MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Automated Terminology Networks for the Integration of Heterogeneous Databases

Xiaoyan Wang[a], Hui Nar Quek[a], Michael Cantor[a,b], Pauline Kra[a], Aylit Schultz[b], Yves A. Lussier[a,b]

[a] Department of Biomedical Informatics, [b] Department of Medicine,
College of Physicians and Surgeons, Columbia University, New York, NY 10032, US

Abstract

As cross-disciplinary research escalates, researchers are facing the challenge of linking disparate biomedical databases that have been developed without common indexes. Manually indexing these large-scale databases is laborious and often impractical. Solutions involving mediating terminologies have been proposed, but coordination of terms from the databases of interest to these mediating terminologies is also laborious, and regular synchronization between indexes is an additional problem. In this study we describe a novel method of linking heterogeneous databases using terminology networks constructed with automated mapping methods. Linkage was established between two disparate biomedical databases (SNOMED-CT and HDG), using two relevant intermediating databases (UMLS and OMIM). One gold standard of 514 distinct matches is used as proof-of-principle. In our study, the fully manually curated network (baseline index) and one automated terminological pathway (HDG-OMIM-SNOMED) perform at high precision and low recall, while the direct automated terminological pathway (HDG-SNOMED) provides higher recall and lower precision. In conclusion, as hypothesized, 1) Manually curated pathways provide high precision, but offer low recall, 2) the automated terminology pathways can significantly increase recall at acceptable precision. Taken together, our conclusion may suggest the combined manual and automated terminology networks could offer recall and precision in an incremental manner.

Keywords:

networks, databases, terminologies, system integration.

Background and Significance

One of the important goals of biomedicine is to represent and integrate knowledge in the fields of molecular biology and clinical medicine comprehensively so that the databases and applications of each field can enrich one another [1,2]. However, rapidly evolving biomedical databases present an unprecedented problem of integration in order to retrieve useful information across different domains.

The difficulty in tackling the database interoperation problem can be attributed to various reasons [3, 4]. First, data is represented heterogeneously in different databases since each database maintains its own data and provides its own interface independently. This problem of heterogeneous data afflicts the representation schema, as well as the scale and the granularity of the data. In addition, capabilities and formats of each database vary. Finally, naming conventions and standards are so different across fields that common indexes and terminology are rarely co-developed and shared. For example, linking biological terminologies to medical terminologies is quite different from linking medical terminologies among themselves [5]. On the other hand, the Unified Medical Language System (UMLS®) is addressing this problem by providing relationships across an increasing number of medical and biological terminologies (e.g. Gene Ontology® (http://www.geneontology.org) has been integrated into the UMLS in 2003 AB version of the UMLS). However, concurrency and synchronization between ever evolving terminologies remains an important, costly and time-consuming issue in metathesauri.

The database intermediation problem can be addressed via the creation of a mediated schema - a set of virtual relations and mappings among 2 or more diverse data sources. The mediated schema can be used to translate a single query into the appropriate format for each specific resource, and then consolidate the various return values into a single, easily interpretable result set. Such strategies can capitalize on the increasingly comprehensive UMLS as a mediating terminology. Towards this end, an earlier work investigated the linkage of MIM, GENEBANK and the UMLS [6]. However, updates in source terminologies occur more frequently than updates to the UMLS, leading to a mediated schema that lags behind updates in source databases. In addition, our group has also undertaken a variety of lexico-semantic techniques, including the incremental use of hybrid techniques intended to automatically link two terminologies [7]. Limited effort has been deployed for the development of high throughput methods for the linkage of rapidly evolving biomedical databases, and to our knowledge, none have explored computational methods involving automated networks of terminologies to dynamically generate putative indexes.

The objective of this feasibility study is to demonstrate the significance of automated terminology networks to dynamically map rapidly evolving heterogeneous biomedical databases that do not share complete cross-indexes.

Materials

Databases to be cross-indexed:

(i) *SNOMED-CT* [8] is a comprehensive concept-based health care terminology. We used the version released in July 2002. This version of SNOMED-CT contains 333,325 concepts. SNOMED-CT contains a cross-index with the older version of SNOMED 3.5 which contains about half as many concepts. For each SNOMED concept, there is one concept term and there may

be several synonym terms associated with the concept as well. SNOMED-CT will be added to UMLS- Metathesaurus and available under certain conditions worldwide.

(ii) HDG [9] is a manually compiled database of human disease genes. For each distinct disease gene record, HDG contains at least one disease name (term); each of the 921 disease gene records of HDG is also mapped to an OMIM unique identifier (concept). The full database has been published in the journal *Nature*, and is available publicly [9].

