Pacific Symposium on Biocomputing 2005: World Scientific; 2005. p. 103-114.

GENESTRACE: PHENOMIC KNOWLEDGE DISCOVERY VIA STRUCTURED TERMINOLOGY

MICHAEL N. CANTOR*
Department of Medicine, Beth Israel Medical Center, New York, NY 10003

INDRA NEIL SARKAR*
Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024
& Department of Biomedical Informatics, Columbia University, New York, NY 10032

OLIVIER BODENREIDER
Lister Hill National Center for Biomedical Communications,
National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

YVES A. LUSSIER §
Departments of Biomedical Informatics and Medicine,
Columbia University, New York, NY 10032
Email: yves.lussier@dbmi.columbia.edu

The era of applied genomic medicine is quickly approaching accompanied by the increasing availability of detailed genetic information. Understanding the genetic etiology behind complex, multi-gene diseases remains an important challenge. In order to uncover the putative genetic etiology of complex diseases, we designed a method that explores the relationships between two major terminological and ontological resources: the Unified Medical Language System (UMLS) and the Gene Ontology (GO). The UMLS has a mainly clinical emphasis; Gene Ontology has become the standard for biological annotations of genes and gene products. Using statistical and semantic relationships within and between the two resources, we are able to infer relationships between disease concepts in the UMLS and gene products annotated using GO and its associated databases. We validated our inferences by comparing them to the known gene-disease relationships, as defined in the Online Mendelian Inheritance in Man's morbidmap (OMIM). The proof-of-concept methods presented here are unique in that they bypass the ambiguity of the direct extraction of gene or disease term from MEDLINE. Additionally, our methods provide direct links to clinically significant diseases through established terminologies or ontologies. The preliminary results presented here indicate the potential utility of exploiting the existing, manually curated relationships in biomedical resources as a tool for the discovery of potentially valuable new gene-disease relationships.

The GenesTrace system may be accessed at the following URL:

http://phene.cpmc.columbia.edu:8080/genesTrace/index.jsp

* These authors contributed equally to the work
§ Corresponding author

1. Introduction

Much of biological hypothesis generation follows the proverbial model of trying to discover the "golden needle in the haystack." Discovering significant genes is becoming a daunting task as various genome projects are creating sequence data information at accelerating rates. Thus, it is quickly becoming an intractable problem for the biomedical scientist to stay abreast all putative genes that may hold hidden keys toward the understanding of disease. Barring exhaustive wet-lab and in vivo experimentation, the use of knowledge bases may offer some insight to this task. Novel bioinformatic methods are required to elucidate putative genes that may be related to the etiology of disease. We describe one such method that may offer such insight using relationships that exist within and among terminologies and ontologies of biomedical knowledge.

The emergence of the Gene Ontology™ (GO) [1] and its related databases is an important advance in discovering elusive gene-disease relationships, as it provides a standardized, easily searchable repository of biological information. A great deal of research is focused on developing or improving methods for gene and sequence annotation (see below); however, relatively little research has looked at the equally, if not more, complex idea of relating GO to the level of clinical diseases. This project attempts to bridge that gap by exploring links between GO and its annotation databases to clinical concepts that are in the Unified Medical Language System® (UMLS®) [2]. Ultimately, this research aims to highlight possible gene-disease relationships via the mappings of structured terminologies (both clinical and biological) contained in the UMLS.

A large amount of biomedical knowledge is represented in free-text form, such as MEDLINE abstracts. Extracting important information from these resources is an extremely active area of research. PubGene, for example, is a database for gene-expression analysis, extracted from a weighted network of gene co-occurrence data in MEDLINE [3]. PubGene's gene-gene relationships were validated by comparison with the Online Mendelian Inheritance in Man (OMIM) database [4]. The utility of literature associations were also validated by a comparison to microarray data [3]. The authors noted problems with the ambiguity of gene names and symbols while creating the PubGene network. Additionally, the network did not attempt to characterize the relationships represented in the co-occurrence network.

