

Ontology-based Annotation and Query of Tissue Microarray Data

Nigam H. Shah, Daniel L. Rubin, Kaustubh S. Supekar and Mark A. Musen
Stanford Medical Informatics, Stanford University School of Medicine and the National
Center for Biomedical Ontology, Stanford University, Stanford, CA 94305, USA.

*Corresponding Author: nigam@stanford.edu Ph: 650-725-6236.

Abstract

The Stanford Tissue Microarray Database (TMAD) is a repository of data amassed by a consortium of pathologists and biomedical researchers. The TMAD data are annotated with multiple free-text fields, specifying the pathological diagnoses for each tissue sample. These annotations are spread out over multiple text fields and are not structured according to any ontology, making it difficult to integrate this resource with other biological and clinical data. We developed methods to map these annotations to the NCI thesaurus and the SNOMED-CT ontologies. Using these two ontologies we can effectively represent about 80% of the annotations in a structured manner. This mapping offers the ability to perform ontology driven querying of the TMAD data. We also found that 40% of annotations can be mapped to terms from both ontologies, providing the potential to align the two ontologies based on experimental data. Our approach provides the basis for a data-driven ontology alignment by mapping annotations of experimental data.

Introduction and Background

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 pathology^[1]. 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 sections^[2]. Commercial digital-imaging systems can rapidly store thousands of images resulting from such sections. The Stanford Tissue Microarray Database (TMAD) provides a central repository for data from TMA's akin to the Stanford Microarray Database (SMD) for gene expression arrays.

Superficially the datasets generated from TMA and gene expression arrays appear similar in that both are matrix type data and each entry in the matrix provides information about the expression of a biological entity (gene or protein) in a particular sample. However, gene expression arrays query a large number of genes in *one* sample or patient,

whereas Tissue microarrays query a large number of samples/patients for *one* protein. The key query dimension in TMA data is a tissue sample, rather than a gene. As a result, queries such as 'find all genes that have a function X' get morphed to a query such as 'find all tissue samples that have a particular diagnosis'. However, because of the lack of a commonly used ontology to describe the diagnosis for a given TMA sample in TMAD – analogous to the Gene Ontology for the function of gene products – it is not possible to perform such a query.

The key challenge is to create consistent terminology labels for each sample/record in the TMAD 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 diagnosis of a particular sample to ontology terms that allow us to formulate refined or coarse search criteria^[3].

In the current work, we describe methods to map the text annotations for records in the TMAD to the NCI thesaurus and SNOMED-CT ontologies, present the results of the mapping effort and discuss how the mapping enables better querying of the data in TMAD. Since ontologies may overlap in content, terms in different ontologies can map to the same data, and thus enable ontology mapping. We demonstrate the potential utility of our approach to derive data-driven mapping between ontologies.

Methods

Overview of data in TMAD

The Tissue Microarray Database (TMAD) contains data from immunohistochemical analysis of a large number of tissue samples that were studied with tissue microarrays. The TMAD provides tools for quick upload, storage and retrieval of the TMA images and the analysis of immunohistochemical staining results^[4]. Each sample record in the TMAD contains free-text annotations – entered by the experimenter – for several fields such as the organ system, the source of the sample, the antibody used for staining, and the staining result. Among these

fields are up to five diagnosis terms (one principal diagnosis field and four sub diagnosis fields) describing the sample as well as a label for the organ and organ system from which the sample is derived. There are separate tables in the database for keeping track of user logins, experiment details, array constructions etc which are not relevant to the current work.

Ontologies for Annotating TMA Data

The NCI Thesaurus is an ontology providing broad coverage of the cancer domain, including cancer-related diseases, findings and abnormalities [5]. In certain areas, such as cancer diseases and combination chemotherapies, the NCI Thesaurus provides the most granular and consistent terminology available. The Thesaurus currently contains over 34,000 concepts, structured into 20 taxonomic trees. It is published under an open content license. The NCI thesaurus can be obtained at <ftp://ftp1.nci.nih.gov/pub/cacore/EVS/>.

We downloaded version Thesaurus-05.09g in the tab delimited text format. This text file contains columns for an id, name, parents, synonyms and definition for each NCI thesaurus term. The parents and synonyms columns contain the immediate parents of the term and its synonyms separated by '|' respectively.

