Stud Health Technol Inform 2003:62-67.

An Evaluation of Hybrid Methods for Matching Biomedical Terminologies: Mapping the Gene Ontology to the UMLS®

M.N. Cantor[a], I.N. Sarkar[a], R. Gelman[a], F. Hartel[b], O. Bodenreider[c*], Y.A. Lussier[§a*]

a. Department of Medical Informatics, Columbia University New York, NY 10032 USA
b. Center for Bioinformatics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20892 USA
c. National Library of Medicine, NIH, DHHS, Bethesda, MD 20892 USA

Abstract

Integration of disparate biomedical terminologies is becoming increasingly important as links between biological science and clinical medicine grow. Mapping concepts in the Gene Ontology™(GO) to the UMLS may help further this integration and allow for more efficient information exchange among researchers. Using a gold standard of GO term – UMLS concept mappings provided by the NCI, we examined the performance of various published and combined mapping techniques, in order to maximize precision and recall. We found that for the previously published techniques precision varied between (0.61-0.95), and recall varied from (0.65-0.90), whereas for the hybrid techniques, precision varied between (0.66-0.97), and recall from (0.59-0.93). Our study reveals the benefits of using mapping techniques that incorporate domain knowledge, and provides a basis for future approaches to mapping between distinct biomedical vocabularies.

Keywords:

Medical Informatics; Bioinformatics; Gene Ontology, Unified Medical Language System, Mapping Terminologies

1. Introduction

With the rapid advancement of scientific knowledge, the links between clinical medicine and the biological sciences are also growing. The rapid advancement of technology, which has led to the improved ability to analyze complex biological systems, has led to an even greater need for cooperation and integration between the fields [1,2]. A standard method of knowledge representation could streamline interactions between physicians and biological scientists, leading to improved cooperation and possibly a more rapid pace of scientific discovery.

The Unified Medical Language System (UMLS)[1] has the most expansive breadth of many, if not most, existing medical vocabularies. The increasing rate of discovery of genetic information, through projects such as the Human Genome Project (HGP), has led to an increasing need for the representation of genes and gene products in the UMLS, which is mainly focused on clinical medicine. Other authors have explored initial strategies for the mapping of genomic data into the UMLS [3]. The set of UMLS 'Lexical Tools' provided by the National Library of Medicine (NLM) [4] helps with the task, but with varying results. At

* These authors contributed equally to this work
§ Corresponding author
[1] http://umlsinfo.nlm.nih.gov

the time we started this experiment, no terminology developed specifically for bioinformatics (e.g., Gene Ontology)[2] had been integrated into the UMLS[3].

Attempts at mapping other terminologies to the UMLS have resulted in limited success, with match rates of 13-60% [5,6]. A recent attempt to match GO terms to the UMLS, using a lexcio-semantic approach, found a match rate from 2% (for gene symbols), to 44% (for molecular functions) [7]. Additional work has investigated approaches to mapping biomedical resources such as OMIM and GENBANK into medical information structures [8]. This prior work demonstrates both the desire and need for the integration of various terminologies, representing both clinical and basic science. An evaluation of several lexical and information extraction techniques permits analysis on two levels: first, on the inherent compatibility between the candidate vocabularies; and second, on the performance of the techniques themselves.

Recently, we undertook an evaluation of published methods for mapping the GO to the UMLS [9]. We then hypothesized that the combination of these methods, or addition of supplemental information for matching criteria, would improve overall performance, and lead to different combinations of recall and precision. Knowledge of these values could be used when considering different query situations. For example, manually curated queries over a small data space would benefit from a balance favoring recall, while those over a larger dataset would benefit from higher precision.

The objective of this study is to find the most effective method for accurately capturing and mapping the largest possible subset of GO to UMLS CUIs. With this in mind, we provide a quantitative evaluation of five different, previously published approaches to mapping the two terminologies [4,7], as well as hybrid approaches that incorporate different combinations of the methods.

2. Materials and Methods

Since the gold standard employed in this study was created in 2001, we used the 2001 version of the UMLS, created and maintained by the National Library of Medicine. The 2001 version of the UMLS Metathesaurus® consists of about 800,000 unique concepts (797,359) from over 60 diverse terminologies[4]. A 'Concept Unique Identifier' (CUI) represents each individual concept in the UMLS. Since there may be multiple text string variants (UMLS terms) affiliated with each CUI, the variants are identified by 'String Unique Identifiers' (SUIs). In the 2001 UMLS there are 1,728,075 SUIs.

