The Open Biomedical Annotator

Clement Jonquet, PhD[1], Nigam H. Shah, M.B.B.S, PhD[1] and Mark A. Musen, MD, PhD[1]
[1]Stanford Center for Biomedical Informatics Research and the National Center for Biomedical Ontology, Stanford University, Stanford, CA

Abstract

The range of publicly available biomedical data is enormous and is expanding fast. This expansion means that researchers now face a hurdle to extracting the data they need from the large numbers of data that are available. Biomedical researchers have turned to ontologies and terminologies to structure and annotate their data with ontology concepts for better search and retrieval. However, this annotation process cannot be easily automated and often requires expert curators. Plus, there is a lack of easy-to-use systems that facilitate the use of ontologies for annotation. This paper presents the Open Biomedical Annotator (OBA), an ontology-based Web service that annotates public datasets with biomedical ontology concepts based on their textual metadata (www.bioontology.org). The biomedical community can use the annotator service to tag datasets automatically with ontology terms (from UMLS and NCBO BioPortal ontologies). Such annotations facilitate translational discoveries by integrating annotated data.[1]

Introduction & background

The wealth of publicly accessible biomedical data is beginning to enable cross-cutting integrative translational bioinformatics studies.[2] However, translational discoveries that could be made by mining biomedical resources are hampered because most online resources typically do not use standard terminologies and ontologies to annotate their elements (i.e., experimental data sets, diagnoses, diseases, samples, experimental conditions, clinical-trial descriptions, published papers). Currently, a researcher studying the allelic variations in a gene would want to know all the pathways that are affected by that gene, the drugs whose effects could be modulated by the allelic variations in the gene, and any disease that could be caused by the gene, and the clinical trials that have studied drugs or diseases related to that gene. The knowledge needed to address such questions is available in public biomedical resources; the problem is finding that information. The research community agrees that ontologies are essential for data integration and translational discoveries to occur.[3]

However, the variety of biomedical data is very large and the data are often annotated with free text meta-data by the researcher who created the dataset. The problem is that these text metadata are unstructured and rarely described using standard ontology terms available in the domains. This situation creates a challenge of producing consistent terminology or ontology labels for each element in public biomedical resources. Such labels would enable the identification of all related elements at a given level of granularity. For example, the Gene Ontology (GO) is widely used to describe the molecular functions, cellular location, and biological processes of gene products and allows the integration of these descriptions across several databases. A similar query on the disease dimension is currently not possible because of the lack of a common terminology to describe disease involvements for gene products.

One mechanism of achieving ontology-based annotation is map existing textual metadata describing the resource element to ontology terms allowing formulation of refined or coarse search criteria.[4,5]

The annotation of biomedical data with biomedical ontology concepts is not a common practice for several reasons:[6]

- Annotation often needs to be done manually either by expert curators or directly by the authors of the data (e.g., when a new Medline entry is created, it is manually indexed with MeSH terms);
- The number of biomedical ontologies available for use is large and ontologies change often and frequently overlap. The ontologies are not in the same format and are not always accessible via application programming interfaces (APIs) that allow users to query them programmatically;
- Users do not always know the structure of an ontology's content or how to use the ontology to do the annotation themselves;
- Annotation is often a boring additional task without immediate reward for the user.

We have previously reported on a system for ontology-driven indexing of public resources for translational bioinformatics.[1] In this paper, we

present an annotator Web service that allows scientists to utilize available biomedical ontologies for annotating their datasets automatically. The *Open Biomedical Annotator* (OBA) Web service processes the raw textual metadata and tags them with relevant biomedical ontology concepts and returns the annotations to the users. Annotations are scored according to the context from which they have been generated. The OBA Web service utilizes ontologies for annotation of biomedical data in order to facilitate interoperation, search and translational discoveries.

