

Ontology-driven Indexing of Public Datasets for Translational Bioinformatics

Nigam H. Shah, M.B.B.S., PhD, Annie P. Chiang, PhD, Atul J. Butte, MD, PhD,
Rong Chen, PhD and Mark A. Musen, MD, PhD
Stanford Medical Informatics, School of Medicine, Stanford University, Stanford,
CA 94305, USA

Abstract

The volume of publicly available genomic scale data is increasing. Genomic datasets in public repositories are annotated with free-text fields describing the pathological state of the studied sample. These annotations are not mapped to concepts in any ontology, making it difficult to integrate these datasets across repositories. We have previously developed methods to map text-annotations of tissue microarrays to concepts in the NCI thesaurus and SNOMED-CT¹.

In this work we generalize our methods to map text annotations of gene expression datasets to concepts in the UMLS. We demonstrate the utility of our methods by processing annotations of datasets in the Gene Expression Omnibus. We demonstrate that we enable ontology-based querying and integration of tissue and gene expression microarray data. We enable identification of datasets on specific diseases across both repositories. Our approach provides the basis for ontology-driven data integration for translational research on gene and protein expression data.

Introduction and Background

The amount and diversity of genomic scale data has been steadily increasing for the past several years. This increase has enabled integrative translational bioinformatics studies across these datasets². Currently, the predominant genomic level data is gene expression microarrays. Recently, other forms of genomic scale measurements have been gaining acceptance, one of them being Tissue Microarrays. Tissue Microarrays allow for the immunohistochemical analysis of large numbers of tissue samples and are used for confirmation of microarray gene-expression results as well as for predictive pathology³. A single tissue microarray (TMA) paraffin block can contain as many as 500 different tumors, enabling the screening of thousands of tumor samples for protein expression using a few array sections⁴. Databases such as the Stanford Tissue Microarray Database (TMAD)⁵ provide a central repository for data from TMA's akin to the

Stanford Microarray Database (SMD) and Gene Expression Omnibus (GEO) for gene expression arrays.

It has been noted before, that little attention is being paid to the problem of developing mechanisms to integrate the results from these – gene expression microarrays and tissue microarrays – complementary data types^{4,6}. Several reviews have suggested that it is essential to address this issue and synchronize the analysis, interpretation and data standards for these data^{4,6}.

In order to develop integrative translational bioinformatics approaches to interpret these datasets, there is a strong and pressing need to be able to identify all experiments that study a particular disease. A key query dimension for such integrative studies is the *sample*, along with a gene or protein name. As a result, besides queries that identify all genes that have a function X – which can be reliably answered using the Gene Ontology (GO) – we need to conduct queries that find all samples/experiments that study a particular disease and/or the effect of an experimental agent. However, because of the lack of a commonly used ontology or vocabulary to describe the diagnosis, disease studied or experimental agent applied in a given experimental dataset it is not possible to perform such a query.

The challenge is to create consistent terminology labels for each experimental dataset in the public repositories that would allow the identification of all samples that are of the same type at a given level of granularity. (e.g., *All carcinoma* samples versus *all Adenocarcinoma in situ of prostate* samples, where the former is at a coarser level of detail). One mechanism of achieving this objective is to map the text-annotations describing the diagnoses, pathological state and experimental agents applied to a particular sample to ontology terms allowing us to formulate refined or coarse search criteria^{7,8}.

Butte et al have previously applied a text-parser (GenoText) to determine the phenotypic and experimental context from text annotations of GEO experiments². They report that text-parsing is still an inefficient method to extract value from these

annotations⁹. In later work, Butte et al explored the use of PUBMED identifiers of the publication associated with GEO experiments and their assigned Medical Subject Headings (MeSH) to identify disease related experiments⁹. They were able to relate 35% of PUBMED associated GEO series to human diseases. Only half or so of the GEO experiments have PUBMED identifiers and the remaining are inaccessible to this approach, possibly necessitating alternative methods⁹.

We have previously developed methods to process such text-annotations for tissue microarrays and map them to concepts in the NCI thesaurus and the SNOMED-CT ontologies^{1, 10}. In the current work we generalize our methods to process text annotations for gene expression datasets in GEO (as well as TMAD) and map them to concepts in the UMLS. We present results on the accuracy of our mapping effort and demonstrate how the mapping enables better query and integration of gene expression and protein expression data. We discuss the utility of our approach to derive integrative analyses.

Methods

Overview of annotations in GEO

The Gene Expression Omnibus (GEO) is an international repository of microarray data run by the National Center for Biotechnology Information (NCBI). In this work we use the November 2006 release of GEO, which contained 108371 samples, 4593 GEO experiment series and 1080 GEO datasets, 369 of which are human. In this analysis we focus only on the human datasets. Each GEO dataset has a title and a description field that contain text entered by the person uploading the dataset. Moreover, GEO datasets can have an additional 24 descriptors (such as agent, cell line, and species) along with their subset descriptions. In the current work we process the text from the title, description and agent descriptors of GEO datasets.