Intermediating Terminologies:

(iii) *OMIM* [10] is a catalog of human genes and genetic disorders. OMIM focuses primarily on inherited and heritable genetic diseases. We used the 2002 version of OMIM which contains 14280 entries, including 8733 human gene loci. Each OMIM unique concept identifier contains two distinct fields in which disease terms are found: the "*Title*", and the "*Disorder*". The "*Title term*" field contains gene products and diseases with no semantic class to distinguish between the terms, while all disorder terms can be considered as one semantic class subsumed by "diseases".

(iv) *UMLS.* We used the 2002AB version of the UMLS, created and maintained by the National Library of Medicine. This version consists of 871,584 unique concepts over 60 diverse terminologies. For each UMLS concept, there is one concept term and there may be several synonym terms associated with the concept as well. Disease terms of UMLS are grouped together as a semantic class. The UMLS Metathesaurus includes 208,454 concepts linked to SNOMED International 3.5 (1998 version) and 250 concepts linked directly to terms of OMIM (1993 version).

Methods

Creating networks to link disparate databases

Networks between databases can be manually curated (e.g. via shared cross-indexes) or automated (e.g. via lexical or semantic computational methods). When concept mapping occurs at the stage of indexing or cataloging and is conducted manually, we will refer to this practice as "manual curation" (MC). In contrast, "automated mapping" (AM) will refer to the mapping of terms associated with the concepts of two terminologies using computational methods. Figure 1 illustrates a network of terminology relationships between the databases to be cross-indexed (HDG, SNOMED-CT) and the intermediating terminologies (OMIM, SNOMED 3.5, UMLS). The arrows in the figure show the available types of mapping (MC, AM). Automated mapping was conducted using previously published methods that include lexical and semantic constraints as described below. Several properties of the terminology network have been explored including types of mapping and number of intermediaries. Distinct mapping strategies generate different types of terminology paths that we categorized as follows: (a) purely MC-based (Table 1, P1), (b) purely AM-based (table 1, P2-7). Combined AM-based and MC-based pathways are beyond the scope of the current study. Similarly the number of intermediating terminologies investigated were: (a) zero (Table1, P2), (b) one (Table1, P3, P4, P5), (b) two (Table1, P6, P7), and (b) three (Table1, P1).

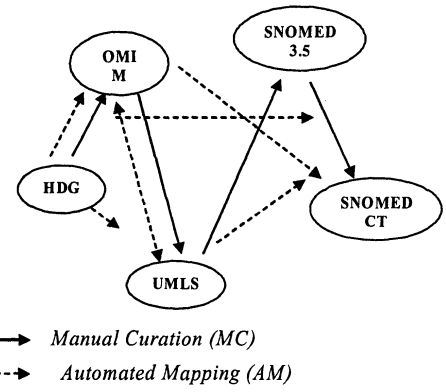

Figure 1 - The network created to link disparate databases

Automated Mapping is performed using two widely published lexical methods: exact matching (EM) and the National Library of Medicine Normalization (Norm) matching categorization of both the original terms & the target concepts to exclude semantically irrelevant mappings. EM and Norm have been used with semantic constraints and the resulting *lexico-semantic methods* have been shown to be more accurate in Bodenreider's laboratory [15] and in ours [7, 13]. For this project, valid AM is semantically relevant when SNOMED-CT terms are diseases/disorders (descendants of SNOMED-CT code 64572001).

Evaluation of terminology pathways

A Gold Standard (GS) linking HDG to SNOMED was produced by the agreement of two experienced knowledge engineers working independently at mapping every HDG concept to SNOMED concepts, which were used as GS. Agreement was observed for 514 distinct HDG records.

First Quantitative evaluation: Accuracy of Concept Maps (ACo). We measured the accuracy of each of the mapping methods described in Table 1 using precision, recall and general accuracy in the resulting HDG-SNOMED concept pairs. As lexico-semantic methods evaluate term-pairs, they are further transformed in a concept-oriented view (since multiple terms can be associated in one concept in SNOMED-CT and in HDG). Each of the mapping methodologies was compared to the gold standard. Relevant pairs ('True Positive'; TP) are pairs found by the linking method that are also in the GS; non-relevant ('False Positive'; FP) matches are those that are not found in the GS; relevant, but *not* retrieved ('False Negative'; FN) are in the GS but not matched by the linking method. Non-relevant, but not retrieved pairs ('True Negatives'; TN) are neither in the GS nor matched by the linking method. In this experimental setup, TN = (total_HDG-SNOMED_pairs)-(TP+FN+FP), where the total_HDG-SNOMED_pairs are the combinations between all HDG and SNOMED concepts used in the mapping experiment (514 HDG concepts and 70,831 disease concepts in SNOMED-CT). Recall was calculated as the ratio of the number of distinct HDG-SNOMED concept pairs that were identified by the mapping method that match HDG-SNOMED concept pairs in the Gold Standard (GS), divided by the total number of pairs In the