Methods

The OBA Web service's workflow is composed of two main steps (Figure 1). First, the user's free text is given as input to a *concept recognition tool* along with a dictionary. The dictionary (or lexicon) is a list of strings that identifies ontology concepts. The dictionary is constructed by accessing biomedical ontologies and pooling all concept names or other string forms, such as synonyms or labels that syntactically identify concepts.[1] The choice of the set of ontologies used to create the dictionary depends of the type of biomedical data the OBA Web service is used to annotate. For instance, if a user wants to annotate gene-expression datasets with disease names, then SNOMED-CT and the NCI Thesaurus could be used. The tool recognizes concepts by using string matching on the dictionary.[2] The output is a set of *direct annotations*.

This primary set of annotations serves as input for the *semantic expansion components*, which expanse the annotations extracted from the first step using the structure and/or semantics of one or more ontologies. For example:

- An *is_a transitive closure* component traverses an ontology parent-child hierarchy to create new annotations with parent concepts of the concepts involved in direct annotations. For instance, if data are directly annotated with the concept `melanoma` from NCI Thesaurus, this semantic expansion component can generate new annotations with concepts `skin tumor` and `neoplasms` because NCI Thesaurus provides the knowledge that `melanoma` *is_a* `skin tumor` and `skin tumor` *is_a* `neoplasms`. The maximum level in the hierarchy to use is parameterizable.
- A *semantic distance* component uses a given notion of concept similarity (or semantic distance)[7,8] to obtain related concepts and create new annotations. For instance, if a text is directly annotated with the concept `melanoma` from Mesh, this semantic expansion component can generate new annotations with concepts `apudoma` and `neurilemmoma` because Mesh specifies these three concepts as siblings in the hierarchy. The maximum distance (threshold) and the type of semantic distance (path/graph based or information content based) to use are parameterizable.
- An *ontology-mapping* component creates new annotations based on existing mappings between different ontologies. For instance, if a text is directly annotated with the concept `NCI/C0025202` (`melanoma` in NCI Thesaurus), this semantic expansion component can generate new annotations with concepts `SNOMEDCT/C0025202` (`melanoma` in SNOMED-CT) and `38865/DOID:1909` (`melanoma` Hunan disease) because the UMLS and the NCBO BioPortal provides the mapping information. The type of mapping to use is parameterizable.

The OBA web service is designed in manner that allows multiple semantic expansion components to be plugged-in, selected, and parameterized by a user when requesting the service.[3] As the result of the second step, the direct annotations and several sets of *semantically expanded annotations* are extracted, scored and returned.

Annotations performed with the OBA service have implicit semantics that say *this dataset is about (or deals with) this concept*. Concepts are identified by UMLS Concept Unique Identifier (CUI) [4] or National Center for Biomedical Ontology (NCBO) Uniform Resource Identifier (URI). An annotation *context* asserts whether the annotation is direct or semantically expanded. In the latter case, the component used to produce the expanded annotation is described along with the concept from which the new annotation is derived. For example, the annotation `[C0431097–ISA_CLOSURE-C0025202]` states that given text was annotated with the concept

[1] A *concept* is unique in an ontology (class). A *term* is a particular string form that identifies a concept. Usually, a concept has several terms (e.g., name, synonyms, label).

[2] Note that the concept recognizer does not execute any natural language processing techniques (stemming, spell-checking, morphological variants). However, this is not a major drawback as biomedical terminologies often contain syntactic variants for concepts as synonyms/terms.

[3] The service response time depends on the selected components as each consumes resources at a different level.

[4] NCBO is collaborating with National Library of Medicine to implement a license checking mechanism for UMLS licensed terminologies.

