

# GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM

John D. Osborne<sup>\*1</sup>, Simon Lin<sup>1</sup>, Warren A. Kibbe<sup>1</sup>, Lihua (Julie) Zhu<sup>1</sup>, Maria I. Danila, Rex L. Chisholm<sup>1</sup>

<sup>1</sup>Robert H. Lurie Comprehensive Cancer Center and the Center for Genetic Medicine, Northwestern University, Chicago, IL 60611

<sup>2</sup>Christ Advocate Medical Center, 4440 95th Street, Oak Lawn, IL, 60453

## ABSTRACT

**Motivation:** The human genome has been extensively annotated with Gene Ontology for biological functions, but minimally computationally annotated for diseases.

**Methods:** We used the Unified Medical Language System (UMLS) MetaMap Transfer tool (MMTx) to data mine gene-disease relationships from both the GeneRIF and OMIM databases. We utilized a comprehensive subset of UMLS structured as a directed acyclic graph (the Disease Ontology) to filter and interpret results from MMTx. The data mining methodology was validated against the Homayouni gene collection using recall and precision measurements.

**Results:** The validation data set suggests a 91% recall rate and 97% precision rate of disease annotation using GeneRIF, in contrast with a 22% (recall) and 98% (precision) using OMIM. Our thesaurus-based approach allows for comparisons to be made between disease containing databases and allows for increased accuracy in disease identification through synonym matching.

**Contact:** j-osborne@northwestern.edu

## 1 INTRODUCTION

High throughput microarray technology generates a vast amount of data that is analyzed with a variety of different algorithms, including those that apply literature mining to discover biological relationships (Jenssen, Laegreid et al. 2001; Hoffmann and Valencia 2004). These can include the inference of gene to gene relationships (Homayouni, Heinrich et al. 2005), gene to disease relationships (Hu, Hines et al. 2003) or other types of relationships (Rindfleisch, Tanabe et al. 2000; Srinivasan and Libbus 2004; Cokol, Iossifov et al. 2005).

Knowing the relationship of a group of genes to disease is particularly useful for microarray analysis, where such biological knowledge can be better used to reduce the list of genes produced from such an experiment to be validated or be used to identify promising gene candidates from linkage analysis (Perez-Iratxeta, Wjst et al. 2005; Platts, Moldenhauer et al. 2005; Tiffin, Kelso et al. 2005). Unfortunately obtaining a comprehensive and up to date mapping of genes to disease is not easy. One source that has been used (Andersson, Petersen et al. 2005; Masseroli, Galati et al. 2005) is the Online Mendelian Inheritance in Man (OMIM,2000)

because it is curated and contains a high level of details described for many of its phenotypes. While curation in OMIM can be expected to reduce errors, it also leaves OMIM with time lag. Furthermore, the vocabulary of OMIM is predominately text based, far from comprehensive and is difficult to use (Becker, Barnes et al. 2004; Masseroli, Galati et al. 2005; Smith, Goldsmith et al. 2005). As described in the OMIM FAQ, it is simply not possible to download a list of diseases from OMIM and users of OMIM have in cases resorted to mining the Clinical Synopsis free text section of OMIM for disease discovery (Masseroli, Galati et al. 2005).

Another source of gene-disease mappings from linkage studies is the "Genetic Association Database" (GAD) (Becker, Barnes et al. 2004) which aims to "collect, standardize and archive genetic association study data". However the structure of the classification system GAD used to classify its diseases is not apparent. Diseases are classified in 10 broad classes including an "other" class (some remain unclassified), an unknown number of broad phenotype classes below those and a further number of narrow phenotype classes. The lack of an apparent ontology behind its disease classification makes it hard to determine just how many and what types of diseases GAD contains. For instance searching for "Crohn's disease" will return 25 results but searching for "regional enteritis" returns no results.