Table 1: Linking Paths derived from the network

Path name	Intermediating terminologies (#)	Complete Path
P1	3	HDG = OMIM = UMLS = SNOMED3.5=SNOMED-CT
P2	0	HDG → SNOMED-CT
P3	1	HDG → UMLS → SNOMED-CT
P4	1	HDG → OMIM (Disease terms) → SNOMED-CT
P5	1	HDG → OMIM (Title terms) → SNOMED-CT
P6	2	HDG → UMLS → OMIM → SNOMED-CT
P7	2	HDG → OMIM → UMLS → SNOMED-CT

A = B Manual Curation / Mapping of terms via a common index between databases A and B.
A→B **Automated Mapping / lexico-semantic mapping of terms between databases A and B.**

GS, TP/(TP+FN) [16]. Precision was measured as the ratio of the number of distinct HDG-SNOMED concept pairs returned by the mapping method that match HDG-SNOMED concept pairs in the GS, divided by the total number of putative HDG-SNOMED concept pairs found by the mapping method, TP/(TP+FP) [16]. General accuracy is calculated as (TP+TN)/(TP+TN+FP+FN).

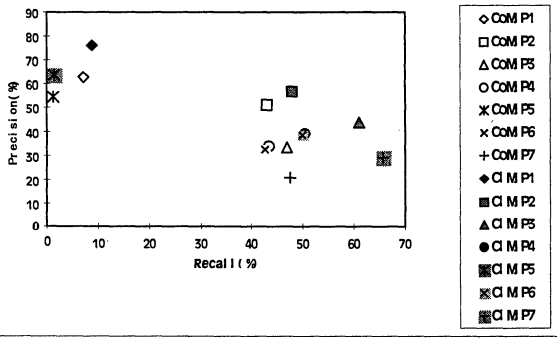

Figure 2 - Precision versus recall of each of the linking paths in the network

The computational methods that we propose each provide one-point outputs on a precision-recall graph, which remains a succinct and clear way of presenting our results. While the traditional use of the precision-recall graph generally requires methods that can produce an 11-point curve, more than one point output per method is not applicable in this experimental setup.

Second Quantitative evaluation: Accuracy of Class-Based Map (Acl). Due to the high level of granularity of the SNOMED terminology, an additional accuracy score was calculated for the class of a concept. For the purpose of this score, the mapping of a HDG concept to an ancestor or a descendant of the associated SNOMED concept in the GS was considered a "True-positive" class-based mapping. Recall, precision and general accuaray were calculated on this basis.

Qualitative evaluation. Intermediating terminology pathway terms and concepts are not evaluated in the accuracy score based on HDG-SNOMED concept pairs. We have therefore manually analyzed the full pathway maps of the manual curation pathway (P1) and sample that of the automated mapping techniques.

Results

Concept-based Quantitative Evaluation. As described in Figure 1 and Table 1, manual curation utilizes the internal mapping of OMIM and SNOMED 3.5 in UMLS, which simulates the linking of HDG and SNOMED via a common and pre-existing index, and hence sets the baseline for the performance of paths derived from the network. The accuracies of concept map and class-based map using the network are summarized in Figure 2. As how the CoM and ClM are defined in this study, CoM is a subclass of ClM and therefore not independent. Our analysis shows that the manually curated pathways provided the highest precision (62.7% and 76.2% for CoM and ClM, respectively), and the poorest recall (7.1% for CoM, 8.7% for ClM). The direct mapping of HDG to SNOMED (P2) provided an intermediate accuracy as compared to other techniques (42.9% for recall and 50% for precision using CoM). Paths involving one level of intermediating terminologies either gave higher recall (such as P3 and P4) with the sacrifice of precision, or vice versa (P5), as compared to the direct path (P2). Both paths (P6 and P7) containing two levels of intermediating terminologies give higher recall but lower precision, compared to the direct path.

Class-based Quantitative Evaluation. The ancestor-descendent relationships in SNOMED-CT allow us to explore the class-based mapping when an exact matching pair is not available from source to target databases. As is seen in Figure2, all pathways show increased recall and precision with the class-based accuracy, some showing better improvements than other (e.g. P5's precision increases from 54.5% to 63.6%, P7's recall increase from 47.47% to 65.75%).