Fuzzy set theory has also been used as an attempt to characterize candidate disease genes. This approach exploits relationships between two types of annotations, MeSH headings for MEDLINE articles and GO terms for protein sequences. The goal of two methods based on fuzzy set theory is to "derive relationships between pathological conditions and terms describing protein function." [5,6] The evaluation of one system showed an association between the authors' scoring system and the likelihood of a gene-disease association [5].

Similarly, the creators of the MedGene database looked at co-occurrences between genes and MeSH disease terms, and then using those relationships, analyzed microarray data [6].

Several groups have also looked at various automated methods for annotating genes and protein sequences with gene ontology functions. The Gene Ontology Annotation (GOA) project uses manual mappings between GO terms and either protein domains or SWISS-PROT keywords to automatically assign GO terms to a sequence containing previously annotated domains [7]. Raychaudhuri et al. [8] used statistical document classification methods to analyze abstracts from the medical literature. From this analysis, they were able to associate a set of GO terms to the genes mentioned in the abstracts. Building on their previous work [5], Perez et al. [9] developed a system to associate keywords to genes or protein sequences using mappings between SWISS-PROT keywords, MESH terms associated with MEDLINE abstracts, and GO terms. Their system demonstrated better performance with mapping SWISS-PROT terms that GO terms. The difference was attributed to the ambiguity introduced into GO mappings by its ontological structure.

Here, we describe a novel method that builds on previous attempts at bridging biomedical terminologies to infer putative genes implicated in disease etiology. Our method, GenesTrace, uses biological and clinical terminologies contained in the UMLS to induce modal relationships. We hypothesize that this putative phenomic network can further be filtered and mined to reveal buried knowledge. The method we propose is fundamentally different from previously cited ones. For example, instead of using fuzzy logic or statistical methods, our proposed method infers genes-to-diseases relationships by constructing an original network of relationships between curated ontologies and databases and then selecting paths in the network, which fulfill valid semantic constraints. From this proof-of-concept study, we will further describe how GenesTrace's inferences may provide some guidance towards subsequent investigations of the genetic etiology of complex diseases.

2. Methods

2.1. *Materials*

UMLS Database. We used the 2003AB version of the UMLS Metathesaurus®, which contains approximately 900,000 concepts from 102 biomedical source

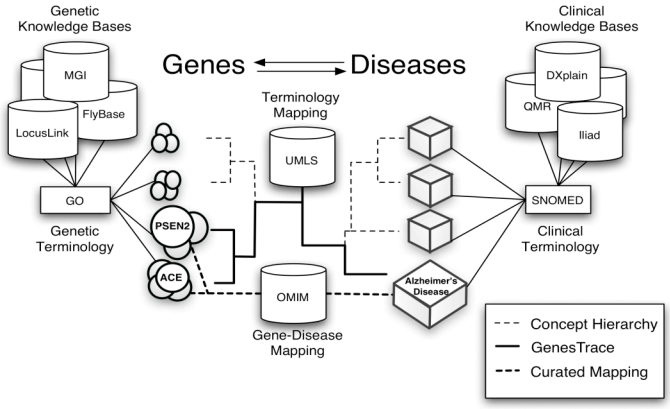

Figure 1: ***Gene-Disease Mapping***. Gene Terms from Genetic Knowledge Bases, such as MGI, LocusLink, or Flybase, are codified using genetic terminologies, such as GO (left). Disease Terms from Clinical Knowledge Bases, such as DXplain, QMR, or Iliad, are codified using clinical terminologies, such as SNOMED (right). Collectively, these terminologies are mapped as concepts in a hierarchical manner, as in the UMLS (center). Manually curated knowledge databases, like OMIM, map some gene product concepts with disease concepts (bold dashed line). The proposed method, GenesTrace, exploits terminology mappings, which relate disease concepts to gene concepts (bold solid line).

terminologies[2]. Many of these source terminologies have a strictly clinical focus. No single terminology in UMLS 2003AB is as specific to molecular biology as GO. With the inclusion of GO, the UMLS spans all the various knowledge representation levels of structural and functional concepts in biomedicine – as previously conceptualized by Blois – from molecular information to clinical and disease terms [10].

