A Protein Annotation Framework Empowered with Semantic Reasoning

Jemma X. Wu, Edmond J. Breen, Xiaomin Song,
Brett Cooke, and Mark P. Molloy

Australian Proteome Analysis Facility, Maquarie University, Sydney, Australia
{jwu,ebreen,xsong,bcooke,mmolly}@proteome.org.au

Abstract. This paper presents an association discovery framework for proteins based on semantic annotations from biomedical literatures. An automatic ontology-based annotation method is used to create a semantic protein annotation knowledge base. A semantic reasoning service enables realisation reasoning on original annotations to infer more accurate associations. A case study on protein-disease association discovery on a real-world colorectal cancer dataset is presented.

Keywords: Protein annotation, bioinformatics, semantic reasoning

1 Introduction

To bridge the gap between the biomedical science and bioinformatics, many biomedical ontologies have been created in the past few years. Ontology-based semantic annotation for biomedical entities are of interest to both biomedical researchers and general public. Meanwhile, the biomedical domain has a large and fast-growing amount of literature resources, among which MedLine[1] is the primary publication repository for biomedical research. Ontology-based biomedical text annotation has shown promising progress and several tools have been successfully developed and evaluated in biomedical text mining problems[2, 5, 4]. However, these generic text-based biomedical annotation tools only provide concept level annotations. The ability to do protein-oriented semantic annotation will greatly benefit the proteomics research by enabling easy protein association discovery. Also, traditional text-based annotation tools tend to create excessive annotations and some tools expand the raw annotations by using semantic reasoning[3]. Inferring of the most informative and accurate annotations will be very valuable to efficient and accurate association discovery.

This paper proposes an integrated high performance framework that leveraging protein annotations and semantic reasoning to an informative protein-biomedical concepts association Knowledge Base(KB). Starting from a list of proteins, the system automatically retrieves a pool of MedLine citations and annotates the proteins using pre-defined biomedical ontologies. A realisation reasoning service is applied to infer more accurate protein association information. In our preliminary study, the focus is on the discovery of potential *protein-disease associations*. A case study on discovering protein-disease associations for a real-world colorectal cancer tissue protein dataset is presented.

[1] http://www.ncbi.nlm.nih.gov/pubmed

2 SPRAM: A Semantic PRotein Annotation framework based on MedLine

We propose a *Semantic PRotein Annotation method based on MedLine* (SPRAM) which produces semantically inferred protein annotations based on biomedical literatures and ontologies (Fig.1). *SPRAM* starts with a list of proteins of in-

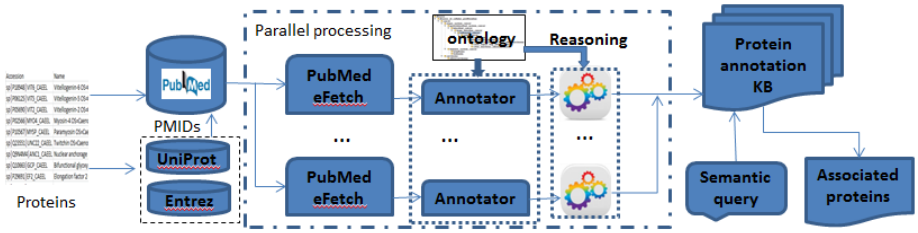


Fig. 1. The framework of Semantic Protein Annotation method based on MedLine.

terest. A list of mapped MedLine publication IDs (PMIDs) for each candidate protein is produced by searching the UniProt[2] and Entrez Gene[3] databases. A parallel process runs to execute PubMed article fetching, semantic annotation and realisation reasoning based on biomedical ontologies, such as Human Disease Ontology (HDO[4]), for the pool of candidate PMIDs. The MedLine citations retrieved by the PubMed's Eutil service are further filtered by the co-occurrence of protein names. We choose the BioPortal's annotator service[5] with no semantic expanding as our annotator. The raw annotated concepts are post-processed by a realisation reasoning service to generate a set of clean and accurate protein annotations and inserted into the knowledge base. Biologists can then issue semantic queries to retrieve all proteins which are associated with one or more concepts in the ontology.

3 Semantic reasoning for protein annotations

In biomedical ontology annotations, very often an instance is annotated with multiple classes with subclass relationships in the ontology. To the best of our knowledge, existing biomedical annotation tools with semantic reasoning functionalities only do semantic expanding[3]. There has been no prior work on drilling down the annotations to most specific concepts by using the semantic reasoning. Despite, in many cases, the most specific classes of a protein, can more accurately represent their biomedical categorical information. For example, the traditional protein Gene Ontology analysis that shows the distribution of biological process or molecular functions nearly always bias towards the top-level classes in the ontology[1].

[2] http://www.uniprot.org/
[3] http://www.ncbi.nlm.nih.gov/gene
[4] http://disease-ontology.org/
[5] http://www.bioontology.org/wiki/index.php/Annotator_Web_service

We developed a specialised realisation reasoning service for dynamically generated annotations. Different to the traditional Description Logic most specific concept reasoning, our algorithm works on a dynamic set of annotations on the fly instead of assertions in a static KB. Only the most specific annotations will be stored in the KB. The algorithm takes a set of semantic protein annotations, ϵ, and an ontology, $\mathcal{O}$, that ϵ is based on. A most specific class set, $\epsilon^{'}$, is initialised to be an empty set, $\emptyset$. For each class $t \in \epsilon$ for each protein, find all subclasses of t in $\mathcal{O}$, i.e., $\{C_i\}$ where $C_i \sqsubseteq t \in \mathcal{O}$. Class t is added to $\epsilon^{'}$ if $\{C_i\} \cap \epsilon = \emptyset$, i.e., t is the most specific annotation in ϵ given ontology $\mathcal{O}$. The algorithm outputs the most specific class set $\epsilon^{'}$ which will be inserted into the nascent KB.