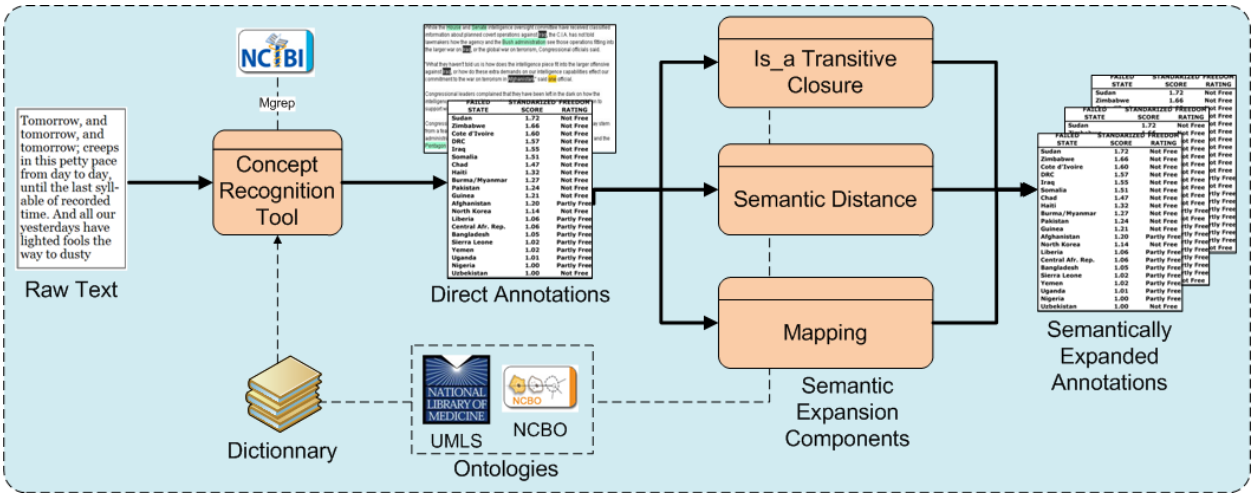

Figure 1. OBA web service workflow. First, direct annotations are created from raw text based on syntactic concept recognition according to a dictionary that use terms (concept names and synonyms) from both UMLS and NCBO ontologies. Second, different components expand the first set of annotations using ontologies semantics.

`C0431097` ('malignant melanocytic lesion') using the is_a relations of the concept `C0025202` ('melanoma'). The scoring algorithm takes into account the context (direct, expanded, level, distance, etc.) and the frequency of annotations to evaluate which concepts annotates the best the given data. Annotations can be returned to the user in different formats (text, tab delimited, XML, or OWL). The description of the results returned by the OBA Web service is available.[9]

Results

We have implemented the service using (at the moment of writing), all the (English) ontologies in UMLS (more than 94) and a subset of the NCBO BioPortal ontologies (more than 36),[5]. Those ontologies offer a dictionary of 2,627,933 concepts and 5,177,973 terms. The service uses *Mgrep,*[10] a concept recognizer with a high degree of accuracy (>95%) in recognizing disease names[11] developed by the National Center for Integrative Biomedical Informatics (NCIBI) at the University of Michigan. Mgrep implements a novel radix-tree-based data-structure that enables fast and efficient matching of text against a set of dictionary terms. Mgrep was parameterized to match all the possible concepts.[6] We have conducted[12] a comparative evaluation of Mgrep with the gold standard in the biomedical community MetaMap. For space reason, the results of this evaluation (in terms of precision, speed of execution, scalability and customizability) are described in another publication.[13]. In the second step of the workflow, our biomedical annotator currently uses an is_a transitive-closure component and leverages UMLS metathesaurus CUI-based mappings in order to expand the annotations created by Mgrep. The service is publicly available. It is deployed as a SOAP (Simple Object Access Protocol) and RESTful (REpresentational State Transfer) Web service.

We evaluated our biomedical annotator for the purpose of annotating a wide range of open biomedical resources.[1,14] For example, we annotated a set of 1,050,000 PubMed citations (title, abstract and other metadata), creating 174,840,027 annotations (18% direct, 82% expanded with is_a relations). We obtained an average of 160 annotating concepts per citation and approximately 99% of our set was annotated (with at least 1 concept), demonstrating the service's utility.

We have used the annotator service internally to process several online datasets and have constructed an *Open Biomedical Resources* (OBR) index that allows a user to search for biomedical data annotated with ontology concepts.[1,14] The OBR index is directly queriable in the NCBO BioPortal ontology repository (http://bioportal.bioontology.org/). For example, searching for "melanoma" in BioPortal returns, among others, the concept `DOID:1909` from the human disease ontology. A user can access the 13 ArrayExpress experiments, the 673 clinical trials, the

[5] Not all the NCBO BioPortal ontologies are fully usable through the REST web services API.

[6] If a text contains "cutaneous melanoma," two annotations are generated: one with 'melanoma' one with 'cutaneous melanoma' because the dictionary contains the two terms.