In the process of integrating GO into the UMLS, many GO terms became new concepts. In many instances, however, related concepts already existed in the Metathesaurus, to which these new concepts were not necessarily linked by explicit relationships. When appropriate, we mapped the new GO concepts to their existing relatives. We also used a table, as supplied by one of the authors, of gene products in the GO databases that were directly represented in the UMLS as an alternate target for disease relationships. Our mappings and inferences were based on the April 2003 distribution of GO [1] and its associated databases. The GO databases include genes and gene products and their associated GO terms, which are represented in a number of important biological databases. We searched four model organism databases (fly, yeast, worm, mouse) and used SwissProt-TREMBL for human genes and products.

[2] http://umlsks.nlm.nih.gov

OMIM Database. The Online Mendelian Inheritance in Man database (OMIM) was used as the gold standard for our validation steps. OMIM contains over 14,000 detailed free-text entries about human genes and genetic disorders [4]. Additionally, OMIM provides *morbidmap*, which gives "the chromosomal location, gene symbol, method(s) of mapping, and disorder(s) related to the specific gene [4]," as well as specific mutations of identified genes. In this study, we used the April 2003 versions of OMIM and its *morbidmap*.

2.2. *General Steps of the GenesTrace Method*

The proposed method, **GenesTrace,** reveals relationships (*traces*) between a disease and a gene according to the following three-step process: (see *Figure 1*).

1. **Identify a Disease** that exists in the UMLS as a concept;
2. **Determine Relationships** between a UMLS disease and other UMLS concepts, such as those in biological terminologies such as GO, using both the symbolic relationships (hierarchical and associative) and the statistical relationships (co-occurrence information);
3. **Identify Putative Genes** that use these related concepts and then use the terminology to determine the list of putative genes through links to valuable knowledge contained in biological databases (e.g., FlyBase, WormBase, MGI, etc.).

2.3. *Inferring Gene-to-Disease Relationships*

To take advantage of the knowledge resources presented by the UMLS, GO, and GO's associated databases (hereafter referred to as "GODB"), we developed a series of methods that related clinical concepts of disease, represented in the UMLS, to gene products represented in GODB.

The first step in our analysis was to select the set of source concepts in the UMLS. We were able to take advantage of a previously described method that transforms UMLS's full graph organization in a directed acyclic graph to obtain all entries that were descendants of the concept "disease" (C0012634), which led to a data set of approximately 200,000 concepts [11]. For each "disease" concept in this list, we then obtained a set of related concepts, using two knowledge source tables from the UMLS: MRREL and MRCOC. MRREL consists of semantically related concepts; MRCOC contains co-occurences between concepts in MEDLINE. We set a minimum threshold of five co-occurrences in MEDLINE as represented in MRCOC to be considered as a significant entry.

Next, we obtained the subset of the related concepts that were represented in GO. We used two sources: the GO terms already represented officially as concepts in the UMLS and the set of experimental mappings of gene products to

UMLS concepts. In order to provide the most inclusive set of results, we used all relevant concepts for GO terms, since there was some overlap among the new, GO-specific concepts and previous concepts (i.e. "acetylcholinesterase", C0001044 and "acetylcholinesterase activity" [from GO], C1149827). Once a merged set of relevant GO terms was established, we systematically obtained the gene products that were associated with each GO term from the GO databases. This was done using SQL queries based on the Perl GO Application Programming Interface. Valid gene-to-disease relationships were then inferred by extracting associations supported by the highest levels of evidence in GO, *IDA* (Inferred from Direct Assay), and *TAS* (Traceable Author Statement).

