Mapping Biomedical Literature Annotations to Extract Gene-Disease Associations

[1,2] Warren A. Cheung[*], [3] B.F. Francis Ouellette and [2] Wyeth W. Wasserman

[1]Bioinformatics Program, University of British Columbia, Vancouver, BC
[3] Centre for Molecular Medicine and Medicine, Vancouver, BC
[3] Ontario Institute for Cancer Research, Toronto, ON
*Author correspondence: e-mail wcheung@cmmt.ubc.ca

INTRODUCTION

The identification and ranking of candidate genes is a critical step in the search for the genetic basis of disease. Integrated approaches to the computational analysis of diverse data collections offer the possibility to evaluate links between genes and diseases. We focus on the analysis of biomedical literature, with the ultimate goal of identifying human genes which play a previously unknown functional role in the pathology of disease.

The novel system we describe draws on data from the PubMed online repository of biomedical literature, Entrez Gene as a validated source for genes and Medical Subject Headings (MeSH) as a structured vocabulary for diseases. We combined the data in an integrated database and developed a relationship evaluation framework. We present a statistical scoring method based on over-representation analysis to evaluate the strength of observed gene-disease associations. The gene-disease relationships extracted were validated against a reference collection extracted from known relationships in the Online Mendelian Inheritance in Man (OMIM).

Related Work

Of the many diverse methods for predicting gene-disease associations present in the literature, few utilise the MeSH annotation of PubMed articles in their analysis. G2D[2] links MeSH disease terms to genes, by following by linking MeSH disease terms with co-occurring MeSH chemical compound terms, which are then linked to Gene Ontology(GO) functional annotations. These functional annotations related to disease are then compared to the GO functional annotations of proteins, and a fuzzy set theory approach was used to score these relations. This method was recently updated[3] to incorporate other protein interaction information. Another method that uses MeSH annotation of PubMed abstracts is BITOLA[1], which extracts the frequency and number of articles supporting co-occurrence of disease, other MeSH terms and automatically extracted gene names. They also employ an ad-hoc scoring method based on the number of supporting articles to rank their gene-disease predictions.

Our method looks for direct gene-disease associations. We implement an automated system to extract the gene-disease relationships, using GeneRIF literature annotations of genes and MeSH annotation of PubMed articles. We provide a statistical model to evaluate the relationships extracted, and compare our results to validation set automatically generated from OMIM.

METHODS AND MATERIALS

Entrez Gene (Feb 2007), OMIM (Feb 2007) and Homologene (Aug 2007; build 57) were downloaded directly from the NCBI FTP repository, as was the MeSH 2007 collection. All PubMed entries with an annotated term in the MeSH disease category were downloaded (Nov 2007) via the Entrez E-Utils interface. The Unified Medical Language System (UMLS) Metathesaurus 2007AB was used to map OMIM terms to their MeSH equivalents. Data analysis was performed via direct SQL queries to a relational database (MySQL 5 and MS Access 12). Statistical scoring was performed using the R statistics package 2.4 and MS Excel 12.

RESULTS AND DISCUSSION

From Genes to Diseases

We consider the 38 604 known or hypothetical human genes in Entrez Gene as our validated gene set. 9766 of these genes have been annotated with GeneRIFs. A GeneRIF describes the function of a gene, providing a brief description of the link and the PubMed identifiers of supporting biomedical literature. Curators at the National Library of Medicine annotate all PubMed articles using MeSH terms.

MeSH is a structured controlled vocabulary, with terms organized by categories, subcategories and so on in a tree-like structure of increasing specificity. One of the top-level categories is Diseases (Category C),

which comprises 4299 of the 24 355 MeSH terms of varying degrees of specificity, from the very general (e.g. “Nervous System Disease”) to the specific (e.g. “Alzheimer Disease”). We consider a subset of PubMed entries, comprising the 8 785 701 articles annotated with a Diseases MeSH term.

