Integration of Neuroimaging and Microarray Datasets through Mapping and Model-Theoretic Semantic Decomposition of Unstructured Phenotypes

Spiro P. Pantazatos, BS[1a], Jianrong Li, MSc[2a], Paul Pavlidis, PhD[1b], Yves A. Lussier, MD[2c]
[1]Dept. of Biomedical Informatics, Columbia University, New York, NY, USA; [2]Center for Biomedical Informatics, Department of Medicine, University of Chicago, Chicago, IL, USA

Abstract

An approach towards heterogeneous neuroscience dataset integration is proposed that uses Natural Language Processing (NLP) and a knowledge-based phenotype organizer system (PhenOS) to link ontology-anchored terms to underlying data from each database, and then maps these terms based on a computable model of disease (SNOMED CT®). The approach was implemented using sample datasets from fMRIDC, GEO and Neuronames and allowed for complex queries such as "List all disorders with a finding site of brain region X, and then find the semantically related references in all participating databases based on the ontological model of the disease or its anatomical and morphological attributes". Precision of the NLP-derived coding of the unstructured phenotypes in each datasets was 88% (n=50), and precision of the semantic mapping between these terms across datasets was 98% (n=100). To our knowledge, this is the first example of the use of both semantic decomposition of disease relationships and hierarchical information found in ontologies to integrate heterogeneous phenotypes across clinical and molecular datasets.

Introduction

Increasingly, there is an understanding that well-managed, comprehensive databases and their interoperability will be necessary for important further advancement in neuroscience [1]. However, in contrast to the reliance on and advancements of informatics in other biosciences, such as molecular biology and genomics, for which data is primarily text-based, the tremendous complexity of neuroscience data is a major impediment in consistent informatics integration and implementation [2]. There have been many proposed solutions to this problem, most of which rely on the labor-intensive and time-consuming development of compatible metadata models of phenotypes that formally describe entities, attributes and the relationships between them in the underlying data (see http://phenos.bsd.uchicago.edu/public/supplement-1-AMIA2009.doc, hereafter referred to as *supplement*). One promising and complementary approach has been to use Ontologies employing Description Logic (DL), such as those that have been introduced into biomedical domains, as a flexible and powerful way to capture and classify biological concepts and potentially be used for making inferences from biological data [3, 4].

A major challenge to the use of DL ontologies in mediating between diverse databases is the differences in concepts and terms used to describe the underlying data in each database [5]. This has been addressed by the development of automated methods for the lexical mapping of terminologies and medical vocabularies onto a major medical DL ontology used to link disparate information systems, typically the UMLS [6-8], but also SNOMED as was recently done for ontology-based query of tissue microarray data [9].

The current effort differs from previous approaches because we are mapping very distinct datasets (that may not share many concepts) to SNOMED, which allows for the use of both hierarchical relationships and semantic decomposition between the anatomies and morphologies related to a disease to find relevant relationships across scales of biology. In effect, the proposed approach is also more effectively utilizing a 'reference model' of disease, such as that contained in SNOMED.

Materials and Methods

This paper presents a query model that can be thought of as an equivalent of a *mediated schema* [10] (described in *supplement*) that was created for the genetics domain, but one adapted for higher

[a] These authors contributed equally to this work.
[b] Current affiliation: Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada
[c] To whom correspondence should be addressed. Email: Lussier@uchicago.edu

relevance and utility for neuroscience. Given the wide range of biological scales, heterogeneous data types and contexts in neuroscience, it would be too difficult to map out all relevant entities and the relationships between them as was done for *mediated schema.* Instead, we chose to adapt a pre-existing, comprehensive ontology as our semantic model and explored how to best utilize it to allow for flexible and useful query formulation in neuroscience applications. SNOMED CT® is a comprehensive clinical terminology consisting over 366,000 concepts with unique meanings and formal logic-based definitions organized into hierarchies covering a broad range of human pathologies and anatomies and the relationships between them. We chose to use SNOMED CT® due to its depth of biological scale and comprehensiveness in human pathologies in general and specifically in psychiatric disorders [11, 12].

The current method employed five general steps (described further below): 1) conceptualization of the general query model, that defines the traversable paths (hierarchical relationships and semantic switches) used in mapping relationships between terms contained in each database 2) mapping of database terms to SNOMED via NLP and coding 3) mapping rules of relatedness (according to the general query model) and 4) query construction and implementation and 5) evaluation. Mapping of database terms to SNOMED was conducted using PhenOS, a knowledge-based phenotype organizer system [13], which was also used in assigning phenotypic context to Gene Ontology Annotations [14]. The architecture is outlined in Figure 1.