From the results of each disease-specific query, we created a database wherein each row contained a disease concept, the concept related to the disease and to a GO term, the GO term and accession number represented by the related concept, and additional descriptive information for each gene product (i.e., gene name, gene unique identifier, source database). We also noted the source of the relationship, which was either statistical (co-occurrence), semantic, or both. For semantic relationships, we noted the type(s) of relationship represented (i.e., parent, sibling, etc.)

A second method consisted in finding concepts that were gene products related to the "disease" concepts. For the gene products that were directly represented in the UMLS, according to the experimental mapping mentioned above, we incorporated similar relevant information (e.g., the disease concept, the related concept, the GO product represented by the related concept, gene name, unique identifier, and source database).

2.4. *Evaluation*

We created a sub-table in our database consisting of the concepts corresponding to OMIM diseases represented in the UMLS. These concepts were obtained using string matching (both exact and normalized) with semantic checking, between entries in "*omim.txt"* and the set of UMLS concepts.

Gold Standard. Next, we compared – using lexico-semantic mapping techniques – the set of genes listed in each disease's *OMIM morbidmap* entry to the set of corresponding genes and gene products that had been associated with the disease in our database. We then tabulated the number of genes associated with each disease. True Positives (TP) were defined as instances where the gene was associated with a concept it is related to in *OMIM's morbidmap*. False Positives (FP) were defined as instances where the gene was associated with any other concept. False Negatives (FN) were instances where genes that were known to be associated with a disease were not retrieved using GenesTrace. The

precision and recall of our system was measured as TP/(TP+FP), and TP/(TP+FN), respectively.

3. Results

3.1. *Sample Trace*

The genes trace for UMLS concept C0002395, "Alzheimer's Disease", is presented here as an illustrative example of a GenesTrace and its efficacy using OMIM's *morbidmap*. From the UMLS, we retrieved 128 distinct concepts related to Alzheimer's Disease using MRREL. An additional 993 concepts were retrieved from MRCOC. The relationships from MRREL were divided among seven semantic classes. Mapping these concepts to the GO database led to 102 distinct GO terms: 62 were found in the molecular function axis, 25 in the biological process axis, and 16 in the cellular component access. From GO, GenesTrace found 102 associated GO terms annotating 10,774 distinct molecular products. For this specific trace, we noted that all of the associations were the results of relationships in MRCOC. Of the 12 genes associated with Alzheimer's disease found in *OMIM's morbidmap*, GenesTrace returned 3; however, only 6 of the 12 genes from *OMIM's morbidmap* were represented in the GODB. 242 other genes were also associated with this source concept. Thus we assessed the number of TP to be 3; FN to be 3, and FP to be 242, giving a precision of 1.2%, and a recall (sensitivity) of 50% Of note, this is close to the lowest value of precision in our range of results.

3.2. *Validation*

Out of the 200,000 disease concepts in the UMLS Metathesaurus, 1,407 were associated with at least one associated GO term. We found at least one gene in the database related to 142 of the 1,407 disease concepts. Globally, we retrieved 124 distinct genes in the context of being related to their specific disease concept, and 290 distinct genes erroneously associated with concepts, for a precision of 30% and recall of 8.8%. Overall, there were an average of 3.1 gene products per disease concept in this dataset (range 1-30; 89% had 1-3 genes). Of note, only 978 of the genes in OMIM's morbidmap existed in the GO databases. For specific diseases, our system had a wide range of precision and recall, from 100% each (1 TP, no FP's or FN's) for several diseases (Multiple Endocrine Neoplasia type 1, Neurofibromatosis type 2), to a precision of 1% with a recall of 100% (Ovarian Tumors), to a precision of 100% with a recall of 50% (Hurler Syndrome, Familial Hypobetalipoproteinemia). *Figure 2* shows the distribution

