information (NCBI) [9]. In addition to basic information about genes (e.g., symbols and names), Entrez Gene provides cross-references to other components of the Entrez system and to external resources. In particular, Entrez Gene compiles bibliographic references for each gene (with links to PubMed). These references are specific to a given gene cited as evidence for functional annotations, from the GeneRIFs[2], protein databases and model organism databases (GOA[3] for human genes and MDG[4] for mouse genes).

Medline and MeSH

Medline is a bibliographic database developed at the US National Library of Medicine, containing over 16 million citations. Medline citations are indexed with terms from the Medical Subject Headings (MeSH) thesaurus and are available through PubMed in the Entrez system.

The current version of MeSH contains 25,186 descriptors, 4,409 of which correspond to diseases (somatic and mental disorders). The hierarchical structure of MeSH makes it easy to aggregate diseases into top-level disease categories (24 including mental disorders).

Methods

Acquiring phenotype information

Starting from the list of orthologs, we use links established between Entrez Gene and PubMed to retrieve Medline citations for each gene, from which we extract the MeSH index terms corresponding to diseases. An overview of the links among resources is presented in Figure 1.

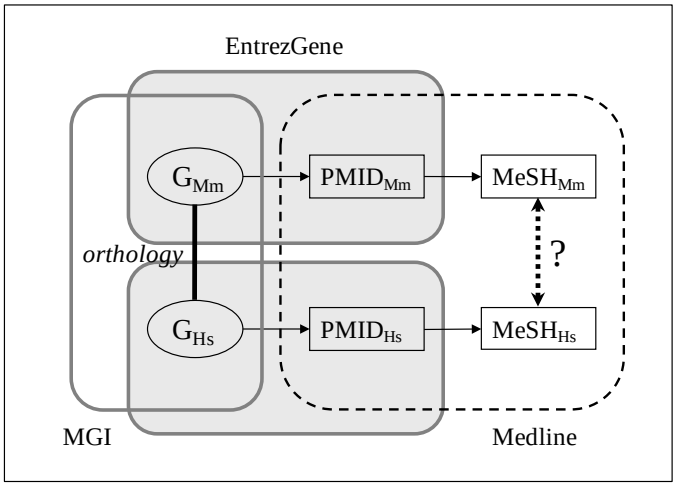

Figure 1. Overview of links among resources

[2] http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html
[3] http://www.ebi.ac.uk/GOA/
[4] http://www.informatics.jax.org/mgihome/projects/overview.shtml

List of orthologous genes

The list of human genes orthologous to mouse genes was downloaded from MGI[5]. Of the 17,796 pairs of genes, 24 were eliminated because no cross-reference to Entrez Gene was provided. Three additional pairs were discarded, in which multiple human orthologs were listed for a given mouse gene. In this case, the curated orthology pair was selected over the computed one. A total of 17,769 pairs of genes remained. Entrez Gene identifiers were available for all 35,538 genes.

Bibliographic references for genes

The association between genes from Entrez Gene and Medline citations from PubMed was queried from the Entrez system using the Entrez Programming Utilities[6]. More specifically, we used the *elink* service to retrieve links from the *gene_pubmed* table. The resulting XML file was processed with XSLT transformation in order to extract pairs of identifiers for the gene (Entrez Gene ID) and the related Medline citations (PubMed identifier or PMID). At least one bibliographic reference was available for all but 552 genes. Orthologous human and mouse genes were associated with a total of 357,008 unique Medline citations.

Disease categories for genes

These 357,008 Medline citations were downloaded in XML format using the *efetch* service and post-processed with XSLT transformation in order to extract the MeSH descriptors from each citation. A total of 5,862,632 MeSH descriptors were extracted and mapped to their identifiers and top-level categories using a local MeSH database. The position of each MeSH descriptor in the hierarchy is indicated by a "tree number". The first node of the tree number refers to the top-level of the hierarchy. For example, the tree number for the descriptor *Megacolon* is C06.405.469.158.701, indicating that it is part of the category *Digestive System Diseases* (C06). Diseases can be related to more than one category.

Linking genes to disease categories

As shown in Figure 1, the various datasets under investigation all share some identifiers: Entrez Gene identifier between orthology data and Entrez Gene, PubMed ID between Entrez Gene and Medline, and MeSH identifiers between Medline and the MeSH thesaurus. These shared identifiers make it possible to chain the various datasets in order to create a link between a given gene and the MeSH disease categories through bibliographic references for this gene. The MeSH disease categories for a given mouse gene can easily be compared to the MeSH disease categories for its human ortholog.