960 articles in PubMed, or the 10 GEO datasets related to this concept that OBA has annotated.

Use cases supported

They are many use cases for the OBA Web service. The first use was to create the OBR index, which is described in a separate publication[1,14]. The service is currently being evaluated for use in several external workflows: (1) Researchers working on Trialbank (www.trialbank.org) at the University of California, San Francisco, create annotations for HIV/AIDS clinical trials in order to provide a Web application for visualizing, and comparing the trials. They are evaluating the use of OBA to process the 'health condition', 'intervention' and 'outcomes' fields for trial records from clinicaltrials.gov. (2) Researchers at the University of Indiana are evaluating the utility of embedding the service in their research management system called Laboratree (http://laboratree.org); so that any textual annotation created in Laboratree would also have corresponding ontology term annotations.(3) Developers at Collabrx (http://collabrx.com) are embedding the service in their Rex platform for processing user generated content; and will evaluate the suitability of using medical dictionaries for processing such content. (4) Researchers at the Jackson Lab (www.jax.org) are evaluating the utility of the OBA service in triaging articles for curation based on the ontology terms recognized in their title and abstract. Each of these groups get better interoperability of their data by using ontology annotations created with OBA. We are currently working on specific evaluations of OBA when used by each of these groups.

There are many other groups who are potential users of the annotator service. For example, Cancer nanoparticle research groups at Stanford and Washington Universities aim to use the annotator service for creating ontology-based annotations for the caNanoLab. And in the Ontology Development Information Extraction project, researchers at the University of Pittsburgh are developing a set of tools for extracting meaning and codifying medical documents that can enhance the annotator service (http://www.bioontology.org/collaboration.html).

Related work

In the biomedical domain, automatic annotation or indexing of online resources is an important topic. A number of publicly available concept recognizers identify entities from ontologies or terminologies in text. For examples, see IndexFinder[15], MetaMap[13], CONANN[16], and Mgrep[10,11]. MetaMap, which identifies UMLS metathesaurus terms in text, is generally used as the gold standard for evaluating these tools. Our choice for Mgrep was made based on criteria for flexibility, speed and scalability as described before. Note that CONANN is very similar to OBA and is also available online. CONNAN aims to identify the best possible matches, whereas Mgrep in the OBA identifies the greatest number of concepts. Plus, CONANN uses term frequency to filter results. However, CONNAN is limited to UMLS and does not perform any semantic expansion step. Indeed, the knowledge contained in ontologies is rarely used to expand annotations, which gives to OBA a significant advantage. Note that the use of ontology semantics to enhance search is an active area of research.[5,,18]

Discussion and future work

The OBA Web service distinguishes itself from previous efforts for several reasons:

- It is a Web service that can be integrated in current programs and workflows;
- It uses public ontologies both to create annotations and to expand them;
- It has access to one of the largest available sets of publicly available biomedical ontologies from the UMLS metathesaurus and the NCBO BioPortal repository.

Current response times performed by the OBA Web service are ~20–25 seconds for 500 words. However, we are performing further technical improvements to OBA, such as: (1) keeping the dictionary loaded into memory between service calls (Mgrep constraint) and (2) loading the pre-computed hierarchy table into memory – in order to ensure fast response times for users

Future work will concentrate on three main issues that will determine the continued success of OBA Web service: (1) enhancement of the concept-recognition step by using natural languages processing techniques and eventually recognize 'relations,' (2) enhancement of the customizability of the service (parameters and ontologies used), and (3) enhancement of the semantic-expansion step by developing new components that use the knowledge in ontologies to relate concepts.

Conclusion

Ontology-based annotation of biomedical data plays a crucial role for enabling data interoperability and the making of translational discoveries.[1] This situation is also true for e-science generally. The need to switch from the current Web to a semantic

Web with semantically rich content annotated using ontologies has been clearly identified.[19] Meeting this need requires services (usable by humans and software agents) that can be integrated into existing data curation and annotation workflows.

