A Protein Annotation Framework Empowered with Semantic Reasoning

Jemma Wu, Edmond J. Breen, Xiaomin Song, Brett Cooke, and Mark P. Molloy
Australian Proteome Analysis Facility, Macquarie University, NSW, Australia

Introduction

The ability to do protein-oriented semantic annotation will greatly benefit the proteomics research by enabling easy protein association discovery. Inferring of the most informative and accurate annotations will be very valuable to efficient and accurate association discovery.

We proposes an integrated high performance framework leveraging protein annotations and semantic reasoning to an informative protein-biomedical concepts association Knowledge Base(KB). The distinguishing features of our system include (1) semantic protein-oriented knowledge mining based on MedLine (2) fuzzy matching of various protein alternative names (3) automated inference of the most specific protein annotations (4) query transformation via semantic expansion.

A case study on discovering protein-disease associations for a real-world colorectal cancer tissue protein dataset is presented.

SPRAM: A Semantic PRotein Annotation framework based on MedLine

A list of mapped MedLine publication IDs (PMIDs) for each candidate protein is produced by searching the UniProt [1] and Entrez Gene [2] databases. A parallel process runs to execute PubMed article fetching, semantic annotation and realisation reasoning based on biomedical ontologies for the pool of candidate PMIDs. The MedLine citations retrieved by the PubMed's Eutil service are further filtered by fuzzy matching of protein names and their synonyms with the article title and abstract. We choose the BioPortal's annotator service [3] 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.

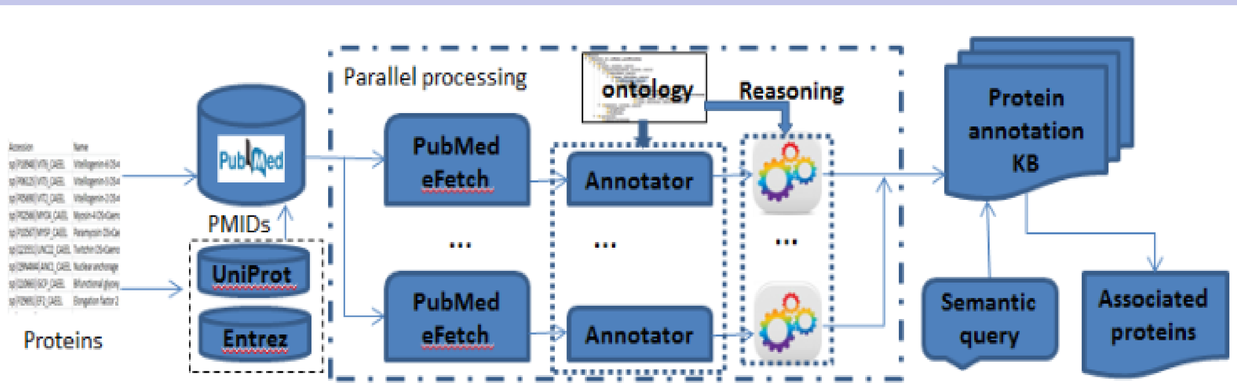


Fig. 1. The framework of Semantic Protein Annotation method based on MedLine.

Most specific annotations

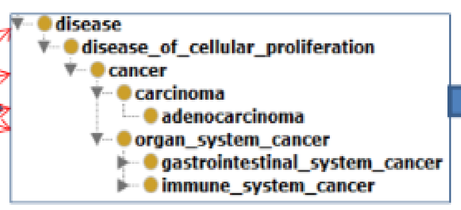
In biomedical ontology annotations, very often an instance is annotated with multiple classes with subclass relationships in the ontology. Despite, in many cases, the most specific classes of a protein, can more accurately represent their biomedical categorical information.

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.

