Humboldt-University of Berlin

Exploiting Link Structure to Discover Meaningful Associations between Controlled Vocabulary Terms

exposé of diploma thesis of Andrej Masula

13th October 2008

supervisor: Louiqa Raschid, Ulf Leser

1 Motivation

A lot of biological research produces a huge amount of linked data. These links between objects in different data sources or resources, for instance *Online Mendelian Inheritance in Man* (OMIM) [12] or *Entrez Gene* [8], are usually created manually. Furthermore data objects are annotated with additional information to describe them as precise as possible. To make annotations semantically unambiguous, it is important to use a standard controlled vocabulary (CV) as ontologies are. If data objects in different resources are linked to each other and are annotated with CV terms, the terms that are associated to multiple objects can be used to mine new information.

Recent scientific research has discovered a methodology to identify and rank assocations between CV terms of different data objects which are related by a link [6]. This *LSLink methodology* (Life Science Link) calculates a confidence and support in the association between two CV terms and uses it for ranking with respect to the other associations. As a result users may then explore those meaningful associations, that exeed a threshold in confidence and support. As a result of a validation run, it could be shown, that in most cases the explored associations were meaningful. Most of them were well known, but some associations were classified as meaningful and unknown. These are the interesting results of this analysis, which can be considered as truely new knowledge. In a follow-up paper the LSLink methodology was extended to consider the structure of ontologies and patterns of annotation [5].
However, until now, this methodology only has been used for analyzing data objects which are connected by a direct link. Here the analyses examined objects of the Entrez Gene database annotated with terms of the *Gene Ontology* (GO) [3] as well as *PubMed* objects [9] annotated with terms of the *Medical Subject Headings* controlled vocabulary (MeSH) [15].

There is a lot of interest to mine significant associations between pairs of CV terms, whereas the underlying data objects are linked by a path that comprises more than two data objects. One has to take into account, that with consideration of additional ressources, the complexity of the graph which is spanned over the linked CV terms increases significantly.
The consideration of OMIM as an additional data source for analyzing the association rules could be beneficial, because OMIM objects are widely annotated on proteins in UniProt [14] or proteins and genes in Entrez [7]. OMIM focuses on the relationship between phenotype and genotype and contains information on all known mendelian disorders in humans. If a protein is involved in a disease and has a link to a appropriate OMIM disease dataset, one can use this dataset to explore links to related OMIM disease datasets, other NCBI data sources and external resources.

As an additional extension, one could involve further ontologies which are dealing with special applications. For instance, there is much interest in research to link annotations with clinical data. To satisfy one aspect of this concern, the Disease Ontology (DO)

has been developed [2]. DO was designed to facilitate the mapping of diseases to particular medical billing codes such as ICD9CM (International Classification of Diseases, 9th Revision, Clinical Modification) [10] and SNOMED (Systematized Nomenclature of Medicine-Clinical Terms) [4]. This ontology has practical relations to the NuGene Project [11], where DO is used to retrieve tissue samples from the tissue bank.
In the only application of the DO outside the nuGene consortium we are aware of [13], the links between proteins and DO terms were set manually. This data set is available.

2 Goals

This thesis has two goals. The method described in [6] [5] should both be improved and applied to a different data set. Regarding the first goal, we plan to extend the current framework such that also paths between annotated objects are considered, and not only direct paths. To this end, we want to integrate more data sources for links, such as OMIM. Using indirect connections will also force us to rethink the scoring method for the associations. Regarding the second goal, we plan to research associations between diseases or clinical phenotypes and genes. To this end, we could either compute associations between GO terms and OMIM entries, or between GO terms and DO terms.

3 Related Work

This research is based on the work on the LSLink methodology [6] [5]. This methodology makes use of the linked data sources Entrez Gene, PubMed and OMIM to discover potentially unknown but significant linked term pairs as described above in Motivation. Their approach and promised future work has a strong focus on extending the ontology mapping capabilities in terms of covering aggregation in GO and MeSH simultaneously and aggregating further structural levels not only child nodes. But the baseline still remains the same, i.e. analyzing genes and directly linked PubMed documents. This research should extend the existing solution by opening it to cover any data source in any graph distance.

4 Methods

This research should lead in a scoring function that is able to evaluate nodes and edges in a graph with respect to their relevance to a certain data object. Starting with this data object it depends on the concrete research interest to decide which links are to be followed and how the final graph will look like. This research will be evaluated by applying it on a model case. A potential model case is shown in Figure 1. Like it was mentioned before one goal is to make use of the Disease Ontology in this research. One potential application is described as follows. At first one could track links from a gene or protein (green frame) to linked OMIM objects. Afterwards one can make use of links to genes (red frames) that are annotated to the OMIM objects. The newly discovered genes may contribute links to PubMed objects. These PubMed objects are likely referenced by interesting genes that are not yet associated to the OMIM disease (blue frames). Maybe the new gene objects are