We used Semantic Web technologies (RDF, the Resource Description Framework and related tooling) to process these datasets, because it is well adapted to handling linked data. In practice, the various datasets were converted to "N triple" format. A total of 1.4 million triples were generated and loaded in the open source triple store Virtuoso[7]. We used the

[5] ftp://ftp.informatics.jax.org/pub/reports/HMD_Human5.rpt
[6] http://eutils.ncbi.nlm.nih.gov/corehtml/query/static/eutils_help.html
[7] http://virtuoso.openlinksw.com/

RDF query language SPARQL to query the triple store (Figure 2). Two queries were used to retrieve all MeSH disease categories for each gene in a given pair of orthologous genes. A total of 106,454 such associations were exported for further analysis.

```
PREFIX mesh:   <http://nlm.nih.gov#MeSH>
PREFIX pubmed: <http://www.ncbi.nlm.nih.gov/PUBMED:>
PREFIX gene:   <http://www.ncbi.nlm.nih.gov/GENE:>
PREFIX mor:    <http://mor.nlm.nih.gov#MOR:>
select distinct ?gene_h ?gene_m ?species ?mesh_name_h
from <http://mor.nlm.nih.gov/PHENOTYPE>
where {
       ?gene_h mor:orthologous_with ?gene_m .
       ?gene_h mor:species ?species .
       ?gene_h mor:described_in ?pmid_h .
       ?pmid_h mor:indexed_with ?mesh_h .
       ?mesh_h mor:has_category ?dis_cat_h .
       ?dis_cat_h rdfs:label ?mesh_name_h
}
```

Figure 2. SPARQL query for retrieving associations between human genes and disease categories

Comparing phenotype information

Two major types of analyses were carried out, from the perspective of genes and from that of phenotypes. For each pair of orthologous genes, we compared the vectors of disease categories between the two genes in the pair in order to determine if phenotype annotations tend to be the same at the level of disease categories. We considered that a gene was associated with one disease category as soon as the gene was associated with one disease from this category. In practice, we recorded which disease categories were common to the two genes in the pair and which were specific to each species. (No similarity metrics were used to compare the vectors of disease categories for pairs of orthologous genes).

For each disease category, we computed the number of pairs of orthologs for which both genes are annotated to this disease category and, conversely, the number of pairs for which the phenotype annotation was specific to one species.

Results

Of the 17,769 pairs of orthologs under investigation, 11,111 exhibited at least one phenotype annotation for each gene in the pair. The remaining pairs were discarded because no phenotype annotation was available for the mouse gene (1241 pairs), the human gene (3625 pairs) or both genes (1792 pairs).

Perspective of genes

Overall, 26,459 disease categories were common to both orthologous genes, 21,854 were specific to human genes and 20,032 were specific to mouse genes. The number of disease categories per pair of orthologs ranges from 0 to 15, with a median of 2 disease categories common to both genes. There is a median of 2 disease categories specific to the human and mouse orthologs. The percentage of disease categories in common between human and mouse genes is 55% for human genes and 57% for mouse genes. Details of the distribution are shown in Figure 3.

Among the 11,111 pairs of orthologous genes with both human and mouse phenotype annotations, 81% have at least one one phenotype annotation in common, 80% have at least one phenotype annotation specific to the human gene and 84% have at least one phenotype annotation specific to the mouse gene.

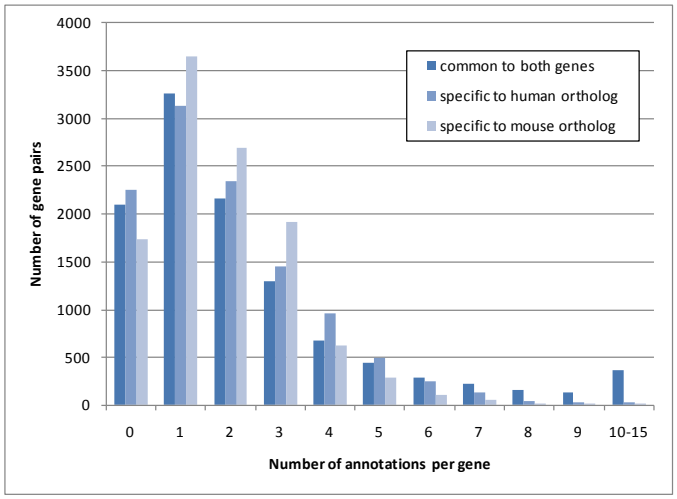

Figure 3. Distribution of the number of phenotype annotations per gene (for the 11,111 pairs of genes)