However, this format is not optimal for searching parent-child terms rapidly as well as for the purpose of pattern matching against individual synonyms. Therefore we reorganized the thesaurus into three separate tables which we named nci_term, nci_children and nci_synonyms. The layout of these tables and the relationships among them are show in Figure 1.

The nci_term table contains the Thesaurus-05.09g as is. The nci_children table contains one row

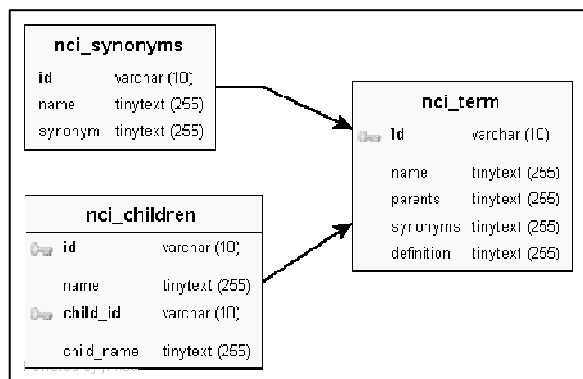


Figure 1. The figure shows the re-organized layout for the NCI thesaurus. The nci_term table contains all the fields from the Thesaurus-05.09g file. The nci_children table contains one row for each direct parent-child relationship. The nci_synonyms table contains one row each for each synonym of a term.

for each direct parent-child relationship and is indexed on both the parent and child ids. The nci_synonyms table contains one row each for each synonym of a term and is indexed on the synonym field. This reorganization allows us to rapidly search the synonyms of a term as well as identify the parents and child terms of a given term. The tab-delimited files corresponding to these tables are available from the authors.

We downloaded UMLS 2005 AA and created a MySQL database containing the SNOMED-CT ontology using the Metamorphosys tool as described in the UMLS documentation [6].

MAPPING TMA Data to Ontologies

In order to map existing annotations of samples in the TMA database to ontology terms, we created a database containing the TMAD data, the NCI thesaurus and the SNOMED-CT (derived from UMLS). We used Perl scripts to process the existing descriptions of tissue samples and to generate strings for matching with ontology terms. We identified several heuristics to increase the accuracy of our matches. The Perl scripts used for matching are available from the authors.

Generation of annotation permutations

Each record in the TMA database has an organ, a diagnosis and four sub diagnosis terms associated with it. For example, a record might contain the entries *breast, carcinoma, ductal, <null>, in situ, <null>* for the organ (O), diagnosis (d0), sub diagnosis 1 (d1), sub diagnosis 2 (d2), sub diagnosis 3 (d3) and sub diagnosis 4 (d4) fields respectively. We refer to the diagnosis related entries (d0, d1, d2, d3, d4) as a *term-set*.

We generated all possible permutations of the non-null entries in every term-set. In theory, we would generate over a million permutations for the 8519 records contained in the TMA database. However, these 8519 records correspond to 783 unique term-sets, many of which contain fewer than five entries, and in practice we end up with about fifteen thousand permutations. We used the pair of a particular term-set permutation and the associated organ in the search for an ontology term to associate with that TMA record

Heuristics for increasing match accuracy

When searching the NCI thesaurus or SNOMED-CT, instead of searching for an ontology term for each permutation-organ pair, we first filtered out the non-informative permutations using heuristics that identify such permutations. For example, there are a number of uninformative records where *d0 = normal and d1=19w* which we identify using regular expressions. We also “tweak” a permutation to

convert it to the most useful form. For example, if the first word in a permutation is *carcinoma* and the second word is *adeno* then flip the order and merge the two words to make the first word to be *adenocarcinoma*. In another example, if the first word in a permutation is *carcinoma* and the second word is *squamous* then flip the order but keep the words separate.

We use such a processed permutation and search for an exact match with a term or a synonym in the NCI thesaurus and SNOMED-CT. If there is no match, then we drop the right most term (similar to right truncation in the UMLSKSS search) and repeat the search. When, during this process, the permutation contains two words or fewer, we add ‘*of <organ>*’ to the search string where *<organ>* is obtained from TMAD.

We also use simple heuristics to weed out bad matches. For example, if a matched term contains *uterine* but the organ associated with the record is not ovary, uterus, or fallopian tube, then we discard the match. If the matched term contains *mouse*, then we discard the match.

