Databases and ontologies

Metab2MeSH: annotating compounds with medical subject headings

Maureen A. Sartor1,2,†, Alex Ade1,‡, Zach Wright1, David States3, Gilbert S. Omenn1,2,3,4, Brian Athey1,2 and Alla Karnovsky1,2,*

1National Center for Integrative Biomedical Informatics, 2Department of Computational Medicine and Bioinformatics, 3Department of Internal Medicine, 4Department of Human Genetics and School of Public Health, University of Michigan, Ann Arbor, MI 48109 and 5University of Texas Health Science Center, Houston, TX 77030, USA

1INTRODUCTION

In recent years, many tools have been developed for automated annotation of genes, transcripts and proteins and linking them to published biomedical literature. The resulting information has proved valuable for a range of bioinformatics tools and research to published biomedical literature. The resulting information has proved valuable for a range of bioinformatics tools and research to published biomedical literature. However, tools attempting to link small molecules to normal and diseased physiology and published data relevant to biologists and clinical investigators, are still lacking. With metabolomics rapidly emerging as a new omics field, the task of annotating small molecule metabolites becomes highly relevant. Our tool Metab2MeSH uses a statistical approach to reliably and automatically annotate compounds with concepts defined in Medical Subject Headings, and the National Library of Medicine’s controlled vocabulary for biomedical concepts. These annotations provide links from compounds to biomedical literature and complement existing resources such as PubChem and the Human Metabolome Database.

Availability: http://metab2mesh.ncibi.org

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 13, 2011; revised on March 6, 2012; accepted on March 29, 2012

ABSTRACT

Summary: Progress in high-throughput genomic technologies has led to the development of a variety of resources that link genes to functional information contained in the biomedical literature. However, tools attempting to link small molecules to normal and diseased physiology and published data relevant to biologists and clinical investigators, are still lacking. With metabolomics rapidly emerging as a new omics field, the task of annotating small molecule metabolites becomes highly relevant. Our tool Metab2MeSH uses a statistical approach to reliably and automatically annotate compounds with concepts defined in Medical Subject Headings, and the National Library of Medicine’s controlled vocabulary for biomedical concepts. These annotations provide links from compounds to biomedical literature and complement existing resources such as PubChem and the Human Metabolome Database. Several tools geared towards chemists such as Compounds In Literature (CIL) and WENDI, allow users to search compounds using names, smiles strings or compound structures and identify relevant publications (Gruning et al., 2011; Zhu et al., 2010).

Rather than linking compounds to all available literature, we decided to utilize Medical Subject Headings (MeSH) to annotate compounds with the most commonly associated medical/biological terms and vice versa. We believe that this approach helps to provide more focused results and aids in finding relevant relationships and literature. MeSH is the NLM controlled vocabulary used to manually index articles for MEDLINE/PubMed. MeSH covers a broad range of biological topics and medical terms, with such general headings as Disease, Anatomy, Chemicals and Drugs, and Phenomena and Processes. There are currently 26 142 MeSH descriptors that are organized into a hierarchical structure. In addition, there are 83 MeSH qualifiers that can be used to narrow a descriptor to a more specific topic.

NLM provides a list of chemical substances that contain a wide range of synonyms used in biomedical publications that can be linked to the compound synonyms used in PubChem. Using the compounds and their occurrences in PubMed literature, we performed statistical tests to estimate the significance of the associations between compound—MeSH descriptor pairs. The resulting associations, supporting data and literature can be accessed through our web-based tool, Metab2MeSH. Web services are also available (http://ws.ncibi.org/m2m.html) and allow for batch queries. Users have the ability to identify either the MeSH descriptors significantly associated with a given compound or the compounds associated with a selected MeSH descriptor. Examples of both use cases are provided below.

2METHODS

The PubChem Compound and Substance databases and NLM Pubmed database were downloaded, parsed and loaded into a relational database. Pubmed substances were extracted from the substance index for each researchers and as input data for other tools, that provides access to 30 million compounds and 85 million substances, and includes biomedical, structure, safety and additional information, including links to Pubmed abstracts via synonym mapping (Wang et al., 2009). ToxNet is another tool, developed by National Library of Medicine (NLM), that allows searches for compounds, and provides links to relevant literature through multiple sources (Wexler, 2001).
Anthropology, Education, Sociology and Social Phenomena, Technology, Industry, and Agriculture. This resulted in a total of 21,818 unique MeSH descriptors. Next, we determined the total number of PubMed articles annotated with at least one of these MeSH descriptors and at least one PubMed substance that maps to a PubChem synonym (6 million), which served as the complete background set of articles for our tests. We also determined the PubMed articles annotated with each MeSH descriptor and each PubMed substance, and the total number of articles for each compound.