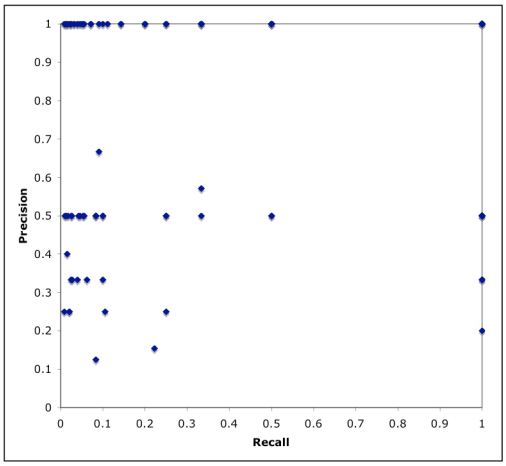

Figure 2: ***Distribution of Precision and Recall Values***. All values, including overlapping points, shown for retrieving pertinent genes for diseases contained in OMIM.

of the precision and recall for all the diseases examined. Interestingly, almost 30% of diseases have a precision and recall of 100%. Overall, average precision for diseases was 51%, with an average recall of 79%. Two cases are illustrative of the complexity of the results. The results for Alzheimer's Disease, mentioned above, are typical of results for complex diseases. Of note, among the gene products were *ACE, ACE1,* and *PSEN2,* all of which were related to Alzheimer's Disease in OMIM's *morbidmap*. On the other end of the spectrum, for concept C0027832, "Neurofibromatosis 2", we retrieved only 2 gene products, *GOT1* and *NF2*; however, as *NF2* is the only related gene in *morbidmap* this led to a precision and recall of 100%.

3.3. *GO-UMLS-OMIM Specific Results*

The results of our gene-disease mappings also revealed some interesting properties of the relationship between GO and the UMLS. As shown in *Tables 1 to 4*, GenesTrace found different types of relationships between either GO terms and UMLS diseases. While few GO terms are associated with 100 diseases or more, 75% of them are linked to 8 diseases or less (*Table 1*). Similarly, 60% of the diseases are associated with only one GO term, but 116 diseases are linked to 10 GO terms or more (*Table 2*). Similar findings were observed between gene products and diseases. While 23 gene products are associated with 200 diseases or more, 75% of them are linked to 25 diseases or less (*Table 3*). Similarly, 80 diseases are associated with more than 1000 gene products, but most diseases are linked to at most 100 gene products (*Table 4*).

Table 1: Percentage of GO Terms associated with individual diseases.

GO Terms (%)	Disease(s)
1-25	1
26-50	1-3
51-75	3-8
76-100	8-132

Table 2: Percentage of individual diseases associated with individual GO terms

Diseases (%)	GO Term(s)
1-25	1
26-50	1
51-75	1-2
76-100	2-102

Table 3: Percentage of Gene Products associated with individual diseases.

Gene Products (%)	Disease(s)
1-25	1-3
26-50	3-13
51-75	13-25
76-100	25-331

Table 4: Percentage of individual diseases associated with gene products.

Diseases (%)	Gene Product(s)
1-25	1-4
26-50	4-15
51-75	15-70
76-100	70-10,774

4. Discussion

Our methods expand on those of other authors in several ways. Perhaps most importantly, we bypass the ambiguity involved in extracting gene names or symbols directly from MEDLINE articles by using pre-defined concepts in the UMLS and GO. Additionally, we use more robust statistical (co-occurrence) data than direct extraction from MEDLINE. Our methods use co-occurrence data that it is calculated on the conceptual, rather than textual, level. In addition to imposing a minimum threshold on the frequency of co-occurrence, the quality of the associations selected is ensured in part by the fact that co-occurrences recorded in the Metathesaurus are limited to the "starred" (major) MeSH descriptors in MEDLINE. Since we also use manually-curated semantic relationships extracted from the UMLS, the characteristics of the relationships we use are also richer than pure co-occurrence data. GenesTrace leverages the above annotation methods and their results, in order to exploit the knowledge contained in GO, the GO annotation databases, and the UMLS. In doing so, the purpose of GenesTrace can be seen in parallel to the annotation projects mentioned above; instead of annotating sequences, however, GenesTrace provides automated methods of annotating diseases.