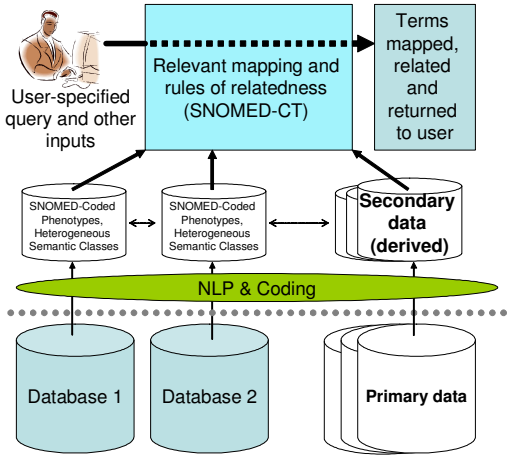

Figure 1. Overall scheme for heterogeneous database integration. Natural Language Processing & Coding (PhenOS) was first used to assign terms (and their corresponding SNOMED codes) to underlying data (Primary data) for each of the participating databases. These were organized into tables (Secondary data) whose fields were then related and mapped using ancestor-descendant and translation tables generated from SNOMED-CT (Data mapping).

1) Query Model. For simplicity we focused on three main classes within the SNOMED ontology: Anatomy (i.e. cingulate gyrus, hypothalamus), Abnormal Morphology (i.e. neoplasia, inflammation) and Disease (i.e. Alzheimer's, encephalitis), abbreviated by **A**, **M** and **D,** respectively. Formally these classes are descendants of three nodes of the SNOMED ontology: *brain tissue structure*, *diseases of brain* and *morphologically abnormal structure*. Diseases (**D**) can be related to Anatomies (**A**) through the linkage concept "has finding site", and Diseases (**D**) can be related to Abnormal Morphology (**M**) through "has associated morphology". The general query model is depicted in Figure 2.

The query model is flexible and general enough to allow for many different types of loosely defined queries. In essence, all queries possible within the model are delineated by traversing the edges on the 'x-y plane', and databases to be included are chosen along the 'z-axis'. Up and down arrows connect more broad and more specific concepts within a class through 'is a' (or 'part of' for anatomy) parent-child relationships. Horizontal arrows represent possible semantic switches and connect the three different classes with each other (D connected to A through 'has finding site', D connected to M through 'has associated morphology') and these can be traversed in both left and right directions. Table 1 (*supplement*) depicts all possible query types along the 'x-axis' and their potential utility.

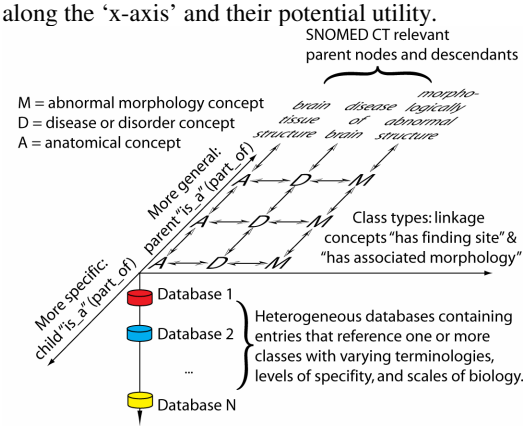

Figure 2. General Query Model. The SNOMED ontology extends along the 'y-axis'; parent nodes are 'most positive'. The relatable semantic classes extend along the 'x-axis'; Anatomies (A) can be related to Diseases (D), which can be related to Abnormal Morphologies (M). Participating databases extend down along the 'z-axis'. Each axis can be extended further; extension down the 'y-axis' is accomplished as more specific terms are added to SNOMED with upcoming revisions, relatable semantic classes could be added along the 'x-*axis'*

(i.e. Disease can also be related to class 'Organism' through linkage concept "causative agent"), and more databases can be added along the 'z-axis'.

2) NLP and Coding. For each database a table was created (via PhenOS) which consisted of database terms linked to a SNOMED ID code and their accession numbers to underlying data ('secondary data' in Figure 1). This was done for Brain, Neuronames, fMRIdc and GEO. (Note: for 'Brain', which consisting mostly of brain disease terms, no accession numbers were included. Example entries from two tables are given in Table 2 (*supplement*).