The general accuracies of all the terminology pathways are 99,99X%, where X varies from one terminology pathway to another. The reason for which the general accuracy is of little value to discriminate between different terminological pathways is that the enormous TN count (ref. Methods) appears both at the numerator and denominator of the calculation, for each pathway.

Qualitative Evaluation. Every mismatched (according to the GS) HDG-SNOMED-CT pair of concepts was manually reviewed in the MC set P1.

Table 2: Categories of Mismatches Observed in the P1 pathway from the terminology network

Category of Mismatch		Example of Each Category		
Name	**Count (%)**	**HDG Disease**	**SNOMED Disease in the automating matching**	**SNOMED Disease in the Gold Standard**
Ret	36	Galactosemia (230400)	Galactosemia (disorder) [Ambiguous] (38177000)	Galactosemia (disorder) (190745006)
CBI	42	Pseudohypoparathyroidism, type Ia (103580)	Pseudohypoparathyroidism (disorder) (58976002)	Pseudohypoparathyroidism type I A (disorder) (58833000)
Amb	12	Meningioma,NF2 related, sporadic, Schwannoma, sporadic (101000)	Neurofibromatosis, type 2 (disorder) (92503002)	Intracranial meningioma (disorder) (302820008)
Red	10	Apert syndrome (101200)	Apert's syndrome (disorder) (63661009)	Acrocephalosyndactyly (disorder) (268262006)

In addition, a subset of the mismatched pairs of other sets was also manually curated. Table 2 shows examples of these mismatches taken from P1 that can be categorized in four classes: (i) *retired concepts (Ret) in SNOMED* [17]. 36% of mismatches in P1 are attributable to concepts in SNOMED 3.5 (UMLS concepts that have been retired in SNOMED-CT) (e.g. Table 2 Ret, an ambiguous concept (38177000) in SNOMED 3.5 has been replaced in SNOMED-CT with a new concept 190745006, which is not reflected in UMLS); (ii) *Class-based indexing (CBI)* in MC (e.g. Table 2, the network finds the ancestor of the matching concept in SNOMED-CT), 42% of mismatches fall in this category for CoM, and are considered matched by the ClM; *(iii) Ambiguity (Amb) in HDG.* More than one concept shares the same code in the database (e.g. Table 2 Amb, two disease sharing the same MIM number in HDG), 12% of mismatches in P1 are ambiguous; and (iv) *Redundancy in SNOMED (Red).* More than one concept shares the same meaning in a terminology and is represented by multiple codes (e.g. Table 2, "Apert syndrome " has been modeled in two different concepts in SNOMED-CT). About 10% of mismatches in P1 are redundant.

Discussion

As hypothesized, the manual curation provided high precision and low recall, probably due to the rapidly evolving OMIM and SNOMED-CT terminologies, (each terminology has more than doubled since its inclusion in UMLS). Generally, in the automated maps, multiple pathways did not lead to higher precision, probably due to increased noise for each successive automated map. In contrast, in the manually curated map the precision remained high regardless of its highest number of intermediating terminologies. Interestingly, one automated pathway (P5) provided a precision approaching that of the manual curation. Notably, P4 and P5 used the same intermediary pathway but different terminology fields. P4 used a field containing unique diseases and disorders, while P5 used a term field which also contained gene products. Surprisingly, P5 outperformed P4 while no semantic constraints could be fabricated over P5 since OMIM does not have semantic classes. One explanation could be that the "Title" field of OMIM is more often explored than the "Disease" field and therefore more "normalized" due to increased feedback from the community of OMIM users.

We have also demonstrated that terminology pathways are non-commutative methods. In P6 and P7 the same terminologies were used in different sequence resulting in better precision for P6 and better recall for P7. This specific result might be explained by the fact that concepts in OMIM did not benefit from as much feedback and knowledge HDG, which engineering as the ones in UMLS, thus coupling has known ambiguities, with OMIM leads to lower precision than coupling HDG with UMLS.

Our group has previously reported increased accuracy using multiple strategies with automated direct mapping methods between two terminologies [7, 13]. Using this previously published incremental approach combined with terminology pathways (e.g. P1, P5, P6), increases the recall to 48.1% and the accuracy to 53%, results comparable to P2. The class-based evaluation measures the mapping of one terminology to a class of the second. Every mapping technique from MC to AM, regardless of its number of intermediating terminologies, was improved, which probably indicate that HDG and SNOMED have different granularities or scope for their concepts and that mapping conceptually from one to another is impractical for many concepts, while classifying a concept of one terminology into the other is attainable as shown by the higher accuracy rate of the class-based evaluation. This observation also brings up an intrinsic problem with conceptual mapping, including MC methathesaurus-based approach.