Inherently, using the UMLS and GO as knowledge sources produces a very large search space. Of the approximately 200,000 initial concepts from the UMLS, for example, 22,040 (11%) had at least one corresponding entry in the GO database. The entire GenesTrace database was comprised of approximately 3 million rows, with 1,326 distinct GO terms and 16,984 distinct gene products, with an average of 129 products for each disease entry. Theoretically, there would be a maximum of approximately 21.5 million gene-disease combinations in the OMIM sub-table of the database, considering the 15,294 distinct gene products and 1,407 distinct diseases represented there. Our methods significantly reduce the combinatorial space for these gene-disease relationships,

(i.e., 688,126 rows in the OMIM sub-table) however, by more than an order of magnitude, and therefore provide a more efficient starting point for high-throughput analysis of the complicated genetic interactions underlying complex diseases.

Our results not only reveal the potential utility for the GenesTrace system, but also point to the limitations of the process. The success of the traces is heavily based on the quality and accuracy of annotation for the corresponding gene products [7]. Indeed, the precision and recall values reported in this study may be affected by varying annotations for homologous genes across multiple organisms. Perhaps the most significant current limitation, however, is the need for filtering of results for complex diseases, such as the almost 11,000 gene products retrieved for Alzheimer's Disease. Investigating the most commonly occurring gene products within the set is the most basic approach, but due to the likely low signal-to-noise ratio (e.g., 3 of the top 4 products for the Alzheimer disease query (Aryl Hydrocarbon Receptor; Uracil-DNA Glycosylase; and E2F-1 transcription factor) are involved in neuronal development or apoptosis, but are not specifically implicated in Alzheimer's disease), other methods would need to be applied. Further complicating methods such as ours is the lack of uniformity in gene names or gene product representation, as evidenced by the relatively small number of genes in *OMIM's morbidmap* also present in the GO databases.

Several factors also explain our relatively wide-ranging values for precision and recall. As mentioned above, one of the major reasons may be, the lack of uniform naming of genes. This was particularly evident in mappings between *morbidmap* and the GODB. For example, *morbidmap* has entries for both *VRNF* and *NF1*, which are the same gene, while only *NF1* is in the GODB. Additionally, even with the sub-table of OMIM diseases, the large search space of genes and gene products makes false positives much more likely. Due to the limitations of gene annotations, another possibility is that a set of our FP's are not easily verified FP's, but instead have "buried or undiscovered" relationships to the diseases. For example, of the 242 genes returned as FP's for Alzheimer's disease, 97 return at least one entry in a PubMed search for "[*gene*] AND dementia", where '[*gene*]' is the query gene symbol, and thus may represent examples of previously discovered, yet buried, knowledge. In this case, buried would indicate that the relationship is difficult to retrieve in high throughput, as it is not explicitly annotated in either OMIM or PubMed. However, besides these examples, other inferred relationships may be undiscovered altogether. Similarly, of the 129 FP's for hepatocellular carcinoma, 69 return at least one article (buried knowledge) for a search on "[*gene*] AND hepatoma", where '[*gene*]' is the query gene symbol.

This study has several limitations. We did not exploit the graph structure of the knowledge sources we used. Nor did we contrast putative results as "buried" (if found in PubMed) and "undiscovered" (if not found in PubMed). Appreciation of the graph structure would have enabled us to modify our results based on recursively navigating up or down levels of MRREL, MRCOC, MeSH or GO. We intend to investigate the recursive use of the relationships to expand the database. An additional relevant study that we have initiated is to use statistical comparisons to compare groups of genes mapped to diseases. For instance, housekeeping genes are likely to be associated non-specifically with a large number of diseases. The top five products in terms of frequency, for example, consisted of two stress response genes (*SKN7, SKI1)*, two transcription regulators (*HNT1, E2F1)* and one metabolic regulator (*IATP)*. Future work will include the usage of established information retrieval weighting techniques to stratify results such as term frequency * inverse document frequency (TF*IDF).