3) Mapping rules of relatedness. An ancestor-descendant table (Table 3 - *supplement*) was generated that included all SNOMED concepts under three nodes: *brain tissue structure*, *diseases of brain* and *morphologically abnormal structure* and the distances between them. A translation table (Table 4 - *supplement*) was also generated in which each disease under the node *disease of brain* was mapped to its Finding Site (Anatomy) and/or Associated Morphology (Morphology). In addition, a table (Table 5 - *supplement*) mapping all SNOMED IDs to their descriptions was generated (to be used in carrying out class-based queries.)

4) Query implementation. All of the above tables were imported into Microsoft Access 2003 and were used to recreate seven queries, or navigation paths, possible within the framework outlined by the general query model (Figure 1). Two general types of queries are described: 1) 'pair-wise mapping query', whereby all terms (and accession numbers to underlying data) between two databases that meet the criteria for the specified relationship type are returned and 2) 'class-based query' whereby a user can input a term (either an anatomical, disease or morphology concept), specify the relationship (type of mapping) and retrieve terms that fit the specified mapping from one or more selected databases. An example 'pair-wise mapping query' is depicted in Figure 3A (*supplement*), and answers the query 'Find Anatomy and Abnormal Morphology terms in fMRIDC that are associated with diseases and/or their subtypes that are included in Brain' (*'fMRIDdc to Brain A,M→D↓'*). This was done for each permutation of possible pair-wise mappings between all participating databases, and for seven types of semantic relationships. The numbers of unique pair-wise mappings generated between each database and for seven types of relationships (total 5,497) were used to populate Table 6 (*supplement*), the main point of which is to show the increase in relatedness between databases as more types of relationships are mapped.

The major utility of such a system is in 'class-based queries'. A schematic example of the class-based query "List all diseases with Finding Site 'temporal lobe' and then find references to these diseases (identical or subsuming) in all participating databases", with its navigation path traced over the General Query Model, is shown Figure 4. Figure 5 depicts in more detail the navigation path through SNOMED, used in returning a result for this query. The MS Access query setup for this query is given in Figure 3B with results 3C (*supplement*). In future implementations of the system, class-based queries would be generated for each type of specified relationship on a web interface.

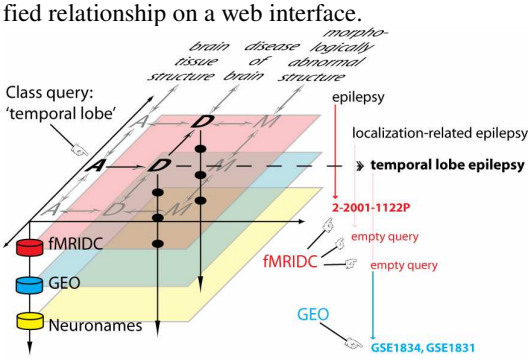

Figure 4. Graphical depiction of the class-based query: "List all diseases with Finding Site 'temporal lobe' and then find references to these diseases (identical or subsuming) in all participating databases." In this example, 'temporal lobe epilepsy' is directly referenced in GEO, but must be expanded to subsuming ancestor term 'epilepsy' to find the closet match in fMRIDC.

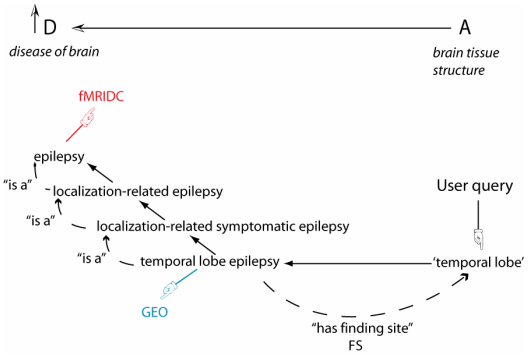

Figure 5. 'Close-up' depiction of semantic navigation path through the SNOMED ontology in answering the class-based query "List all diseases with Finding Site 'temporal lobe' and then find references to these disease (identical or subsuming) in all participating databases." Solid arrows are query navigation path, and dashed arrows are SNOMED directed relationships ("has finding site" and "is a"). "temporal lobe epilepsy" is found to be referenced in GEO, whereas only the more general term "epilepsy" was found in fMRIDC.