With so few sources of genetic association information, researchers have used abstracts and titles from MEDLINE as a data source (Hu, Hines et al. 2003; Hristovski, Peterlin et al. 2005). Although being a current and rich source of information, the free text form of MEDLINE abstracts makes it difficult to identify the proper context in which to associate a gene with a disease (for a review of problems in this area see (Shatkay and Feldman 2003)). This is particularly true when genes are identified by semantically ambiguous gene symbols only some of which may apply to a disease recognized in free text.

An alternative source of gene-disease association data is the GeneRIF database. A GeneRIF entry is a short (up to 255 character) annotation to a gene in the NCBI database. It can include a variety of information including disease associations and be entered by anyone with access to NCBI website willing to provide their email address. The advantage of GeneRIF is that unless there is a GeneRIF mapping error (or the user discusses a disease unrelated to the GeneRIF annotation) any disease described should be correctly mapped to its corresponding gene. Despite this utility, GeneRIF has been infrequently considered as a data source for text

<sup>\*</sup>To whom correspondence should be addressed.

mining, there are only two papers indexed in PubMed containing the keyword “GeneRIF”.

One of those two papers taking advantage of GeneRIF data describes a data mining tool called MILANO (Rubinstein and Simon 2005). While not looking explicitly at gene-disease relationships, the MILANO tool does count occurrences of each GeneRIF annotated gene with a user-defined terms selected from MeSH headlines which includes some disease terms. Their results state that GeneRIFs are “ideal” for microarray analysis because GeneRIF was the only data source of 5 tested (including MEDLINE) that identified all of the p53-affected genes they were testing.

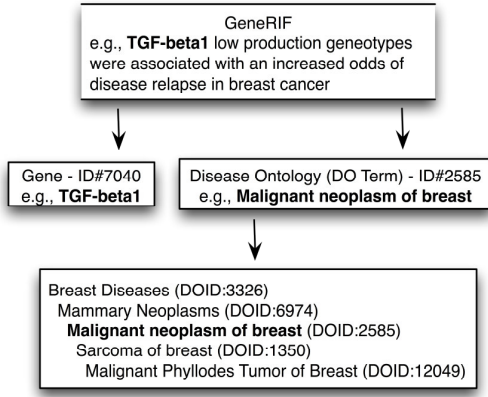


Fig. 1. Sample GeneRIF mapped to DO

Compared with MILANO, we used the Disease Ontology (DO) (Warren A. Kibbe 2006) to identify relevant diseases instead of MeSH terms. The Disease Ontology is a manually inspected subset of UMLS and includes concepts from outside the UMLS disease/disorder semantic network including various cancers, congenital abnormalities, deformities and mental disorders that are important to researchers trying to understand the genetic and molecular basis of a particular disease. While many researchers have mapped diseases to MeSH terms (Jenssen, Laegreid et al. 2001; Srinivasan and Libbus 2004; Hristovski, Peterlin et al. 2005; Perez-Iratxeta, Wjst et al. 2005) or OMIM (Masseroli, Galati et al. 2005; Masseroli, Galati et al. 2005) the disease ontology is much larger in size and should therefore provide greater disease coverage and allow a fairer comparison between GeneRIF and OMIM. Its hierarchal structure also allows more general disease terms to be distinguished from subclasses, in order to account for “over-mapping” of diseases terms to a textually larger database.

Finally, we used a thesaurus based approach (MMTx) for text mining GeneRIFs which has already had success in mining clinical documents for medical problems (Meystre and Haug 2005). It allows us to identify synonyms and provides a convenient mapping to DO diseases. An example of a typical gene-disease mapping is provided in Figure 1.

## 2 METHODS

### 2.1 MMTx

MMTx is a natural language processing engine that identify concepts from free text using a lexicon (Aronson 2001). Briefly, a part-of-speech tagger

labels the noun phrases from the lexical elements created after parsing and tokenization (Fig. 2). These noun phrases and variants of these phrases are used to search the UMLS Metathesaurus to find matching candidates, each of which is given a score; final mappings are generated that best cover the input noun phrase.