We therefore use GeneRIFs to link human genes to the Disease MeSH terms. Specifically, this links 9766 human genes that have at least one GeneRIF annotation to the 41 832 PubMed articles with disease-related MeSH terms. This yields 221 170 distinct human gene-disease relationships, relating human genes to 2895 of the Disease MeSH terms (See Figure 1).

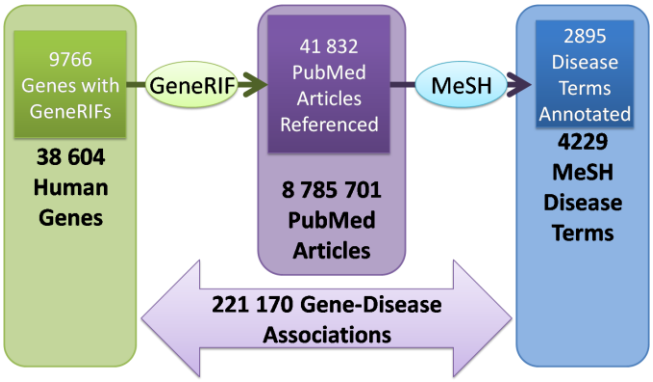

Figure 1: We follow 9766 of the human known or hypothetical genes in Entrez Gene which have GeneRIF references leading to 41 832 distinct PubMed articles annotated with one or more Disease Category MeSH terms. These PubMed articles were annotated by 2895 MeSH different terms from the MeSH Disease Category.

Statistical Scoring

Although the gene-disease relationships all have at least one PubMed article as support, this does not guarantee that the association is valid. Particularly, it is possible that the GeneRIF reflects the association of the gene with a different aspect addressed by the article, and that the occurrence of the disease was by chance. If we assume that such occurrences are random and independent in context of the GeneRIFs, the probability of such occurrences can be modeled by the hypergeometric distribution, and we can compute p-values by a one-tailed Fisher’s exact test.

Specifically, for a given gene and disease MeSH term, we consider the PubMed articles referred to by the GeneRIFs for our given gene. We then compare the number of these referenced articles which have our given disease MeSH term or one of its children in the MeSH hierarchy to the total number of articles referenced. This is compared to all the PubMed articles which has any disease MeSH term, and the number of articles in PubMed with our given disease MeSH term. For example, the gene A2M has GeneRIF references to 31 PubMed articles, 8 of which have annotations for the MeSH term “Alzheimer Disease”. The term “Alzheimer Disease” is annotated on 42 120 of the 8 785 701 PubMed articles annotated with MeSH disease terms. This can be expressed as the using a contingency table (See Table 1) and by the Fisher’s exact test, results in a p-value of 1.97E-12.

Table 1: Contingency Table for the number of PubMed articles with disease MeSH terms referenced by gene A2M and the MeSH term “Alzheimer Disease”

	Referenced by GeneRIFs	Not Referenced by GeneRIFs	Total
Annotated with MeSH term	**8**	**42 112**	**42 120**
Not Annotated with MeSH term	**23**	**8 743 558**	**8 743 581**
Total	**31**	**8 785 670**	**8 785 701**

However, when computing such statistics, we have to account for the large number of tests being done. In this situation, we are subject to the possibility of increased Type I error. To account for this, we employ the conservative Bonferroni correction, accounting for each gene-disease association as being a separate test. Continuing our previous example of the gene A2M and the MeSH term “Alzheimer Disease”, the corrected p-value for the relationship, after applying Bonferroni correction for the 221 170 tests, is 4.36E-07.

After multiple-testing correction, we have 9014 gene-disease associations that are significant (p-value < 0.05), with 3018 of these associations having p-value < 10^{-13} (See Figure 2). Table 2 lists the 50 most significant results with the highest number of unique supporting PubMed articles.