The strengths of the GenesTrace system stems from its ability to combine several sources of manually curated information into a high-throughput system for gene discovery. Additionally, very few systems provide genome-wide approaches to phenotypic discovery, and tools, such as GenesTrace, are foundational steps to comprehend the phenome as they provide an incremental discovery path using established knowledge bases. Furthermore, GenesTrace performs searches based both on semantic and statistical information, and has the ability to exploit the ontological properties of the source materials. The system is also easy to update, as the source materials are updated either quarterly or monthly. Finally, though the system explores direct, clinically diverse gene-disease relationships, providing a high-level view, it is also expandable to virtually any source annotated by GO, down to the sequence level, as well as any source incorporated into the UMLS, such as SNOMED CT.

5. Conclusion

The GenesTrace methodology presented here is a proof-of-concept study that mines gene-disease relationships across biomedical databases. By design, the prototypal method was unfiltered. Consequently, while the results of this study provide sufficient accuracy that attests to the validity of the GenesTrace principle, it yet remains inadequate for active use by researchers. However, as the work progresses in planned complementary studies, we expect the methods presented here to have potential for significant accuracy improvements, which may yield a powerful tool for research and discovery. Specific future studies will include machine learning or filtering metrics such as the frequency of co-

occurrences of the intermediating knowledge, and convergence of distinct gene-disease traces ("knowledge pathways").

Understanding the genetics behind complex diseases is one of the principal goals of genomics research. Many complex genetic phenomena are encoded in biological knowledge bases. Similarly, many clinical manifestations of disease are represented in clinical knowledge bases. Multiple levels of both explicit and implicit pathophysiologic and phenomic knowledge may be buried in mappings across mappings between clinical and biological knowledge bases. By examining these mappings, one can envisage automated systems, like GenesTrace, that may help elucidate testable genetic hypotheses by tracing the links between these clinical and biological knowledge bases.

Acknowledgements

These studies were supported in part by the Department of Medicine, BIMC and the following grants: NIH/NLM 1K22 LM008308-01, NIH/NIAID 5U54 AI057158-02, and NIH/NLM LM-07079-09. The authors thank Inderpal Kohli and Jianrong Li for the development of the GenesTrace interface and database system, respectively.

References

1. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**: p. 25-9.
2. Bodenreider, O. *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic Acids Res 2004;**32**, D267-70.
3. Jenssen, T.K., et al., *A literature network of human genes for high throughput analysis of gene expression.* Nature Genetics, 2001. **28**:p. 21-28.
4. Hamosh, A., et al., *Online mendelian inheritance in man.* Human Mutation, 2000. **15**: p. 57-61.
5. Perez-Iratxeta, et al., *Association of genes to genetically inherited diseases using data mining.* Nature Genetics, 2002. **31**: p. 316-9.
6. Hu, Y., et al., *Analysis of genetic and proteomic data using advanced literature mining.* Journal of Proteome Research, 2003. **2**: p. 405-12.
7. Camon, E., et al., *The Gene Ontology Annotation [GOA] Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro.* Genome Research, 2003. **13**: p. 662-72.
8. Raychaudhuri, S., et al., *Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature.* Genome Research, 2002. **12**: p. 203-14.
9. Perez, A.J., et al., *Gene annotation from scientific literature using mappings between keyword systems.* Bioinformatics, 2004. (April 1) [epub].
10. Blois, M.S., *Information holds medicine together.* MD Computing, 1987. **4**: p. 42-6.
11. Bodenreider, O., *Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications, and prevention.* Proc AMIA Symp, 2001: p. 57-61.