Using these values and the numbers of PubMed articles co-annotated with each compound—MeSH term pair using two-sided Fisher’s exact tests, and p-values were calculated. This resulted in a total of 42 million tests, with 6,646,000 significant associations (p-value < 0.0001). Significant associations are presented in Metab2MeSH sorted by p-value. The False Discovery Rate method was used to adjust p-values for multiple comparisons, and the resulting q-values are provided (Benjamini and Hochberg, 1995). (In certain instances, the association is so significant that the precise p-value cannot be calculated. In these cases, the p-value is set to 0 and the χ² statistic from the equivalent χ² test is used to sort the associations by decreasing significance.) As an alternative measure of strength of association, we provide fold change, defined for a specific compound and MeSH descriptor as follows:

Fold Change = \( \frac{p(c/d)}{p(d/c)} \),

where \( p \) is the number of articles that have the compound and MeSH descriptor pair, \( c \) is the total number of articles that contain the compound, \( d \) is the total number of articles that contain the MeSH descriptor and \( t \) is the total number of articles considered. We combined the above significant results with links to the literature and PubChem compounds in a web interface to create our tool, Metab2MeSH (http://metab2mesh.ncbi.nlm.nih.gov). The data are updated twice a year. Supplementary Materials and Supplementary Figure S1 contain further details about the web interface. The Metab2MeSH data set can also be accessed through our web services (http://ws.ncbi.nlm.nih.gov/m2m.html) which allows users to query the Metab2MeSH database programmatically, and perform batch queries if desired. The resulting computer-parsable XML data include compound-MeSH term pairs with significance values and a list of associated PMIDs (see Supplementary Material for an example web services bulk query).

### 3 RESULTS

To illustrate the uses and features of Metab2MeSH, we present two applications. Suppose we would like to determine whether cyclic AMP (cAMP) is commonly associated with any certain type of cancer. We can query the compound ‘cyclic AMP’ in Metab2MeSH using an exact match search and filter the MeSH headings by Disease to see that 29 diseases are found to be significantly associated with cAMP among the top 1000 displayed. To see the complete list of results (>1000), we can download the tab-delimited text file to view in a spreadsheet. Supplementary Figure S2 shows the Metab2MeSH results for the top diseases. The two diseases most strongly associated with cAMP levels are pseudohypoparathyroidism and neuroblastoma. There are several other types of cancers listed, but neuroblastomas are by far the most strongly associated MeSH Disease term by p-value (\( p = 7.6 \times 10^{-262} \)). ‘Leydig cell tumor’, although less studied in the literature, had the highest fold enrichment (fold = 12.5), with other cancers including Glioma, Osteosarcoma and Adrenal Gland Neoplasms. Clicking the literature link for Neuroblastoma and limiting to Review articles immediately brings our attention to ‘Defects in CAMP-pathway may initiate carcinogenesis in dividing nerve cells’ (Prasad et al., 2003).

Another important feature of Metab2MeSH is the ability to search for compounds associated with a MeSH descriptor, e.g. a disease. Suppose one would like to identify compounds associated with phenylketonuria, a metabolic disease caused by a defect in the enzyme phenylalanine hydroxylase. This enzyme is required to convert phenylalanine to tyrosine. Supplementary Figure S3 shows the top 25 compounds associated with Phenylketonurias. Among these are phenylalanine, endophenyl, tetrahydrobiopterin, sapropertin, peritn and several other compounds that are known to be associated with this disease. This example illustrates that Metab2MeSH prioritizes known relationships. If users are unsure of the MeSH term they require, they can browse the MeSH tree and limiting to Review articles immediately brings our attention to ‘Defects in CAMP-pathway may initiate carcinogenesis in dividing nerve cells’ (Prasad et al., 2003).

In conclusion, Metab2MeSH provides a quick and convenient method for identifying compounds commonly associated with a biological topic or medical term, or for identifying biological topics or medical terms commonly associated with a compound of interest. Although Metab2MeSH is not meant to predict novel associations (due to its dependence on what is in the literature), as the examples illustrate, it is very useful for identifying sets of associations that would otherwise be difficult to find. Metab2MeSH creates a valuable resource.
data set that can be accessed via our website or API, and will likely prove useful in other bioinformatics tools.

**Funding:** National Institutes of Health (U54DA21519); Michigan Diabetes Research and Training Center Pilot and Feasibility Grant (to A.K.); Michigan Nutrition Obesity Research Center (5P30DK089503 - A.K.) and 5R01DK079084 (M.A.S. and B.A.).

**Conflict of Interest:** none declared.

**REFERENCES**