Overview of text annotations in TMAD

The Stanford Tissue Microarray Database (TMAD) contains data from immunohistochemical analyses performed with tissue microarrays. The TMAD provides tools for quick upload, storage and retrieval of the tissue microarray images and the analysis of immunohistochemical staining results¹¹. Each sample in the TMAD contains free-text annotations – entered by the experimenter – for fields such as the organ system, and up to five diagnosis terms (one principal diagnosis field and four sub diagnosis fields) describing the sample. We concatenate the text from these six fields and use it as the annotation of the sample for our work. We refer to experiments

studying samples with the same diagnoses as one dataset. Currently, the TMAD has 10734 samples that can be grouped into 1045 datasets according to their diagnoses.

Indexing with Ontology terms

We downloaded UMLS 2006 AD and created a MySQL database using the Metamorphosys tool as described in the UMLS documentation¹².

In order to map existing annotations in TMAD and GEO to ontology terms, we used the UMLS-Query module developed by our group to process the existing descriptions of the samples and matching them to ontology terms. Fully describing the UMLS-Query module and all of its functionality is beyond the scope of this current work but we describe the key mapping function *mapToId* here. For each text-annotation, we read a sliding window of five words from the text. We generate all possible permutations (5-grams) of these words and look for an exact match to an ontology term. We examine all 5 word permutations because we observed that most disease and drug names are less than 5 words in length. This permutation based method, though accurate, can be made more efficient computationally. We are working on that collaboratively with the National Center for Integrative Biomedical Informatics¹³. We restrict the matches to SNOMEDCT and the NCI thesaurus vocabularies when identifying disease names. The UMLS-Query module along with detailed documentation is available from www.stanford.edu/~nigam/UMLS. We do not employ any natural language processing strategies such as stemming, normalization or noun-phrase recognition. We also do not employ any heuristics or hacks for increasing match accuracy.

The result of the mapping is a table that associates each GEO or TMAD dataset identifier to one or more concepts in the UMLS. We query this table to identify disease related datasets as well as identify matching datasets from the repositories.

Results

We processed the annotations corresponding to the annotations of 369 GEO datasets (GDS) and 1045 TMAD datasets. We then evaluated the ability of our ontology-based indexing scheme to enable the identification of experiments for the following use cases: 1) Accurately identify experiments related to particular diseases 2) Identify gene and protein expression datasets corresponding to diseases from both GEO and TMAD. These use cases were defined based on current research in translational bioinformatics and prior reviews indicating the need for such integrative analyses^{2-4, 9}.

Identifying Disease related experiments in GEO

Out of 369 GDS, we identify 241 disease related experiments. The 241 disease related experiments can be grouped into categories according to the semantic type of the concepts assigned to them. We exclude high level ontology terms such as *cancer*, *syndrome*, *exposure*, *damage*, *toxicity* before performing the grouping. The number of experiments identified as disease related drops to 209 on removing high level terms. Such terms, though accurately mapped and being of the right semantic type, are too high level to enable the correct identification of disease related datasets.

Semantic type	Number of GDS
Neoplastic Process	109
Disease or Syndrome	97
Injury or Poisoning	8
Mental or Behavioral Dysfunction	3

Table 1. Categorization of GEO datasets according to the Semantic type after excluding matches to high level ontology terms.

Table 3 shows examples of the highest and lowest numbers of GEO datasets for the *Neoplastic Process* and *Disease or Syndrome* categories. (We omit rows that subsume other rows, for example leukemia subsumes acute leukemia, which in turn subsumes acute myeloid leukemia)

GDS	Concept name	CUI	Semantic type
Examples of cancers with many GEO datasets			
26	Breast cancer	C0006142	Neoplastic Process
11	Acute myeloid leukemia	C0023467	Neoplastic Process
5	Acute lymphoblastic leukemia	C0023449	Neoplastic Process
Examples of cancers with few GEO datasets			
1	Kaposi's sarcoma	C0036220	Neoplastic Process
1	Acute promyelocytic leukemia	C0023487	Neoplastic Process
1	Pleural mesothelioma	C1377913	Neoplastic Process
Examples of diseases with many GEO datasets			
13	Duchenne dystrophy	C0013264	Disease or Syndrome
6	Arthritis	C0003864	Disease or Syndrome
4	Chronic obstructive	C0024117	Disease or Syndrome

	pulmonary disease		
Examples of diseases with few GEO datasets			
1	Open-angle glaucoma	C0017612	Disease or Syndrome
1	Purpura thrombocytopenic	C0857305	Disease or Syndrome
1	Corneal dystrophy	C0010035	Disease or Syndrome

Table 2. Overview of the number of GEO datasets for concepts in the Neoplastic Process and Disease or Syndrome category.