Results

We mapped the term-sets corresponding to the 8495 records from the TMAD, which specified a diagnosis, to ontology terms in the NCI thesaurus and SNOMED-CT. Out of the 783 distinct term-sets (for the 8495 samples) we were able to match 577 term-sets to the NCI thesaurus and 365 term-sets to SNOMED-CT, corresponding to 6614 and 3465 records respectively. In total we were able to map 6871 records (80%) of annotated records in TMAD (corresponding to 641 distinct term-sets) to one or more ontology terms.

Matches to the NCI Thesaurus

The NCI thesaurus had the widest coverage in providing matching terms for associating with diagnosis term-sets from the TMAD. Term-sets for a total of 6614 records could be mapped to one or more terms from the NCI thesaurus. The type (and the granularity level) of the ontology terms that matched a given term-set varied over a wide range. For example, there are records where the term-set contained just four characters, such as *MMMM (of ovary)* and matched a very specific ontology term such as ‘*Malignant_Mixed_Mesodermal_Mullerian_Tumor*’. At the same time, there are records where the term-set was highly descriptive such as *carcinoma adeno intraductal (of prostate)* which matched an ontology term such as ‘*Prostate_Ductal_Adenocarcinoma*’ and records where the term-set was highly descriptive such as *carcinoma transitional cell in situ (of*

bladder) which matched two ontology terms – ‘*Stage_0_Transitional_Cell_Carcinoma*’ and ‘*Bladder_Carcinoma*’ because no single term existed that would capture all the information.

Matches to SNOMEDCT

The SNOMED-CT provided matching terms for term-sets corresponding to 3465 records, with 3208 records also having a matching term from the NCI thesaurus. This implies that the term-sets from 257 records mapped *only* to SNOMED-CT terms. However, these 257 records correspond to just 70 distinct term-sets. Among these 70 are records where the term-set did not describe any cancer and consisted of words such as *intussusception (of colon)* which matched ‘*Intussusception of colon*’ and *endometrium proliferative (of uterus)* which matched ‘*Proliferative endometrium*’. It is not surprising that these term-sets did not match anything in the NCI thesaurus. There are a few records with very specific term-sets such as *oncocytoma (of kidney)* which matches ‘*Oncocytoma of kidney*’. The rest are records with rather descriptive term-sets, such as *carcinoma adeno gastrointestinal primary (of stomach)* and *carcinoma adeno into muscularis (of colon)* which match high level terms such as ‘*Carcinoma of stomach*’ and ‘*Carcinoma of colon*’ respectively. We do not consider such high-level matches to be correct hits in the evaluation described below.

The 3208 records that matched terms from *both* the NCI thesaurus and SNOMED-CT correspond to 295 distinct term-sets. These records are very interesting because they can provide potential ‘anchor’ points between the two ontologies, as will be described below.

Evaluation by random sampling

In our work, we mapped about 80% of the records in TMAD to one or more ontology terms, making it extremely time consuming for a domain expert to evaluate every matched term manually. Therefore, we devised a sampling strategy where we randomly selected 50 rows comprising a distinct term-set, the associated organ, and the matched ontology term to determine the percentage of matches that were appropriate (or inappropriate) on manual inspection. We do not consider high-level matches such as *carcinoma adeno into muscularis (of colon)* – which matches ‘*Carcinoma of colon*’ – to be appropriate hits in this evaluation (Adenocarcinoma of colon would be an appropriate match in this case).

We repeated this procedure 3 times for both the NCI thesaurus matches and the SNOMED-CT matches. The results of this exercise are presented in the following table:

	NCI		SNOMED-CT	
	Appropriate	Inappropriate	Appropriate	Inappropriate
Set-1	41	9	41	9
Set-2	42	8	43	7
Set-3	46	4	38	12
Total	129	21	122	28
Average (%)	43.0 (86%)	7.0 (14%)	40.66 (81%)	9.33 (19%)

For the 21 NCI Thesaurus matches that were deemed inappropriate, there were 8 term-sets where the SNOMED-CT term provided a better, appropriate match. Whereas, for the 28 SNOMED-CT matches deemed inappropriate, there was just 1 term-set where the NCI Thesaurus provided an appropriate match.

Partial alignment of the NCI thesaurus and SNOMED-CT ontologies

The term-sets that have matching ontology terms from two different ontologies are very interesting because they represent potential ‘anchor’ points that can be used to align the two ontologies. In the case of the NCI thesaurus and SNOMED-CT, the UMLS can be used to identify equivalent terms from the two ontologies and to map them to a common concept id and to declare a relationship that exists

between them. However, these alignments among source terminologies in UMLS are manually created, and not induced from data.