Our in-house software parses in the GeneRIF and OMIM data and uses the MMTx API to generate final mappings between genes and diseases. The strict data model of Unified Medical Language System (UMLS) distribution 2005ac was searched against using the default settings of MMTx with an empirically derived score cutoff value of 700. Results were further filtered using the DO version 3.0 (RC9) to eliminate non-disease biological relationships. In addition, a simple heuristic approach was used to eliminate both non-informative mappings to DO (such as “Disease” or “Syndrome”) and to eliminate text present in GeneRIF that was frequently mis-mapped by UMLS such as Ca++ ions being mapped to cancer terms. The program calling the MMTx API and generating the gene-disease mapping is written in java and available upon request.

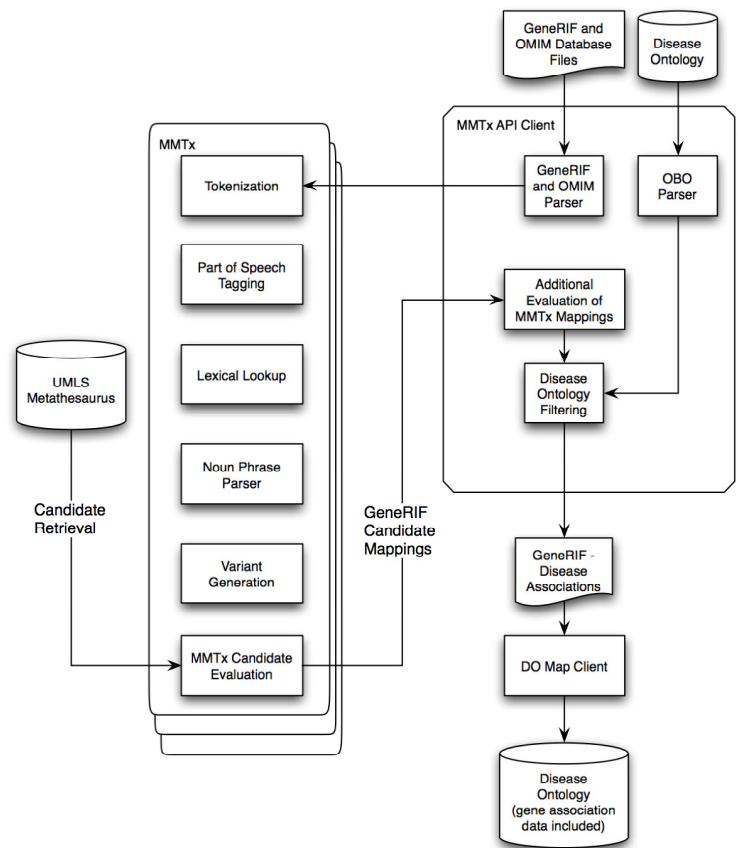


Fig 2. Data Mining Methodology

### 2.2 GeneRIF and OMIM Data

The February 9<sup>th</sup>, 2006 release of both OMIM and GeneRIF were used as input to MMTx. Only validated OMIM data in the formatted “morbidmap” (including disease susceptibilities) file was used as input to MMTx. This is because there are currently only 235 records in OMIM which contain a clinical synopsis of the disease, additional disease information is scattered through other sections of the OMIM record making it hard to determine if the disease mentioned is for an animal model or other comparative purposes.

### 2.3 Scoring and Validation

To evaluate our data mining methodology, and to compare our GeneRIF results with the traditional OMIM resource in detail, we utilized a well-characterized fifty-gene collection by Homayouni et al. (2005) that they used to evaluate semantic indexing of gene functions. This gene collection includes genes in the reelin signaling pathway of Alzheimer's disease and other genes important in cancer biology and development. We call it Homayouni gene collection from here on. The 5 genes with more than 50 diseases mapped to them (*apoe*, *egfr*, *erbb2*, *tgfb1* and *tp53*) were excluded from the test set due to the large number of GeneRIFs requiring manual inspection.