We have presented a service for ontology-based annotation of biomedical data. Our biomedical annotator has access to a large dictionary, which is composed of UMLS and NCBO ontologies. OBA is not limited to the syntactic recognition of terms, but also leverages the structure of the ontologies to expand annotations.

The service workflow is currently used in a project within NCBO to annotate a large number of public biomedical resources.[14] The OBA Web service is available to (and is already being used by) the community to evaluate its utility for creating ontology-based annotation of their data. The service can be customized to their specific needs (in terms of annotations parameters and biomedical ontologies used).

Acknowledgments

This work is supported by NIH grant U54 HG004028 in support of the National Center for Biomedical Ontology, one of the National Centers for Biomedical Computing. We also acknowledge the assistance of Manhong Dai and Fan Meng (NCIBI).

References

1. Shah, N. H., Chiang, A. P., Butte, A. J., Chen, R., Musen, M. A.. Ontology-driven Indexing of Public Datasets for Translational Bioinformatics. AMIA STB, San Francisco (Mar 2008)
2. Butte, A., Chen, R.: Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. AMIA Annual Symp., Washington DC (2006) 106
3. Bodenreider, O., Stevens, R.: Bio-ontologies: Current Trends and Future Directions. Briefings in Bioinformatics 7(3) (Aug 2006) 256–274
4. Shah, N.H., Rubin, D.L., Supekar, K.S., Musen, M.A.: Ontology-based Annotation and Query of Tissue Microarray Data. AMIA Annual Symp., Washington DC (Nov 2006) 709–713
5. Moskovitch, R., Martins, S.B., Behiri, E., Weiss, A., Shahar, Y.: A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. AMIA 14(2) (Mar-Apr 2007) 164–174
6. Shah, N.H.,: Biomedical Data/Content Acquisition, Curation, Chapter in Encyclopedia of Database Systems, Springer-Verlag (2009)
7. Lee, W.J., Raschid, L., Srinivasan, P., Shah, N.H., Rubin, D., Noy, N.: Using Annotations from Controlled Vocabularies to Find Meaningful Associations. 4th Int. Work. on Data Integration in the Life Sciences. Philadelphia, PA, (Jun 2007) 264–279
8. Caviedesa, J.E., Cimino, J.J.: Towards the development of a conceptual distance metric for the UMLS. Biomedical Informatics 37(2) (Apr 2004) 77–85
9. Jonquet, C., Musen, M.A., Shah., N.H.: Help will be provided for this task: Ontology-Based Annotator Web Service. Res. Report, BMIR-2008-1317, Stanford University (May 2008)
10. Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B. Meng, F.: An Efficient Solution for Mapping Free Text to Ontology Terms. AMIA Summit on Translational Bioinformatics, San Francisco (March 2008)
11. Xuan, W., Dai, M., Mirel, B., Athey, B., Watson, S.J., Meng, F.: Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. BioLINK SIG: Linking Literature, Information and Knowledge for Biology, Vienna, Austria (Jul 2007) 55–58
12. Bhatia, N., Shah, N.H., Rubin, D.L., Chiang, A.P., Musen, M.A.: Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap. AMIA STB, San Francisco (March 2009)
13. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annual Symp., Washington DC (Nov 2001) 17–21
14. Jonquet, C., Musen, M.A., Shah, N.H.: A System for Ontology-Based Annotation of Biomedical Data. 5th Int. Work. on Data Integration in the Life Sciences. Evry, France, (Jun 2008) 144–152
15. Zou, Q., Chu, W.W., Morioka, C., Leazer, G.H., Kangarloo, H.: IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. AMIA Annual Symp., Washington DC (Nov 2003) 763–767
16. Reeve, L.H., Han, H.: CONANN: An Online Biomedical Concept Annotator. 4th Int. Work. on Data Integration in the Life Sciences, Philadelphia, PA, (Jun 2007) 264–279
17. Ide, N.C., Loane, R.F., Demner-Fushman, D.: Essie: A Concept-based Search Engine for Structured Biomedical Text. AMIA 14(3) (2007) 253–263
18. Handschuh, S., Staab, S., eds.: Annotation for the Semantic Web. Frontiers in Artificial Intelligence and Applications (96). IOS Press (2003)