The purpose of the Gene Ontology is to develop a coded, structured vocabulary for molecular function, biological processes, and cellular components that can be used across all species [10]. A unique text string, known as a GO term, represents each GO concept, which is in turn referenced by a unique "accession number" (GOID). Significantly, the component sub-vocabularies are independent of the associations between specific gene products and GO terms, leading to flexibility and precision in the use of the framework. For this study, we used the May 2001 version of the GO, since that was the version used by our gold standard.

As a *gold standard* (GS) for mappings between the vocabularies, this study used files provided by The National Cancer Institute (NCI), which contained mappings between a subset of GO (NCI-GO version May 2001), the UMLS, and the NCI's internal metathesaurus. From

[2] http://www.geneontology.org
[3] NLM recently announced the inclusion of GO among the UMLS source vocabularies for mid-2003
[4] http://umlsks.nlm.nih.gov

these files, we derived a subset of 332 distinct GOIDs that had been mapped to CUIs from the UMLS. Of note, the GS contained 314 distinct CUIs, indicating that some CUIs mapped to more than one GOID. Additionally, 6,113 SUIs are associated with these 314 individual CUIs.

Published Methods. Of the four individual methods we used, the simplest method was exact string matching, which finds the lexical matches between UMLS terms and GO terms. Two of the other methods were implementations of lexical tools available from the UMLS(4): *norm* and MMTx, an implementation of MetaMap [11]. After processing with *norm*, we matched each distinct GO term against the normalized form of UMLS terms. We also performed two types of analyses using MMTx, which we termed 'Strict' and 'Loose'. In our 'Strict' MMTx analysis, we considered only GOID-CUI pairs returned with a MetaMap score of 1000. In contrast, 'Loose' MMTx analysis consists of all distinct GOID-CUI pairs provided by the 'Meta-Mappings' regardless of their scores.

For the lexico-semantic (LS) approach we used experimental information on GO-UMLS mappings supplied by one of the authors. These mappings take into account matches on both the lexical as well as the semantic level, as previously described [7]. Specifically, there is an attempt to find an exact lexical match; if none is found, there is an attempt to find a normalized match. After a potential match is found using either method, the semantic constraints are applied, in order to verify that both concepts are of compatible semantic types. These constraints are based on mappings between the three categories in GO and the semantic types present in the UMLS as described in [7]. Using this information, we created a subset of our full GO-UMLS *norm* mapping, eliminating those terms that were not matched both semantically and lexically.

Hybrid Methods. Our hybrid strategies consisted of two distinct approaches, both of which combined the results from the above methods. The first approach involved incremental application of each method to the starting set of terms. For example, we first applied the exact string matching technique, then *norm* or *norm+LS* to the remaining, unmatched terms, and finally 'Strict' MMTx, used last because of its more sophisticated algorithm, keeping cumulative scores of recall and precision. For the incremental techniques involving semantic information, the second step was the equivalent of combining *norm* and the above lexico-semantic techniques. The second approach consisted of using a voting scheme among the three methods. In this case, we looked at the broad set of terms for which there was a correct match using any of the techniques, as well as the more restrictive set, for which the same correct match emerged in all of the techniques.

Using the GS GOID-CUI concept pairs as our gold standard, we performed an evaluation of the precision and recall of each of the methodologies implemented. In our examination of the GS, we defined our results as follows: relevant pairs ('True Positive'; TP) were pairs found by the coupling method that were also in the GS; non-relevant ('False Positive'; FP) matches were those that were not found in the GS; relevant, but *not* retrieved ('False Negative'; FN) were in the original GS but not matched by the coupling method.