Assessing the false positive and false negative error rates for this collection was difficult (Zeeberg, Riss et al. 2004), so two authors were used for scoring the results. To determine gene-disease relationships for Table 1, a false positive was scored only when the disease was identified incorrectly. No effort was made here to assess the appropriateness of the GeneRIF because of the subjective nature of such a process. However for Table 2 estimates were used for calculating precision and recall rates whereby the overall false positive value was corrected to account for false positives arising when a correctly identified disease did not have a relationship to its associated gene as specified in the GeneRIF.

## 3 RESULTS

Table 1 summarizes the gene-disease mappings for both data sources. GeneRIF contains 139438 entries in 496 taxons. There are 74933 entries for human genes, of which 7443 are associated with diseases.

**Table 1.** Summary statistics of disease annotation by OMIM and GeneRIF

	OMIM	GeneRIF
# Genes Associated with Disease	1988	7443
# Disease-Gene Mappings	3354	31462
# Estimated Potential True Mappings	4025	32935
# Diseases/Gene	2.02	4.42

A mapping is the association of one disease with one gene. Corrected mappings accounts for false negatives and positives.

As can be seen from Table 1, the total number of gene-disease mappings is greater both proportionally and absolutely with GeneRIF than with OMIM. For evaluation, we use two commonly used performance metrics in textual data retrieval, which are defined as follows:

Recall = (Number of relevant ones returned) / (Total number of relevant ones)

Precision = (Number of relevant ones returned) / (Total number of returned)

From these formulas, we can see they are closely related to false positive and false negative rates that are used in other fields. A recall rate of 100% and a precision rate of 100% are of ideal situations. For disease annotation, we constructed a truth table of the Homayouni gene collection manually using GeneRIF and OMIM text as a source.

For the Homayouni gene collection, there are 3879 GeneRIFs, with an average of 77.58 GeneRIFs per gene and a median of 22.

On average our algorithm maps 18.8% of GeneRIFs to a disease, with a false positive rate of 12.2% and false negative rate of 9.5%.

**Table 2.** Estimation of recall and precision of disease annotation using the Homayouni gene collection.

	OMIM	GeneRIF
Recall	21.85	90.76
Precision	98.46	96.66

## 4 DISCUSSION

Our results indicate that GeneRIFs are an excellent data source for mining disease-gene relationships. This is primarily the result of both the large number of GeneRIFs relative to OMIM, and the surprisingly high (18.8%) frequency at which diseases are mentioned. OMIM's showing for disease coverage would be improved if we had decided to mine OMIM's free text, but only 235 genes in OMIM have a clinical synopsis section which would bias the results. Using this and other OMIM free text would also increase the number of false positives since OMIM free text frequently includes diseases not directly related to the gene, usually for comparative purposes or in an animal model.

Errors in our method arise from a variety of sources, include problems with MMTx, many of which have already been analyzed (Divita, Tse et al. 2004). The problem of having disease terms present in OMIM or GeneRIF, but missing in DO or UMLS was infrequent but did include some cases such Craniofacial-deafness-hand syndrome. A more significant problem (which made up the bulk of the false positive rate) was the discovery of disease terms in GeneRIF that indicated only a partial, ambiguous or no association to the gene in question. Fortunately, the succinctness of GeneRIF means that this occurs less frequently than in abstracts (data not shown) where diseases not directly related to the described genes may clutter the text. Finally, we found only a single incorrectly assigned GeneRIF in the 1746 GeneRIFs we examined, making this a minor source of error.

One trend that emerged from our analysis is that OMIM performs poorly relative to GeneRIF with newly discovered mappings. A fairly typical case is the alpha-2-macroglobulin gene. While OMIM includes mappings for Alzheimer's disease and pulmonary emphysema (which GeneRIF missed), it excludes potential links to benign prostatic hyperplasia, multiple sclerosis and argyrophilic grain disease. Part of this may be OMIM's stronger requirement for evidence, but failure to keep pace with the continuous stream of new research must also play a role.

We established a data mining methodology in this paper and reported its performance. In a subsequent paper, we will describe a web-based database application to serve the results to data consumers.