5) Evaluation. The evaluation was conducted on a set of 100 randomly chosen mappings (25 from

each datasource), as well as on 50 randomly selected mappings (Table 7-*supplement*) from step 1 of the approach (NLP & PhenOS). Precision was measured as the number of true mappings divided by the total number sampled, TP/(TP+FP). 95% confidence intervals (CI) were also calculated using the binomial formula ($p \pm Zc\sqrt{p(1-p)/n}$).

Results

5,497 unique pair-wise mappings were generated for seven types of relationships between each of the datasets: 1) **Identity -** terms are identical or similar between one dataset and another 2) **Subsuming** – terms in the one dataset subsume terms in the second 3) **Subsumed** – terms in one dataset are subsumed by terms in the second 4) **A,M→D↑** - terms in one dataset are either an Anatomical Structure or Abnormal Morphology and terms in the second dataset are Diseases that subsume diseases that have as finding site or associated morphology the term in the first dataset 5) **A,M→D↓** - terms in one dataset are either an Anatomical Structure or Abnormal Morphology and terms in the second dataset are Diseases that are subsumed by diseases that have as finding site or associated morphology the term in the first dataset 6) **D→A,M↑** - terms in one dataset are Diseases and terms in the second dataset are either an Anatomical Structure or Abnormal Morphology that subsume finding sites or associated morphologies of terms in the first dataset 7) **D→A,M↓** - terms in one dataset are Diseases and terms in the second dataset are either an Anatomical Structure or Abnormal Morphology that are subsumed by finding sites or associated morphologies of terms in the first dataset. Table 6 (*supplement*) shows the number of mappings for each relationship between each pair of datasets.

Based on 100 randomly selected mappings from Table 6 (25 to each datasource), the precision of the method was 98±2.7%. Based on 50 (12-13 from each datasource) randomly selected mappings from tables generated through NLP and PhenOS, precision for stage 1 of the method was 88±9%. Table 8 (*supplement*) shows reasons for common errors (homonymy, correct relations) and examples.

In a sample class query the term "mass" was used to retrieve all subsumed terms and underlying accession numbers from the GEO dataset. Using the symbols from above, this query can be written as: "mass"→ *M*↓ to GDS. This query resulted in 28 unique term and accession number pairs from the GEO dataset (Table 9).

GEO term	GEO accession
leukemia	GDS 461
glioma	GDS 493
astrocytoma	GDS 506
cancer	GDS 512
medulloblastoma	GDS 526

Table 9. Five example results (of 28) from the general class query: "mass"→ M↓ to GDS. This query retrieved all GDS terms and underlying accession numbers subsumed by the term "mass".

Discussion

Seamless integration of complex data types (i.e. imaging, microarrays) is the goal of many brain information resources and databases [15]. However, the technical, theoretical and computational challenges of imaging informatics currently prevent this and will do so for quite a while [16]. Meanwhile, there are efforts to standardize neuroscience data and meta-data models so that heterogeneous data can be joined across many disparate participating databases. An alternative approach has been proposed that bypasses the need for compatible data models and maps metadata between disparate participating databases on a semantic level. An additional advantage of the approach is that it utilizes the comprehensive knowledge encapsulated in the SNOMED ontology to enable queries that heretofore had no method for being answered.

More studies are emerging that attempt to find and interpret correlations between biomarkers (i.e. alleles), imaging, and neuropsychological markers with disease [17]. Ideally, these studies could be extended with questions such as: 1) where in the brain are biomarker-related genes expressed 2) what other genes are coexpressed with these genes and how do they vary by brain region 3) are these genes differentially expressed in tissues undergoing a pathological process (i.e. abnormal morphology such as inflammation or neuronal degeneration) related to the disease and 4) how do the above observations compare across related disorders? To address these questions the proposed approach could be used to quickly survey and retrieve relevant data from online databases. Furthermore, as meta-analysis of microarray and neuroimaging data become more feasible [18], this approach could help organize and retrieve such data in order to facilitate comparisons across tissues and according to the diseases and abnormal morphologies (pathological processes) that affect them in order to identify novel relationships that may elucidate the genesis of psychiatric diseases and disorders.

In addition to the inherent limitations of mapping only on the semantic level, the approach is also limited by mismapping due to the inherent risks in

NLP and text mining. This is further amplified by potential mismapping of the knowledge source (SNOMED) as we explore many more relationships than usual in a DAG. In future studies, we plan to use the BiomedLEE NLP [19] and a more formal schema for representing NLP-derived results [20] that has higher accuracy than text-mining.