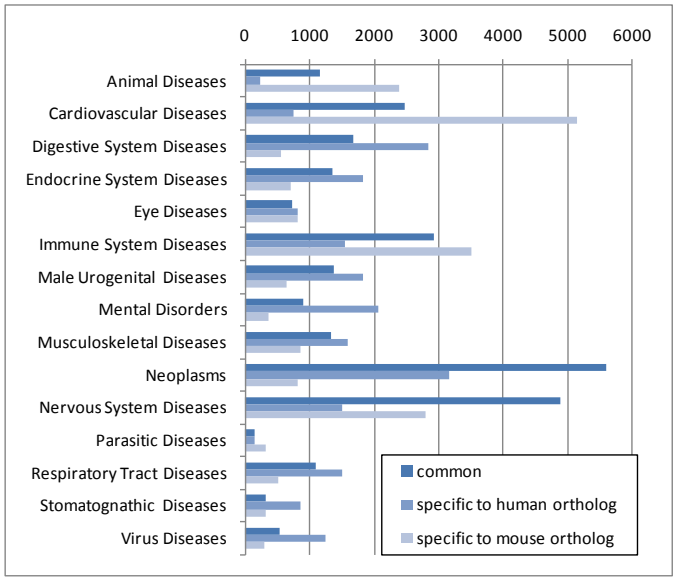

Figure 4. Distribution of the number of gene for each top-level disease category.

Perspective of phenotypes

Of the 24 top-level disease categories in MeSH (including mental diseases), 15 were reached through the phenotype annotations of the 11,111 orthologous genes. As shown in Figure 4, four disease categories represent 54% of all phenotype annotations. These are *Cardiovascular Diseases*, *Immune System Diseases*, *Neoplasms* and *Nervous System Diseases*. Within

each category, the repartition between common and species-specific annotations varies widely, with a majority of common annotations for *Neoplasms* and *Nervous System Diseases*, and mostly mouse-specific annotations for *Cardiovascular Diseases*.

Extended example

A clinical researcher interested in the molecular genetics of congenital hypothyroidism could go to Entrez Gene, type "congenital hypothyroidism" and get back such human genes as TPO, TSHR and TG. A translational researcher would easily find out that their mouse orthologs – Tpo, Tshr and Tg – are also associated with the same phenotype.

The translational researcher might be further interested in identifying mouse genes associated with congenital hypothyroidism for which the human ortholog is not. Such genes would indeed be potential candidates for association with congenital hypothyroidism in human.

Using our framework, we start from the MeSH term *Congenital Hypothyroidism* and identify the PubMed articles indexed to this term for which there is a link to a mouse gene. A list of 28 such mouse genes is obtained. We then use orthology relations from MGI to point to the human orthologs of these genes. Ten human orthologs including TPO, TSHR and TG are associated with *Congenital Hypothyroidism* through one or more PubMed articles and are discarded, because their link to the disease is already known. The remaining 18 genes are potentially novel human genes for congenital hypothyroidism.

Among the 18 genes is THRB, ortholog of the mouse gene Thrb associated with congenital hypothyroidism. Although no GeneRIF or functional annotation links THRB to any PubMed article indexed with *Congenital Hypothyroidism* (as of October 15, 2009), a search on PubMed for "THRB AND congenital hypothyroidism[mh]" yields one recently published article (PMID: 19318451) [10], indexed with *Congenital Hypothyroidism*, confirming the association between congenital hypothyroidism and THRB mutation. This example illustrates the lag between the publication of an article and the availability of the knowledge extracted from this article in model organism knowledge bases.

Finally, there are also 12 human genes associated with *Congenital Hypothyroidism*, for which the mouse ortholog is not. The 12 mouse genes could be investigated for potential association with congenital hypothyroidism in the literature in order to ensure the completeness of phenotype annotations for the mouse genes.

Discussion

Applications

Supporting phenotype annotation

By linking genes to MeSH phenotypes, this framework provides a tool to biocurators responsible for phenotype annotation. Starting from a given gene, it helps retrieve relevant articles based on the MeSH indexing of diseases. Curators responsible for a group of genes can easily scan new disease-related articles and flag them for review. Conversely, curators responsible for a group of diseases are provided with the list of genes in relation to which a given disease was discussed.

This study shows that there is a large reservoir of species-specific annotations, i.e., annotations to a given disease category in one species for which there is no equivalent in the ortholog. This information could be used for inferring phenotype annotation from sequence similarity in a way similar to functional annotations to GO "inferred from electronic annotation" (IEA).

Both supporting curation and transferring annotations from other species contribute to address the incompleteness of annotations [11].

Generating novel hypotheses

In the era of translational research, easy access to comparison of phenotype annotation between orthologs can help researchers generate novel hypotheses. Phenotype annotations specific to the mouse with potential therapeutic implications can be explored in humans. Conversely, the set of mouse-specific genes associated with a given disease in the mouse provides candidates for exploration in humans for the same disease. Moreover, annotations specific to the human species may be used to evaluate the molecular markers involved in processes in a mouse model that mimics essential elements of human diseases and to analyze mutant mice carrying specific alleles [12].