## ACKNOWLEDGEMENTS

The authors thank Dr. Pan Du for discussing related data mining applications using GeneRIF.

## REFERENCES

- Andersson, L., G. Petersen, et al. (2005). "A web tool for finding gene candidates associated with experimentally induced arthritis in the rat." *Arthritis Res Ther* **7**(3): R485-92.
- Aronson, A. R. (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." *Proc AMIA Symp*: 17-21.
- Becker, K. G., K. C. Barnes, et al. (2004). "The genetic association database." *Nat Genet* **36**(5): 431-2.
- Cokol, M., I. Iossifov, et al. (2005). "Emergent behavior of growing knowledge about molecular interactions." *Nat Biotechnol* **23**(10): 1243-7.
- Divita, G., T. Tse, et al. (2004). "Failure analysis of MetaMap Transfer (MMTx)." *Medinfo* **11**(Pt 2): 763-7.
- Hoffmann, R. and A. Valencia (2004). "A gene network for navigating the literature." *Nat Genet* **36**(7): 664.
- Homayouni, R., K. Heinrich, et al. (2005). "Gene clustering by latent semantic indexing of MEDLINE abstracts." *Bioinformatics* **21**(1): 104-15.
- Hristovski, D., B. Peterlin, et al. (2005). "Using literature-based discovery to identify disease candidate genes." *Int J Med Inform* **74**(2-4): 289-98.
- Hu, Y., L. M. Hines, et al. (2003). "Analysis of genomic and proteomic data using advanced literature mining." *J Proteome Res* **2**(4): 405-12.
- Jenssen, T. K., A. Laegreid, et al. (2001). "A literature network of human genes for high-throughput analysis of gene expression." *Nat Genet* **28**(1): 21-8.
- Masseroli, M., O. Galati, et al. (2005). "Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists." *BMC Bioinformatics* **6 Suppl 4**: S18.
- Masseroli, M., O. Galati, et al. (2005). "GFINDER: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists." *Nucleic Acids Res* **33**(Web Server issue): W717-23.
- Meystre, S. and P. J. Haug (2005). "Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx)." *Stud Health Technol Inform* **116**: 823-8.
- Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000. World Wide Web URL: <http://www.ncbi.nlm.nih.gov/omim/>
- Perez-Iratxeta, C., M. Wjst, et al. (2005). "G2D: a tool for mining genes associated with disease." *BMC Genet* **6**: 45.
- Platts, A. E., J. S. Moldenhauer, et al. (2005). "LARAlink 2.0: a comprehensive aid to basic and clinical cytogenetic research." *Genet Test* **9**(4): 334-41.
- Rindflesch, T. C., L. Tanabe, et al. (2000). "EDGAR: extraction of drugs, genes and relations from the biomedical literature." *Pac Symp Biocomput*: 517-28.
- Rubinstein, R. and I. Simon (2005). "MILANO--custom annotation of microarray results using automatic literature searches." *BMC Bioinformatics* **6**(1): 12.
- Shatkay, H. and R. Feldman (2003). "Mining the biomedical literature in the genomic era: an overview." *J Comput Biol* **10**(6): 821-55.
- Smith, C. L., C. A. Goldsmith, et al. (2005). "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information." *Genome Biol* **6**(1): R7.
- Srinivasan, P. and B. Libbus (2004). "Mining MEDLINE for implicit links between dietary substances and diseases." *Bioinformatics* **20 Suppl 1**: I290-I296.
- Tiffin, N., J. F. Kelso, et al. (2005). "Integration of text- and data-mining using ontologies successfully selects disease gene candidates." *Nucleic Acids Res* **33**(5): 1544-52.
- Warren A. Kibbe, J. D. O., Wendy A. Wolf, Maureen E. Smith, Lilhua Zhu, Simon Lin and Rex L. Chisholm (2006). *The Disease Ontology and Browser*. caBIG Annual Meeting, Washington D.C.
- Zeeberg, B. R., J. Riss, et al. (2004). "Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics." *BMC Bioinformatics* **5**: 80.