Identifying Disease related experiments in TMAD

We have previously presented results on processing annotations in TMAD. For the current work, we reprocessed the records in TMAD and performed the evaluation as described before¹. The average precision and recall was 85% and 95% respectively. The annotations of 1045 datasets mapped to 902 disease related concepts in the UMLS. We do not discuss this further because this has been described in our previous work.

Identifying matching GEO and tissue array datasets

From the 902 disease related datasets in TMAD and 241 disease related GDS that we identified in GEO, we were able to identify 45 disease related concepts for which there were datasets in both GEO and TMAD – and hence are potential candidates to support further analysis. Many of them are high level matches (such as Leukemia) that are accurate but too high level to enable correlative analyses.

From this set of 45 matches, we identified the 23 disease related concepts that were at an appropriate level of granularity and have multiple samples in both GEO and TMAD to enable further integrative study. In the table below we show the number of datasets, the number of GEO samples (GSM) and the corresponding number of tissue microarray samples for these.

Disease	GEO datasets	GEO samples	TMAD samples
Acute myeloid leukemia	11	366	3
Malignant melanoma	3	47	43
B-cell lymphoma	3	133	27
Prostate cancer	3	47	15
Renal carcinoma	2	34	185
Carcinoma squamous	2	105	175
Multiple myeloma	2	225	169
Clear cell carcinoma	2	34	63
Renal cell carcinoma	2	34	9
Breast carcinoma	2	3	1277

Hepatocellular carcinoma	1	80	163
Carcinoma lung	1	91	66
Cutaneous malignant melanoma	1	38	41
T-cell lymphoma	1	29	31
Lymphoblastic lymphoma	1	29	30
Uterine fibroid	1	10	19
Medulloblastoma	1	46	9
Clear cell sarcoma	1	35	8
Leiomyosarcoma	1	24	5
Mesothelioma	1	54	5
Kaposi's sarcoma	1	4	3
Cardiomyopathy	1	14	2
Dilated cardiomyopathy	1	14	2

Table 3. Diseases for which there are both gene expression and tissue microarray datasets.

Evaluation

In order to calculate the precision and recall for the task of identifying disease related experiments in GEO, we examined the 241 (209 after removing high level terms) GDS identified as disease related, to determine which of them were correctly identified. One of the authors (AC), went through each matched record and scored them for being a true positive of being a false positive. For calculating recall, we examined the unmatched GDSs (from the total of 369 GDSs) and searched the UMLS manually to identify concept to which the GDSs should have matched; this allowed us to compute the false negative rate. (Note that such exhaustive evaluation is not possible with the current size of GEO, which has over 2000 GDSs). The results of this evaluation are presented in the following table:

Accuracy in identifying disease related datasets			
	Correct	Incorrect	Total
Positive	202 (TP)	39 (FP)	241
Negative	97 (TN)	31 (FN)	128
Precision = 83.8%		Recall = 86.6%	
Accuracy in identifying disease related datasets after limiting high level matches			
	Correct	Incorrect	Total
Positive	188 (TP)	21 (FP)	209
Negative	115 (TN)	45 (FN)	160
Precision = 89.9%		Recall = 80.6%	

Table 4. Accuracy of identifying disease related datasets.

Next, we evaluate the ability to accurately match up the right tissue array datasets with gene expression datasets. Out of the 45 candidate datasets proposed as corresponding between GEO and TMAD, on manual inspection all of them were accurate matches, though

12 were high level terms such as *Cancer, Syndrome, and Sarcoma*. We consider these as false positive because such matched are uninformative for the purpose of matching up disease related datasets across repositories. This gives us a precision of 73%. We were unable to perform a recall analysis for this task because it is extremely time consuming to manually examine all the TMAD and GEO datasets to determine the number of matches that were missed by our method. One notable disease for which datasets exist in both but were not accurately identified is Breast carcinoma. Most datasets in GEO for this disease are labeled with the term Breast Cancer (C0006142) and those in TMAD are labeled with the term Breast carcinoma (C0678222). These terms have different CUIs in the UMLS and hence we were unable to match up these datasets.

Discussion and future work

In the current work we have processed annotations of tissue microarray samples in TMAD as well as annotations of gene expression datasets in GEO and mapped them to concepts in the UMLS. One of the insights in our work is that disease names are rarely longer than five words and it is computationally tractable to perform an exhaustive search for all possible five-word permutations of the text-annotation.