3. Results

Tables 1,2 and figure 1 summarize the analysis of our results using each of the above hybrid

Table 1: Results of Published Methods

Type of Match	Exact String	Norm	MMTx Loose	MMTx T=1000	Lexico-semantic (LS)
Relevant GS Matches (TP)	216	299	272	247	286
Non-Relevant GS Matches (FP)	13	38	170	12	26
Not Retrieved GS Matches (FN)	116	33	60	85	46
Recall/Precision	0.65/ 0.94	0.90/ 0.89	0.82/ 0.61	0.74/ 0.95	0.86/ 0.92

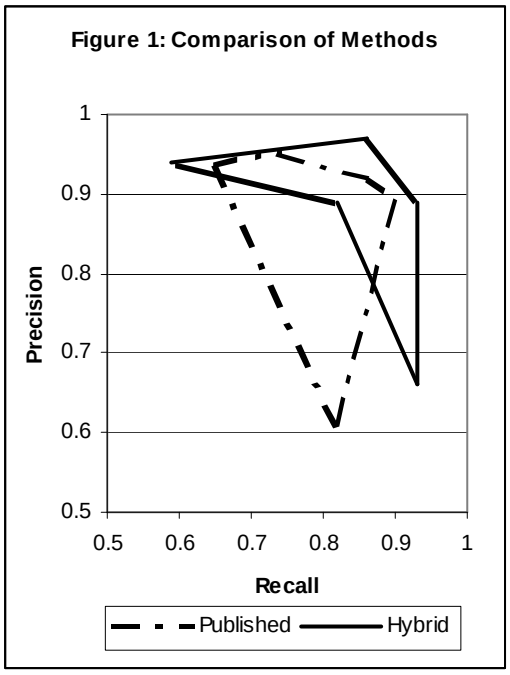

Figure 1: Graphs representing the overall results between the two sets of methods, published and hybrid. Inflections in the graph representing hybrid methods may not exactly match the results in Table 2 due to merging of redundant data points. The hybrid graph represents data points in the incremental methods, as well as the combined results of the "voting" methods.

methodologies, as well as the published methods. Table 1 shows the results, as well as the recall and precision values, for each of the published methods. Table 2 shows the same results for the hybrid methods, both with and without incorporating semantic information. Figure 1 gives a general view of the range of precision and recall for each larger set of methods. As seen in Table 2, incorporating semantic information into the technique initially reduced the set of exact matches; however, both sets of incremental methods provided steadily increasing recall, with only a relatively small drop in precision.

Performance of the voting schemes was somewhat similar to that of the incremental methods. The "union" of methods resulted in a high recall (94%), but suffered in precision (65%), mainly due to the false positives returned by MMTx. Conversely, the methods' "intersection" performed similar to *norm*, with recall of 82% and precision of 89%. Adding semantic information to the intersection of methods increased precision, to 94%, but produced a large drop in recall, to 59%. As with the other semantic methods, this is mainly due to the false positives eliminated in the component steps. Of all the techniques, the highest combination of recall and precision is observed with the *norm+LS* technique. Overall, however, the benefit of the use of hybrid methods is the fact that at the highest levels of precision, they result in an increase in recall of approximately 14%, as seen in Figure 1.

4. Discussion

Since this study demonstrates the feasibility of mapping GO terms to the UMLS, it is conceivable that using a combination of our methods, as well as others that may exist, will enable a complete mapping of these terminologies as well as any future terminology from other related domains. The varied results returned from our hybrid methods may reveal both the strength of the individual lexical tools, such as *norm*; possible weaknesses of the GS, such as

misclassified FP's; and, finally, the possible existence of a core group of terms that are extremely difficult to code in different vocabularies. Additionally, the results reflect the

Table 2: Results of Hybrid Methods

	Methods Without Semantic Information					Semantic Methods				
	Incremental (cumulative scores)			Voting		Incremental (cumulative scores)			Voting	
Type of Match	**EM**	**Norm**	**MMTx**	**U**	**I**	**EM**	**Norm + LS**	**MMTx**	**U**	**I**
Relevant GS Matches (TP)	216	299	310	310	272	204	284	308	308	195
Non-Relevant GS Matches (FP)	13	19	39	157	34	4	9	37	94	13
Non-Retrieved GS Matches(FN)	116	33	22	22	60	128	48	24	24	137
Recall/Precision	0.65 / 0.94	0.90/ 0.94	0.93 / 0.89	0.93 / 0.66	0.82 / 0.89	0.61 / 0.95	0.86 / 0.97	0.92 / 0.89	0.93 / 0.77	0.59/ 0.94

EM= Exact Match; U= Union; I= Intersection

complexity inherent in representing ideas in biomedicine, both in logical, contextual, and biological terms.