We examined the matched terms for the 3208 records that had matches from both the NCI Thesaurus and the SNOMED-CT. For each record, we retrieved and examined the UMLS Concept Unique Identifiers (CUIs) for each matched term from the NCI thesaurus (Nt_i) and from SNOMED-CT (St_j). For a given term-set, if there were more than one Nt_i and St_j then we computed the link-distance between all $Nt_i - St_j$ pairs and used the smallest link-distance. If the CUIs for a $Nt_i - St_j$ pair were identical, or within two links of each other, we considered the terms to be properly aligned.

The CUIs for Nt_i and St_j were identical for 2335 records; were at one link from each other for 403 records and were at two links from each other for 189 records. Overall, $Nt_i - St_j$ pairs from 2927 records (corresponding to 259 distinct term-sets) were appropriately aligned. The CUIs for the $Nt_i - St_j$ pairs for 281 records (corresponding to 36 distinct term-sets), were separated by more than two links.

Mapping to NCI and SNOMED-CT enables better querying/analysis of TMA data.

Even after the mapping is accomplished, the simple assignment of the ontology terms to tissue samples is not immediately useful to the end-user unless these ontology terms are used to drive

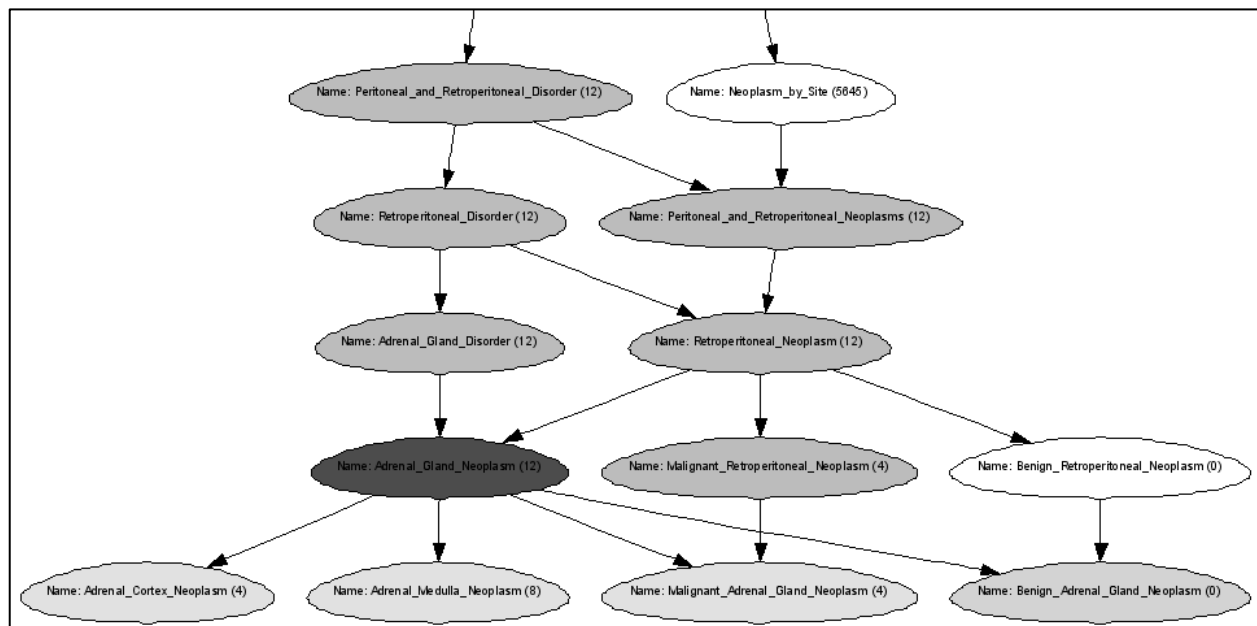


Figure 2. The figure shows a zoomed in region of the DAG view resulting from clicking on the term Adrenal gland neoplasm as described in the example in the main text. The red node is the term that has been clicked by the user, the yellow nodes are the child terms that have at least one sample in the TMA database assigned to that term, grey nodes are child terms with no corresponding samples in the TMAD and burlywood nodes are parent terms with less than 50 samples.

specialized query-interfaces. Plain keyword-based querying of ontology terms is not very useful. Therefore, we are developing a querying mechanism where the user starts with a term, visualizes the 'neighboring' ontology terms of that term in a DAG view and browses up, and down the ontology-term hierarchy to identify a term that is at the right level of granularity. This term then can be used to pull out all the TMAD records that are associated with that term or its child-terms. A prototype of such an interface is implemented at <http://smi-protege.stanford.edu/~nigam>.