Limitations. The precision of the P1 mapping could be improved by translating retired SNOMED 3.5 concepts into current ones using relationships from SNOMED-CT which point retired concepts to their current equivalents (when available). Another limitation of the study was that we did not evaluate concepts from HDG which could be mapped to SNOMED-CT using multiple terms. Therefore the problem of AM in a compositional terminology has not been addressed with these methods. Our results suggest that additional research is required for automated linkage of databases with terminology networks. Additional pathways should also be explored and additional mediation of databases in order to draw conclusions and develop reliable predictors of increased precision.

Conclusions

While the manual curation (e.g. UMLS Methathesaurus) approach remain the gold standard for integrating terminologies, it is rate limiting. This study reveals the feasibility of using automated networks of terminologies to accomplish terminology integration in support of database intermediation. More specifically, incremental usage of terminology pathways can increase the overall precision of an automated multi-strategy method intended to intermediate databases. Automated terminology pathways allow for high-throughput linkages between disparate biomedical databases when manual curation is prohibitive. While sampling a subset of the automated network would present affordable means to evaluate the precision and recall of distinct automated pathways, automated predictive algorithms could also be used to further improve the throughput and reduce the cost [18]. Considering the escalating number of biomedical databases, the potential to accelerate discovery science warrants further investigations. We are currently conducting additional studies 1) to investigate the accuracies of combined MC and AM terminology paths: the primary analysis shows that both recall and precision increase in an incremental manner; and 2) to support compositional mapping, such as when one concept in OMIM is completely and non-redundantly represented by the composition of 2 SNOMED concepts; and 3) to predict the accuracy of alternate terminology pathways of large-scale networks.

Acknowledgements

This project was partially supported under grant 1U54 AI 57159-01 of the National Institute of Allergy and Infectious Diseases (NIAID), the New York State Office of Science, Technology, and Academic Research (NYSTAR)-sponsored Center for Advanced Technology at Columbia University Grant C020054.1425-33 the 1DIB TM 00043-01 contract from the Office of Advanced Telemedicine (OAT) of the Health Resources and Services Administration (HRSA), and the 528753/PO P417322 contract with Virginia Commonwealth University's Medical Informatics and Technology Applications Consortium, a National Aeronautics and Space Administration (NASA) Commercial Space Center.

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] Sujansky W. Heterogeneous database integration in biomedicine. *Jnl Biomed Informatics* 2001;34:285-298.

[4] Stead W, Miller R, Musen M, Hersh W. Integration and beyond: Linking information from disparate sources into workflow. *JAMIA* 2000;7:135-45.

[5] Mork P, Halevy A, Tarczay-Hornoch P. A model for data integration systems of biomedical data applied to online genetic databases. *Proc AMIA Symp* 2001:473-7.

[6] 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

[7] Cantor MN, Sarkar IN, Gelman R, Hartel F, Bodenreider O, Lussier YA. An evaluation of hybridmethods for matching biomedical terminologies: Mapping the Gene Ontology to the UMLS. *Stud Health Technol Inform.* 2003;95:62-7.

[8] Spackman K. SNOMED RT and SNOMED CT. Promise of an international clinical terminology. *MD Computing* 2000;17(6):29.

[9] Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. *Nature* 2001;409:853-5. http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v409/n6822/full/409853a0_fs.html&content_filetype=pdf).

[10]McKusick-Nathans. Online mendelian inheritance in man. Johns Hopkins University and NCBI. Available at http://ncbi.nlm.nih.gov/omim. 2000.

[11]National Library of Medicine. UMLS Lexical Tools. Application and Documentation available at http://umlsks.nlm.nih.gov.

[12]National Library of Medicine. UMLS Lexical Tools. Application and Documentation available at http://umlsks.nlm.nih.gov.

[13]Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking biomedical information and knowledge resources: GO and UMLS. *Pac Symp Biocomputing* 2003;8:427-50.

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

[15]Bodenreider O, Mitchell JA, McCray AT. Evaluation of the UMLS as a Terminology and Knowledge Resource for Biomedical Informatics. *Proc AMIA Symp* 2002:61.

[16]Hersh R. W., *Information Retrieval: A Health Care Perspective.* Springer Verlag; 2nd edition 2002.

[17]Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med.* 1998 Nov; 37 (4-5): 394-403.

[18]Bodenreider O. Strength in numbers: Exploring redundancy in hierarchical relations across biomedical terminologies. *Proceedings of AMIA Annual Symposium* 2003:101-105.

Address for correspondence

Yves A. Lussier, M.D

622 W168th Street, VC5 New York, NY, 10032

yves.lussier@dbmi.columbia.edu