Conclusion

The current work presents a novel method for query implementation that first provides structure over unstructured metadata of fMRI and gene expression datasets through NLP and coding, and then makes use of the modeling in SNOMED to decompose semantic information allowing for mapping between anatomies or morphologies related to disease. This allows for the integration of heterogeneous data with different biological scales, such as arrays and imaging, because the decomposition of a diagnosis or disease to its cell type, anatomical and/or morphological component allows for the spanning of more biological scales than the diagnosis would alone. To our knowledge, this is the first comprehensive implementation of the model of SNOMED's diseases that exploit their semantic decomposition in their otherwise implicit subphenotypes (histological, anatomical, morphological) that can further be mapped to the histological/morphological/anatomical metadata found in other scales in datasets such as microarrays.

Acknowledgments

We acknowlegde John D. Van Horn for valuable input, and the following grants: the NIH/NLM 1K22LM008308 (Semantic Approaches to Phenotypic Database Analysis), and the NIH/NCI 1U54CA121852-01A1 (National Center for the Multiscale Analysis of Genomic and Cellular Networks (MAGNet).

Address for correspondence

Yves Lussier
The University of Chicago
AMB N660B, (MC 6091)
5841 South Maryland Avenue
Chicago, IL 60637
Email: Lussier@uchicago.edu

References

1. Brinkley JF, Rosse C. Imaging and the Human Brain Project: a review. Methods Inf Med. 2002;41(4):245-60.
2. Kotter R. Neuroscience databases: tools for exploring brain structure-function relationships. Philos Trans R Soc Lond B Biol Sci. 2001 Aug 29;356(1412):1111-20.
3. Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. J Biomed Inform. 2005 Apr;38(2):114-29.
4. Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. Pac Symp Biocomput. 2003:624-35.
5. Aronson AR. The effect of textual variation on concept based information retrieval. Proc AMIA Annu Fall Symp. 1996:373-7.
6. Zeng Q, Cimino JJ. Mapping medical vocabularies to the Unified Medical Language System. Proc AMIA Annu Fall Symp. 1996:105-9.
7. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp. 1998:815-9.
8. Cantor MN, Sarkar IN, Gelman R, Hartel F, Bodenreider O, Lussier YA. An evaluation of hybrid methods for matching biomedical terminologies: mapping the gene ontology to the UMLS. Stud Health Technol Inform. 2003;95:62-7.
9. Shah NH RD, Supekar KS, Musen MA,, editor. Ontology-based Annotation and Query of Tissue Microarray Data. AMIA; 2006.
10. Mork P, Halevy A, Tarczy-Hornoch P. A model for data integration systems of biomedical data applied to online genetic databases. Proc AMIA Symp. 2001:473-7.
11. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc. 2003:699-703.
12. Jenders RA. Classification of psychiatric disorders. Jama. 2005 Oct 19;294(15):1899; author reply -900.
13. Lussier YA, Li J. Terminological mapping for high throughput comparative biology of phenotypes. Pac Symp Biocomput. 2004:202-13.
14. Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C. Phenogo: assigning phenotypic context to gene ontology annotations with natural language processing. Pac Symp Biocomput. 2006:64-75.
15. Bowden DM, Dubach MF. NeuroNames 2002. Neuroinformatics. 2003;1(1):43-59.
16. Neerincx PB, Leunissen JA. Evolution of web services in bioinformatics. Brief Bioinform. 2005 Jun;6(2):178-88.
17. Tan HY, Nicodemus KK, Chen Q, Li Z, Brooke JK, Honea R, et al. Genetic variation in AKT1 is linked to dopamine-associated prefrontal cortical structure and function in humans. J Clin Invest. 2008 Jun;118(6):2200-8.
18. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. Nat Biotechnol. 2006 Jan;24(1):55-62.
19. Lussier YA, C. Friedman. BiomedLEE: a natural-language processor for extracting and representing phenotypes, underlying molecular mechanisms and their relationships. ISMB. 2007;(in press).
20. Friedman C, Borlawsky T, Shagina L, Xing HR, Lussier YA. Bio-Ontology and text: bridging the modeling gap. Bioinformatics. 2006 Oct 1;22(19):2421-9.