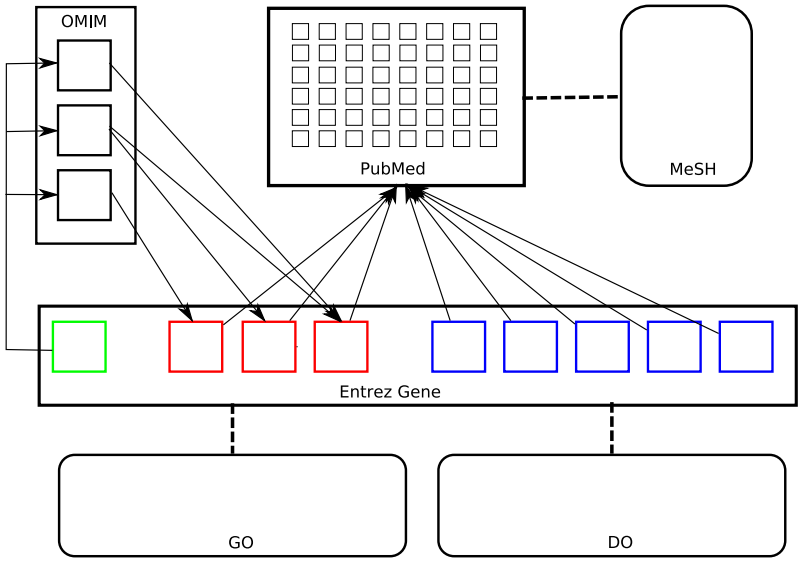

Figure 1: Example for discovery of data objects from different sources

valuable as well but certainly less important than the still explored genes because of its semantical distance to the queried (green) data object. As a result a comprehensive graph will emerge. The red as well as the blue genes have perhaps been previously unrelated to the green query gene but will be now part of follow-up evaluations. This could be a GO - DO, GO - GO or DO - DO examination in terms of a treatment with the LSLink algorithm but in consideration of the data object scores. Some method related issues are discussed shortly below.

Which elements should be part of the scoring method? This is one of the main objectives of this research. From a general point of view it could be stated that each node, i.e. data object, in the graph will have potentially a different weight dependend on its relevance to the requested data object. Some simple measures may be *path length* or *minimum distance to the queried data object*. Furthermore, a discrimination of types of data objects and links between them are useful. For instance, a link from a gene to a protein may have more strength than a link from a gene to an OMIM object. According to this it seems to be valuable to take the *authority flow approach* [16][1] into account.

How should the data object's score expand the LSLink score? It would be very efficient if the data object's score could be combined with the calculated LSLink score by a simple operator. If there was evidence confirming these thoughts it would make it easy to apply the new scoring method to the existing LSLink methodology.

How to map the user's intention into a request? If a user queries a relevant data object according to the LSLink methodology all related relevant data objects are connected

through direct links of equal weight. In brief, the term *relevant data object* means a data object that is in focus of actual examination due to its annotation with CV terms. In contrast this research project assumes that (1) multi-step paths between two relevant data objects can exist and (2) there may exist many paths between two relevant data objects. The user needs to be supplied with the power to filter the paths that are important to him. In fact, handling this concern will be essential for generalizing this approach and providing its practical application. For this reasons it could be a valuable idea to adopt the query language of the *Graph Information Discovery Framework* (GID) [17]. But these issues are not a main objective of this research.

How can the background data set be created efficiently? The amount of data to handle with is expected to get huge. The reason is, building and querying a complete graph which depends on a base set of data objects, e.g. all human genes with GO annotation available, will cost a lot of computational power. Further the background data set has to be renewed at least with each change of the base data objects. These concerns may be subject of further optimization research.

References

[1] Andrey Balmin, Vagelis Hristidis, and Yannis Papakonstantinou. Objectrank: Authority-based keyword search in databases.

[2] Disease Ontology Project. *Disease Ontology*, 2008. `http://diseaseontology.sourceforge.net`.

[3] Gene Ontology Consortium. *Gene Ontology*, 2008. `http://www.geneontology.org/GO.doc.shtml`.

[4] International Health Terminology Standards Develeopment Organisation. *SNOMED-CT*, 2008. `http://www.ihtsdo.org/snomed-ct`.

[5] Woei-Jyh Lee, Louiqa Raschid, Hassan Sayyadi, and Padmini Srinivasan. Exploiting ontology structure and patterns of annotation to mine significant associations between pairs of controlled vocabulary terms. In *DILS*, pages 44–60, 2008.

[6] Woei-Jyh Lee, Louiqa Raschid, Padmini Srinivasan, Nigam Shah, Daniel L. Rubin, and Natasha Fridman Noy. Using annotations from controlled vocabularies to find meaningful associations. In *DILS*, pages 247–263, 2007.

[7] National Center for Biotechnology Information. *Entrez*, 2008. `www.ncbi.nlm.nih.gov/sites/entrez`.

[8] National Center for Biotechnology Information. *Entrez Gene Database*, 2008. `www.ncbi.nlm.nih.gov/sites/entrez?db=gene`.

[9] National Center for Biotechnology Information. *PubMed*, 2008. `http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed`.

[10] National Center for Health Statistics. *ICD-9-CM*, 2008. `http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm`.

[11] NUgene Project. *NUgene*, 2008. `http://www.nugene.org/`.

[12] Online Mendelian Inheritance in Man (OMIM). *OMIM*, 2008. `http://www.ncbi.nlm.nih.gov/Omim/allresources.html`.

[13] Wyatt T. Clark Brandon J. Peters Amrita Mohan Sean M. Boyle Sean D. Mooney Predrag Radivojac, Kang Peng.

[14] UniProt Consortium. *UniProt*, 2008. `http://www.uniprot.org/`.

[15] U.S. National Library of Medicine (NLM). *MeSH*, 2008. `http://www.nlm.nih.gov/mesh/meshhome.html`.

[16] Ramakrishna Varadarajan, Vagelis Hristidis, and Louiqa Raschid. Explaining and reformulating authority flow queries.

[17] Ramakrishna Varadarajan, Vagelis Hristidis, Louiqa Raschid, María-Esther Vidal, Luis Ibáñez, and Héctor Rodríguez-Drumond. Flexible and efficient querying and ranking on hyperlinked data sources.