Limitations

In our present investigation, the level of granularity for phenotypes is limited to a few dozen disease categories, which is too coarse for a detailed comparison of phenotypes across species, as no single gene is responsible for cardiovascular diseases as a whole. This limitation can be easily addressed by our framework in which all MeSH disease terms are recorded before being aggregated into disease categories for analysis. The example presented earlier demonstrates that a disease such as congenital hypothyroidism can be analyzed individually. However, it might also be useful to extend the framework to physiological phenomena (e.g., bone growth) for phenotypes for which no clinical information is available. Such preclinical phenotypes would also be easier to align with annotations to the Mammalian Phenotype Ontology [13]. In future work, we also plan to investigate the use of semantic similarity measures among MeSH terms for the aggregation of diseases into finer-grained classes.

Generalization

In this paper, we focus on phenotype annotation of human and mouse genes. However, the framework we propose can easily be extended to other species. Orthology relations available from Entrez Homologene can be used in the absence of (or in complement to) curated orthology relations. GeneRIFs are recorded regardless of species and functional annotations to

the Gene Ontology (with reference to biomedical articles) are provided by several dozen model organism databases. Finally, MeSH indexing is available throughout Medline. Therefore, links between genes and disease categories can be established for virtually any species. As mentioned earlier, this framework would benefit from being extended to the annotation of physiologic phenomena in addition to diseases, especially for non-mammalian species, such as yeast and fruitfly, for which phenotype information may be recorded at a preclinical level.

Technological considerations

In our experience, Semantic Web technologies including RDF and triple stores greatly facilitate the integration of datasets, especially when shared identifiers are present across datasets. The SPARQL query language was useful for formulating basic queries (e.g., list the disease categories for each gene) and export the data, with satisfactory performance (under one minute). More complex queries could be formulated (e.g., list of the disease categories common to orthologous genes). However, the absence of straightforward mechanisms for expressing negation in SPARQL makes it difficult to formulate queries such as listing disease categories specific to one species. We resorted to simple *ad hoc* programming (including differences between files) in order to carry out the analysis of annotations specific to a given species.

Conclusions

In this paper, we propose a framework for comparing phenotype annotations of orthologous genes based on publicly-available resources integrated through Semantic Web technologies. We show that there is a large reservoir of species-specific phenotype annotations for mouse and human orthologs. The proposed framework supports the curation of phenotype annotation and the generation of research hypotheses based on comparative studies. In contrast to popular resources such as Entrez Gene or GeneWiki, this integrative framework not only supports navigation between gene and phenotype resources, but also the computational analysis of the annotations.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). Our thanks go to Ramez Ghazzaoui who helped set up Virtuoso.

References

[1] Butte AJ. Translational bioinformatics applications in genome medicine. *Genome Med* 2009;1(6):64

[2] Watters JW, McLeod HL. Murine pharmacogenomics: using the mouse to understand the genetics of drug therapy. *Pharmacogenomics* 2002;3(6):781-90

[3] Beck T, Morgan H, Blake A, Wells S, Hancock JM, Mallon AM. Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data. *BMC Bioinformatics* 2009;10 Suppl 5:S2

[4] Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res* 2008;36(Database issue):D724-8

[5] Shaw DR. Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr Protoc Bioinformatics* 2009;Chapter 1:Unit1 7

[6] Hancock JM, Mallon AM, Beck T, Gkoutos GV, Mungall C, Schofield PN. Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mamm Genome* 2009;20(8):457-61

[7] Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. *Nat Genet* 2002;31(3):316-9

[8] Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, et al. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* 2005;3(5):e134

[9] Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2009;37(Database issue):D5-15

[10] Borck G, Seewi O, Jung A, Schonau E, Kubisch C. Genetic causes of goiter and deafness: Pendred syndrome in a girl and cooccurrence of Pendred syndrome and resistance to thyroid hormone in her sister. *J Clin Endocrinol Metab* 2009;94(6):2106-9

[11] Baumgartner WA, Jr., Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 2007;23(13):i41-8

[12] Gondo Y, Fukumura R, Murata T, Makino S. Next-generation gene targeting in the mouse for functional genomics. *BMB Rep* 2009;42(6):315-23

[13] Burgun A, Mougin F, Bodenreider O. Two approaches to integrating phenotype and clinical information. *AMIA Annu Symp Proc* 2009:75-79

Address for correspondence

Olivier Bodenreider (olivier@nlm.nih.gov)

Anita Burgun (anita.burgun@univ-rennes1.fr)