Using this interface, a user can query a term such as "peritoneal neoplasm" to find that there are no samples corresponding to that particular term, but at the same time, the user can see that there are 12 samples corresponding to the parent term (peritoneal or retroperitoneal neoplasm), and can click on the retroperitoneal neoplasm term to find that these are adrenal gland neoplasms (Figure 2) and that four of them are from the medulla, and that eight are from the cortex of which four are malignant. The user can then choose to retrieve all 12 samples corresponding to retroperitoneal neoplasm or retrieve the samples corresponding to specific terms such as adrenal cortex neoplasm.

Conclusion

We have demonstrated that we can effectively map the diagnosis-related TMAD annotation terms to the NCI and SNOMED-CT ontologies. The NCI thesaurus terms have the widest coverage and can provide terms for about 77% of the matches. In our opinion the NCI thesaurus is the ontology of choice for annotating tissue microarray samples.

We described how such a mapping allows a rich querying facility and offers the ability to identify "similar" or "related" tissue microarray samples, even though they may be annotated with different terms. For example, the four neoplasms of the adrenal medulla and eight neoplasms of the adrenal cortex (four of which are malignant) are all related to each other by the fact that they are all retroperitoneal neoplasms.

During the process of mapping the diagnosis term-sets to the NCI and SNOMED-CT, we acquire information that can be used to align terms from the two ontologies. In case of the TMA database, 3208 records had their term-sets mapped to *both* the NCI thesaurus and SNOMED-CT terms. Analysis of these terms showed that NCI and SNOMED terms mapped to 2810 records, and these terms were appropriately aligned as evidenced by their identical (or very close) CUIs in the UMLS.

Discussion and future work

In the current work we have automatically mapped approximately 80% of diagnoses-related annotations for the samples in the Stanford TMAD to terms in ontologies. It is possible to perform a similar procedure for microarray data corresponding to those samples in the Stanford Microarray Database (SMD) to allow integrated analysis of immunohistochemistry and mRNA expression data.

We have shown that a significant proportion of diagnoses-related annotations map to terms from both the NCI thesaurus and SNOMED-CT. This mapping of a single record to terms from *different* ontologies presents a concrete annotation-driven mechanism for aligning related ontologies by using them for annotation. In future work we will examine how this approach can be extended to ontologies that are not closely related as well as compare it with existing alignment approaches such as PROMPT^[7].

Acknowledgements

We acknowledge Robert Marinelli and Matt van de Rijn for useful discussions and access to TMAD data.

References

1. Sauter, G. and M. Mirlacher, *Tissue microarrays for predictive molecular pathology*. J Clin Pathol, 2002. **55**(8): p. 575-6.
2. Rimm, D.L., R.L. Camp, L.A. Charette, *et al.*, *Tissue microarray: A new technology for amplification of tissue resources*. Cancer J, 2001. **7**(1): p. 24-31.
3. Spasic, I., S. Ananiadou, J. McNaught, and A. Kumar, *Text mining and ontologies in biomedicine: Making sense of raw text*. Brief Bioinform, 2005. **6**(3): p. 239-51.
4. Liu, C.L., W. Prapong, Y. Natkunam, *et al.*, *Software tools for high-throughput analysis and archiving of immunohistochemistry staining data obtained with tissue microarrays*. Am J Pathol, 2002. **161**(5): p. 1557-65.
5. de Coronado, S., M.W. Haber, N. Sioutos, M.S. Tuttle, and L.W. Wright, *Nci thesaurus: Using science-based terminology to integrate cancer research results*. Medinfo, 2004. **11**(Pt1): p. 33-7.
6. Bodenreider, O., *The unified medical language system (umls): Integrating biomedical terminology*. Nucleic Acids Res, 2004. **32**(Database issue): p. D267-70.
7. Noy, N.F. and M.A. Musen, *The prompt suite: Interactive tools for ontology merging and mapping*. International Journal of Human-Computer Studies, 2003. **59**(6): p. 983-1024.