Raw annotation set ε for protein O43175

Protein	Disease
O43175	lymphoma
O43175	leukemia
O43175	colon adenocarcinoma
O43175	adenocarcinoma
O43175	cancer
O43175	carcinoma
O43175	microcephaly
O43175	disease
O43175	breast cancer

Partial view of HDO



Annotation set ε' after reasoning

Protein	Disease
O43175	lymphoma
O43175	leukemia
O43175	colon adenocarcinoma
O43175	adenocarcinoma
O43175	microcephaly
O43175	breast cancer

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.

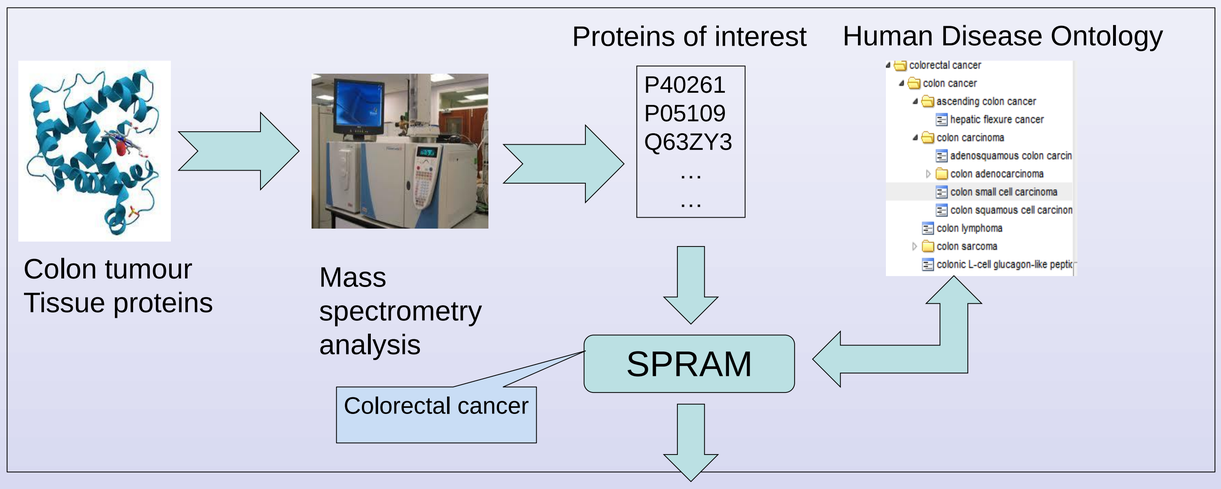
References:

1. Uniprot: http://www.uniprot.org/
2. Entrez Gene: http://www.ncbi.nlm.nih.gov/gene
3. Bioportal annotator: http://www.bioontology.org/wiki/index.php/Annotator_Web_service
4. 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. Molly. Proteomic comparison of colorectal tummours and non-neoplastic mucosa from paired patient samples using iTRAQ mass spectrometry. *Mol Biosyst*, 7(11), 2011.

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

Jankova et al. [4] found 45 up-regulated proteins in colorectal cancer tissues by using Mass Spectrometry protein analysis. 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.

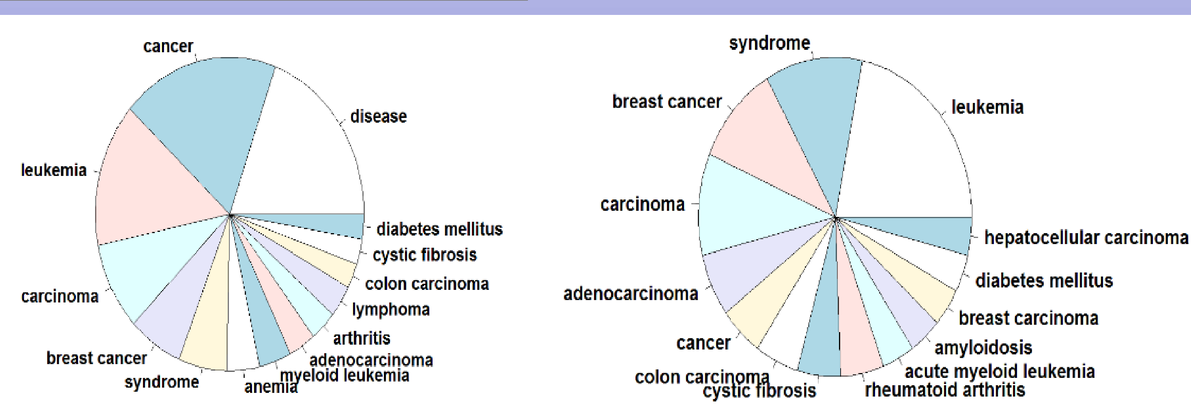
To help biologists achieve these goals, we take these 45 proteins and use our SPRAM workflow to assist discovering the potential diseases associated with these proteins.