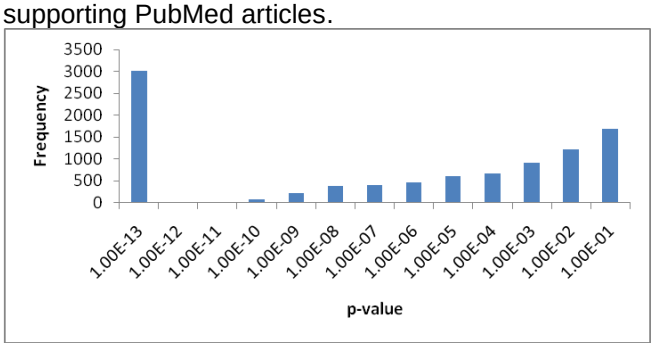

Figure 2: Histogram of p-values ≤ 10^{-1} for the gene-disease associations found by the system.

Table 2: Top 50 results by PubMed evidence (Bonferroni Corrected p-value < 10^{-13})

Entrez Gene ID	Gene Name	MeSH Term	PubMed articles
7157	TP53	Neoplasms	708
7157	TP53	Neoplasms by Site	529
7157	TP53	Neoplasms by Histologic Type	387
7422	VEGFA	Neoplasms	298
5743	PTGS2	Neoplasms	281
7157	TP53	Neoplasms, Glandular and Epithelial	275
7422	VEGFA	Pathologic Processes	274
1029	CDKN2A	Neoplasms	271
7157	TP53	Pathological Conditions, Signs and Symptoms	256
2064	ERBB2	Neoplasms	253
5743	PTGS2	Neoplasms by Site	253
7422	VEGFA	Neoplasms by Site	237
2064	ERBB2	Neoplasms by Site	236
1956	EGFR	Neoplasms	222
7157	TP53	Carcinoma	222
672	BRCA1	Neoplasms	213
1029	CDKN2A	Neoplasms by Site	208
672	BRCA1	Neoplasms by Site	207
1029	CDKN2A	Neoplasms by Histologic Type	206
672	BRCA1	Skin and Connective Tissue Diseases	184
672	BRCA1	Skin Diseases	184
672	BRCA1	Breast Diseases	183
672	BRCA1	Breast Neoplasms	183
1956	EGFR	Neoplasms by Site	181
7422	VEGFA	Neoplasms by Histologic Type	181
351	APP	Nervous System Diseases	176
7157	TP53	Digestive System Neoplasms	176
7422	VEGFA	Metaplasia	171
7422	VEGFA	Neovascularization, Pathologic	171
596	BCL2	Neoplasms	170
351	APP	Central Nervous System Diseases	168
351	APP	Brain Diseases	166
2099	ESR1	Neoplasms	166
999	CDH1	Neoplasms	165
7157	TP53	DNA Damage	164
9370	ADIPOQ	Nutritional and Metabolic Diseases	163
351	APP	Neurodegenerative Diseases	162
5743	PTGS2	Neoplasms by Histologic Type	162
2099	ESR1	Neoplasms by Site	159
7015	TERT	Neoplasms	158
351	APP	Dementia	157
351	APP	Alzheimer Disease	156
351	APP	Tauopathies	156
1026	CDKN1A	Neoplasms	156
348	APOE	Nervous System Diseases	153
595	CCND1	Neoplasms	153
367	AR	Male Urogenital Diseases	150
1956	EGFR	Neoplasms by Histologic Type	150
999	CDH1	Neoplasms by Site	149

Validation

To validate these results, we compared the gene-disease associations derived by our system to the OMIM gene references provided in Entrez Gene. OMIM is a collection of annotations for human diseases with a genetic component, and the genes implicated in these diseases. OMIM tracks and distinguishes both known and suspected relationships. OMIM entries were mapped to MeSH terms via the UMLS Metathesaurus. The UMLS Metathesaurus maps concepts from a variety of vocabularies to its internal concepts. We consider an OMIM entry and a MeSH term as equivalent when both refer to the same UMLS concept. Of the 17503 OMIM entries (genes and diseases) in UMLS, 462 OMIM-MeSH mappings were generated, linking 462 OMIM entries (for genotypes and phenotypes) to 435 MeSH terms (for genes and diseases) (See Figure 3). Applying this mapping to the OMIM references in Entrez Gene resulted in a set of 570 bona fide associations between genes in Entrez Gene and OMIM phenotypes with MeSH terms.