Fig.2 shows an example of the effect of applying realisation reasoning on the disease annotations for a protein with a UniProt ID "O43175". Class "disease", "cancer" and "carcinoma" in the original annotation set are all realised to "adenocarinoma" because the last concept is subsumed by those three concepts and it is also in the original annotation set. This is important because it more accurately represents the biomedical categorical information and reduces complexity.

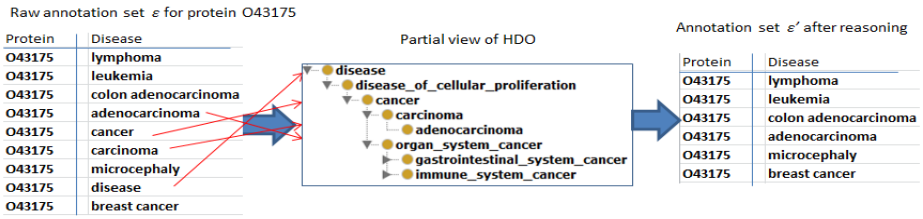


Fig. 2. An example showing the change to the protein disease annotations after realisation reasoning. Left: the original disease annotation. Middle: a simplified partial view of Human Disease Ontology. Right: the disease annotation after realisation reasoning.

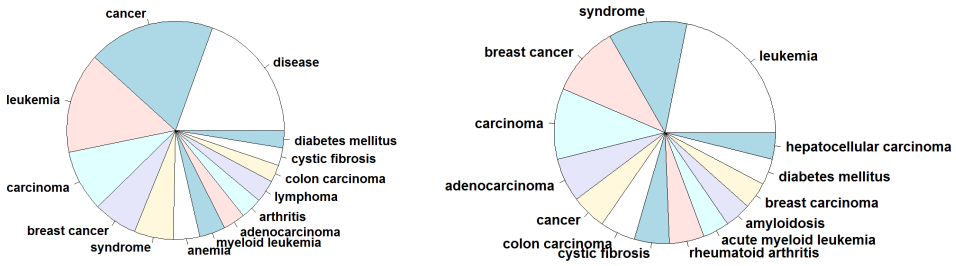
4 Case study: discovery of proteins-disease associations for colorectal cancer tissues

Jankova et al. found 45 up-regulated proteins in colorectal cancer tissues by using experimental protein iTRAQ analysis[1]. The biologists would like to know what diseases are related to these proteins and if the associations to the colorectal cancer have been discovered before.

To help biologists achieve these goals, we take these 45 proteins and use our *SPRAM* workflow to assist discovering the potential diseases associated with these proteins. The result was: 1080 MedLine citations, 354 diseases associations based on HDO, that was reduced to 241 unique associations after realisation reasoning. *SPRAM* returns a set of protein-disease association to the biologists. That includes also the source reference titles and URLs. The biologist can then use this result to help validate these associations easily by tracing back to the references.

Fig.3 shows the changes of the distribution of the diseases associated with these 45 proteins before and after realisation reasoning. The distribution after

realisation reasoning represents more accurate and sensible information. For example, the top distributed concept, disease, was removed and the next, cancer, was greatly reduced, thereby producing a clearer set of associations.



(a) DOA distribution before reasoning (b) DOA distribution after reasoning

Fig. 3. The comparison of the Disease Ontology Annotation (DOA) distributions before and after realisation reasoning for the 45 up-regulated colorectal cancer proteins

To find proteins reported as colorectal cancer related, the biologist issues a query using the concept, "colorectal cancer". The semantic reasoning service rewrites this query into a union of this concept and all of its subclasses. The result shows that 6 proteins (CEA, NNE, HSP 84, NPM, 3-PGDH and UEV-1) reported in the literature as being related to colorectal cancer.

5 Conclusion

This paper proposes an automatic protein-oriented association discovery framework based on semantic annotations from literature. A semantic reasoning service provides realisation reasoning. We demonstrate the usage of our system on protein-disease association discovery using a real-world colorectal cancer protein dataset. In upcoming work, focus will be given to a ranking model of protein associations and customisable selection of protein-PMID mappings.

References

1. L. Jankova, C. Chan, C. Fung, X. Song, S. Kwun, M. Cowley, W. Kaplan, O. Dent, E. Bokey, P. Chapuis, M. Baker, G. Robertson, S. Clarke, and M.P. Molloy. Proteomic comparison of colorectal tumours and non-neoplastic mucosa from paired patient samples using iTRAQ mass spectrometry. *Mol Biosyst*, 7(11), 2011.
2. R. Jelier, M. Schuemie, A. Veldhoven, L. Dorssers, G. Jenster, and J. Kors. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Gen biology*, 9(6), 2008.
3. Clement Jonquet, Nigam H. Shah, and Mark A. Musen. The open biomedical annotator. In *AMIA-TBI'09*, pages 56–60, 2009.
4. H. Lpez-Fernndez, M. Reboiro-Jato, D. Glez-Pea, F. Aparicio, D. Gachet, M. Buenaga, and F. Fdez-Riverola. Bioannote: A software platform for annotating biomedical documents with application in medical learning environments. *Computer Methods and Programs in Biomedicine*, 111(1):139 – 147, 2013.
5. Mariana Neves and Ulf Leser. A survey on annotation tools for the biomedical literature. *Brief Bioinform*, 18, December 2012.