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. Table 1 shows the proteins that are discovered as associated with colorectal cancer.

ProteinID	ProteinName	Disease	ReferenceID	ReferenceTitle
P06748	Nucleophosmin	colon carcinoma	pmid:9150948	A two-dimensional gel database of human colon carcinoma proteins.
P06733	Alpha-enolase	colon carcinoma	pmid:9150948	A two-dimensional gel database of human colon carcinoma proteins.
Q13404	TRAF6-regulated IKK activator 1 beta Uev1A	colon carcinoma	pmid:9418904	Role of UEV-1, an inactive variant of the E2 ubiquitin-conjugating enzymes, in in vitro differentiation and cell cycle behavior of HT-29-M6 intestinal mucosecretory cells.
O43175	D-3-phosphoglycer	colon adenocarcinoma	pmid:10713460	Nucleotide sequence and differential expression of the human 3-phosphoglycerate dehydrogenase gene.
P06731	Carcinoembryonic	colon adenocarcinoma	pmid:3670312	Isolation and characterization of full-length functional cDNA clones for human carcinoembryonic antigen.

Table 1. Protein associated with colorectal cancer



(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

Fig. 3. shows the changes of the distribution of the diseases associated with these 45 proteins before and after realisation reasoning. The top distributed concept, disease, was removed and the next, cancer, was greatly reduced, thereby producing a more specific set of associations.
Fig. 4. shows the top 10 protein-diseases associations among the 45 up-regulated colorectal tumour tissue proteins based on the number of unique supporting Pubmed references. These proteins have potential associations to colorectal cancer.

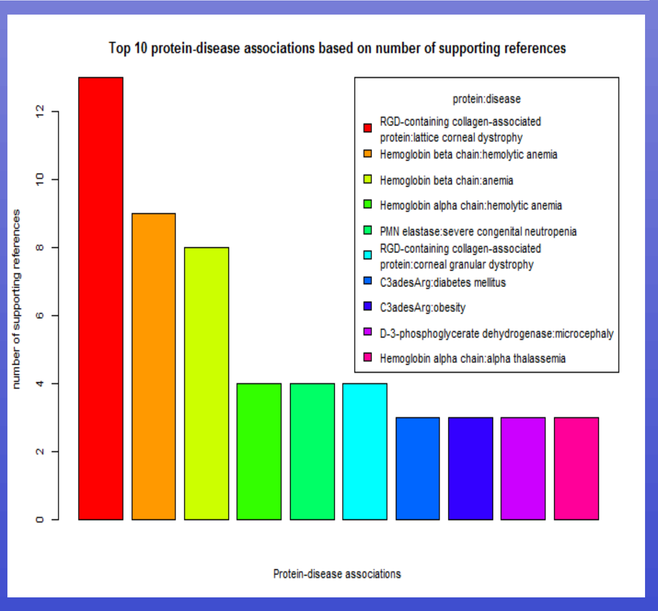


Fig.4 Top protein-disease associations

Summary
We propose a semantics empowered protein association discovery framework, which aims to help biologists post-analyse their interested proteins based on literature. We have applied this workflow to discover colorectal cancer related proteins from a list of 45 up-regulated colorectal cancer tissue.

APAF – Australia's premier proteomics provider since 1995