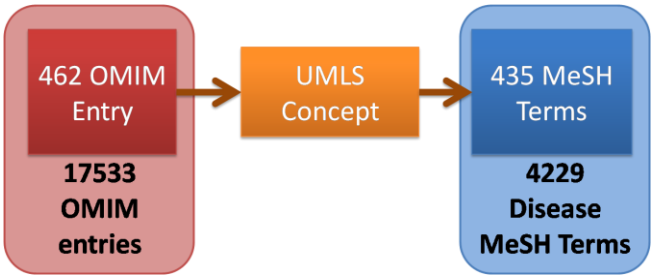

Figure 3: Of the 17533 OMIM entries in UMLS, 462 OMIM entries map to an UMLS concept which in turn maps to one of 435 Disease Category MeSH terms.

Contrasting this result with the gene-disease associations ranked by our system, 379 of the 570 associations were found (66.5% sensitivity), and 223 of these were deemed significant (39.1% sensitivity) (See Figure 4). Of particular interest are the 8635 gene-disease associations that were not in the validation set but were labeled as significant. Further study will be needed to determine whether these predictions have utility as initial indicators of gene-disease associations.

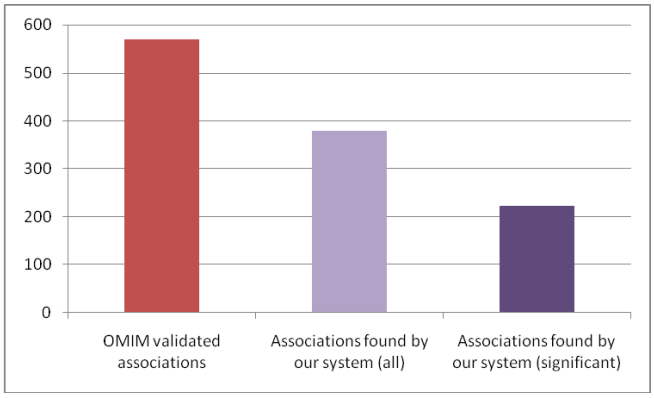

Figure 4: Gene-disease associations from Entrez Gene to Disease MeSH terms, comparing the validated associations in OMIM against associations found by the system.

To facilitate prediction of novel gene-disease associations, beyond extraction of known associations, we will need to extend the system to incorporate additional sources of relations. Such an approach could include the analysis of data for homologous genes studied in model organisms and extending literature analysis to include secondary and tertiary relationships between articles – via citation and textual similarity – to identify more subtle relationships.

CONCLUSION

In this paper, we demonstrate a literature-based disease-gene association system for recovering gene-disease associations using GeneRIFs from Entrez Gene and MeSH annotations of PubMed articles. We validated our results against associations automatically extracted from a reference set, OMIM, successfully extracting the majority (66.5%) of associations, including the 39.1% of associations with sufficiently strong support to be statistically significant.

ACKNOWLEDGEMENTS

This research was funded in part by the MSFHR/CIHR Bioinformatics Training Program and NSERC.

REFERENCES

1. Hristovski, D., Peterlin, B., Mitchell, J.A., Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2-4): 289-298.
2. Perez-Iratxeta, C., Bork, P., Andrade, M. (2002). Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31: 316-319.
3. Perez-Iratxeta, C., Bork, P., Andrade-Navarro, M. (2007). Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Research,* 35(Web Server issue): W212-W216.