One limitation of this study is the fact that the GS, as well as all methods other than LS, used 2001 versions of both GO and the UMLS, while the LS method is based on GO from February of 2002 and the UMLS version 2002AA. Another potential limitation is the structure of the gold standard used in this study, which may have biased the results of the evaluation. The dataset is curated and maintained by the NCI, and is specific to one domain of medicine (oncology). With the large number of potential SUIs for each single term, it is unreasonable to expect manual curation of all combinations. Missing combinations may also have contributed to an inflated number of FP's.

The potentially misclassified FP's are emblematic of the complexity of mapping terms to a composite terminology such as the UMLS. Adding biological terminologies to the UMLS is often a more complex task than adding terms from other medical vocabularies. Normalization of text strings, for example, de-emphasizes numeric values, which are generally much more important in biology than in clinical medicine. One potential approach to this problem of ambiguity is to use the properties of the underlying ontology of a terminology, such as the UMLS Semantic Network, to apply formal predicate logic to relationships among concepts.

5. Conclusions

As the quantity of genetic information continues to grow, and its implications for clinical medicine become apparent, the need for seamlessly integrated biomedical vocabularies will be increasingly manifest. Fully mapping the GO to the UMLS, for example, could allow for the exploitation of the UMLS semantic network to link disparate genes, through their annotation in GO, to unique clinical outcomes, potentially uncovering biological relationships.

This study reveals the inherent difficulties in the integration of vocabularies created in different manners and by specialists in different fields, as well as the strengths of different

techniques used to accomplish this integration. While existing lexical methods perform adequately, techniques that add semantic and other contextual information, such as *norm+LS* and MMTx, provide better precision. Using automated, high-throughput techniques allows for faster creation of mappings between vocabularies, as compared with manual curation, and may save considerable time and effort. Additionally, with increasingly efficient methods, one may implement high-throughput mapping techniques and be able to manually verify their accuracy on a random subset of the terminologies, as long as the sample size gives sufficient statistical power.

Depending on the goals of a vocabulary-mapping project, i.e. whether emphasis is placed on the breadth or accuracy of the resulting mappings, researchers have a variety of techniques from which to choose. Results closer to 100% for precision or recall, however, will require even more sophisticated Natural Language Processing techniques, in addition to at least some human intervention. Links to other vocabularies through the UMLS could also allow for the discovery of new relationships, such as relationships between clusters of genes and clinical phenotypes, using connections to SNOMED, for example. Using proven text mining and other information extraction techniques will allow for further mappings of existing knowledge resources to each other, potentially leading to an accelerated rate of scientific discovery.

Acknowledgements

INS and MNC are funded by National Library of Medicine Medical Informatics Training Grant LM07079-09. This work has also been supported by an NYSTAR grant.

References

[1] Altman RB, Klein TE. Challenges for biomedical informatics and pharmacogenomics. Annual Review of Pharmacology & Toxicology 2002;42:113-133.

[2] Shortliffe E, (eds). PL. Medical Informatics: Computer Applications in Health Care and Biomedicine. New York: Springer; 2001.

[3] Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS semantic network. Proc Am Med Inf Assoc 1999:181-186.

[4] National Library of Medicine. UMLS Lexical Tools. Available at http://umlsks.nlm.nih.gov

[5] Tuttle MS, Suarez-Munist ON, Olsen NE, et al. Merging terminologies. MEDINFO 1995;8(part1)(1):162-166.

[6] Zeng Q, Cimino JJ. Mapping medical vocabularies to the UMLS. Proc Am Med Inf Assoc 1996:730-4.

[7] Bodenreider O, Mitchell JA, McCray AT. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. Proc Am Med Inf Assoc 2002:61-65.

[8] Sperzel WD, Abarbanel RM, Nelson SJ, et al. Biomedical database interconnectivity: An experiment linking MIM, GENBANK, and META-1 via MEDLINE. Proc Annu Symp Comput Appl Med Care 1991:190-193.

[9] Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking biomedical information and knowledge resources: GO and UMLS. Proc Pac Symp Biocomputing 2003.

[10] Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. Gen Res. 2001;11(8):1425-1433.

[11] Aronson AR. Effective mapping of biomedical text to the UMLS. Proc Am Med Inf Assoc 2001:17-21.

6. Address for Correspondence

Yves Lussier, VC-5 Medical Informatics, 622 West 168th Street, New York, NY 10032
email: lussier@dmi.columbia.edu