Mapping text-phrases to UMLS concepts has been performed by other researchers in the past^{14,15}. Most of these approaches are for the purpose of automatic indexing of biomedical literature, and have been shown to be inadequate for processing annotations of high-throughput datasets^{2,9}. It has also been shown that for the task of identifying concepts from annotations of high-throughput datasets, simple methods perform equal or better than Metamap^{2,9,10,13}

In the current work, we use very simple methods to process text-annotations and map them to ontologies to demonstrate that such automatic mapping enables integration of gene expression and tissue microarray datasets.

Currently the only way to query the processed annotations is via SQL queries. Performing complicated SQL queries is not always possible for all end-users and the ontology hierarchies and the mapped annotations should drive specialized query-interfaces¹. One possible approach is to process the user's query text using the same indexing method that mapped the annotations of the datasets and retrieve those datasets that have the largest intersection with the concepts identified in the processed query. In future we will develop interfaces

to search these processed annotations as part of our work at the National Center for Biomedical Ontology, to create resources and methods to (help biomedical investigators) store, view, and compare annotations of biomedical research data.

As we all know, genomic scale data is increasing in volume and diversity. Currently, little attention is being paid to the problem of developing methods to integrate complementary data types such as gene expression microarrays and tissue microarrays based on their annotations^{4,6}.

We believe it is possible to perform ontology based indexing of free-text annotations of high throughput datasets in public databases to enable such integration. If the annotations of multiple databases – such as that of GEO, TMAD, SMD, PharmGKB – are processed in this manner, it will enable users to perform integrated analyses of multiple, high throughput, genomic scale datasets⁶.

Conclusions

We have demonstrated that we can effectively map the text-annotations of microarray datasets in GEO as well as annotations of tissue microarrays in TMAD to concepts from vocabularies in the UMLS. Our results show that we can map disease names and disease related concepts with high precision and recall.

We demonstrate how such mapping to concept hierarchies offers the ability to identify corresponding datasets across different repositories for integrative analyses. We identified 23 candidate datasets for further study.

We have implemented this mapping functionality as a PERL module named UMLS-Query, which is available with documentation at www.stanford.edu/~nigam/UMLS

Acknowledgements

We acknowledge Robert Marinelli and Matt van de Rijn for access to TMAD data. NHS and MAM acknowledge support from NIH grant U54 HG004028. APC, AJB and RC acknowledge support from the Lucile Packard Foundation for Children's Health and the National Library of Medicine (K22 LM008261).

References

- 1 Shah NH, Rubin DL, Supekar KS, Musen MA. Ontology-based Annotation and Query of Tissue Microarray Data. AMIA Annual Symposium; 2006; Washington, DC; 2006. p. 709-13.
- 2 Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. Nat Biotechnol. 2006 Jan;24(1):55-62.

- 3 Sauter G, Mirlacher M. Tissue microarrays for predictive molecular pathology. J Clin Pathol. 2002 Aug;55(8):575-6.
- 4 Rimm DL, Camp RL, Charette LA, Costa J, Olsen DA, Reiss M. Tissue microarray: a new technology for amplification of tissue resources. Cancer J. 2001 Jan-Feb;7(1):24-31.
- 5 Marinelli R.J., Montgomery K., Liu C. L., Shah N. H., Nitzberg M., Zachariah Z. K., et al. The Stanford Tissue Microarray Database Nucleic Acids Res. 2008;Database Issue(Accepted).
- 6 Basik M, Mousses S, Trent J. Integration of genomic technologies for accelerated cancer drug development. Biotechniques. 2003 Sep;35(3):580-2, 4, 6 passim.
- 7 Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform. 2005 Sep;6(3):239-51.
- 8 Moskovitch R, Martins SB, Behiri E, Weiss A, Shahar Y. A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. J Am Med Inform Assoc. 2007 March-April;14(2):164-74.
- 9 Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. AMIA Annu Symp Proc. 2006:106-10.
- 10 Shah NH, Rubin DL, Espinosa I, Montgomery K, Musen MA. Annotation and query of tissue microarray data using the NCI Thesaurus. BMC Bioinformatics. 2007;8:296.
- 11 Liu CL, Prapong W, Natkunam Y, Alizadeh A, Montgomery K, Gilks CB, et al. Software tools for high-throughput analysis and archiving of immunohistochemistry staining data obtained with tissue microarrays. Am J Pathol. 2002 Nov;161(5):1557-65.
- 12 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.
- 13 Dai M, H. SN, Xuan W, Musen MA, Watson SJ, Athey B, et al. An Efficient Solution for Mapping Free Text to Ontology Terms. AMIA Summit on Translational Bioinformatics (accepted poster); San Francisco, CA; 2008.
- 14 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
- 15 Hersh W, Leone TJ. The SAPHIRE server: a new algorithm and implementation. Proc Annu Symp Comput Appl Med Care. 1995:858-62.