Interactive platform for semantic gene expression analysis of Alzheimer's disease

Toshiaki Katayama[1], Masataka Kikuchi[2], Soichi Ogishima[3]

[1] Database Center for Life Science, Research Organization of Information and Systems, Tokyo, Japan
ktym@dbcls.jp
[2] Department of Molecular Genetics, Center for Bioresources, Brain Research Institute, Niigata University, Niigata, Japan
kikuchi@bri.niigata-u.ac.jp
[3] Department of Bioclinical Informatics, Tohoku Medical Megabank Organization, Tohoku University, Miyagi, Japan
ogishima@megabank.tohoku.ac.jp

Abstract. In the course of gene expression analysis, it is required to interpret data by referencing knowledge bases of genetics, pathways, diseases and drugs. However, because those external resources are often stored in distributed databases in various formats, it is hard for biomedical scientists to use them in combination. Semantic Web technologies are suitable for integration of those heterogeneous datasets using Resource Description Framework (RDF) and providing a faceted search interface. In this work, we report an application of semantic data integration with faceted search for interactive analysis of gene expression data of Alzheimer's disease. This framework can be easily extended for other biomedical domains.

Keywords: data integration, linked open data, faceted search, gene expression analysis, Alzheimer's disease

Introduction

Difference of gene expressions of target and control samples can be measured by microarray technology. Those data are primarily processed to find genes having excessive fold change with a significant probability. Interpretation of relationship between those candidate genes and a hypothesis behind a design of the experiment depends on supportive information from existing knowledge. This includes genetics, pathways, diseases and drugs which are independently stored in different databases and only available in incompatible data formats. For integration of heterogeneous datasets, it is becoming standard to use RDF in life sciences [1]. With semantic Web technologies, relations of data can be represented as a linked graph which provides powerful means to explore a set of related information often referred to as faceted

search. Here we report an application which can be used for exploring gene expression data with integrated semantic datasets for Alzheimer's disease.

Results

We introduced RDF for data integration of Affymetrix microarray probes, Human Genome Organization (HUGO) gene annotations, UniProt protein annotations, gene-drug interactions in DrugBank, disease susceptibility genes in AlzGene [2] and molecular interactions in AlzPathway [3] databases for Alzheimer's disease. Most of those datasets except for UniProt are converted into RDF by in-house developed scripts. Based on these datasets stored in a triple store, we developed a new Web application (Fig. 1) that accepts gene expression data and provides interactive interface to explore the list of candidate genes with various facets including gene expression ratio, chromosomal positions of genes, protein interactions with drugs, known genetic diseases, and pathway information.

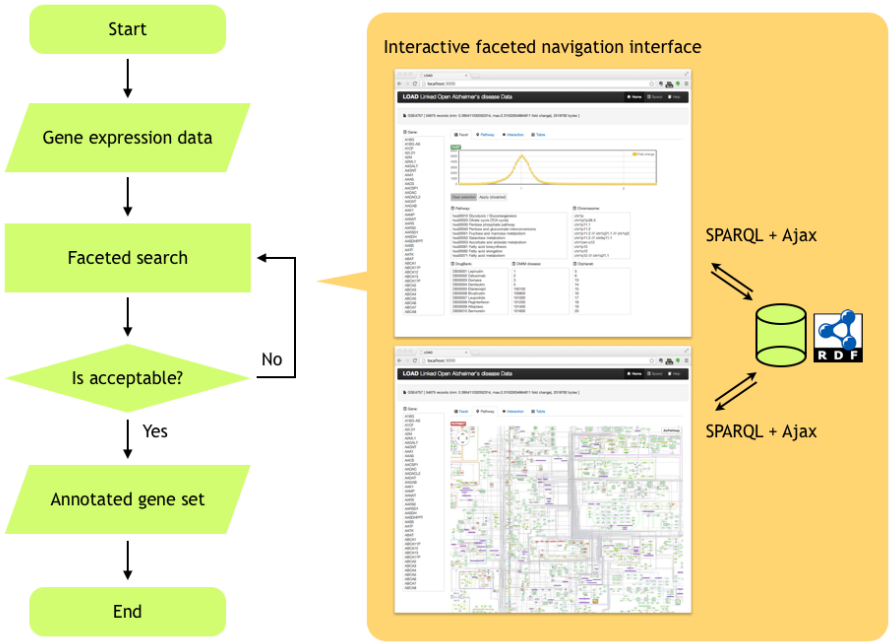

Figure 1. Workflow of interactive faceted analysis.

The server is implemented in Node.js which accesses a backend triple store to perform faceted search by SPARQL Protocol and RDF Query Language (SPARQL). The Web interface is developed using D3.js, Bootstrap and Google Map API.

Discussions

Our application is designed for exploring candidate genes of Alzheimer's disease based on the user-supplied gene expression data by referencing existing knowledge with a faceted navigation. We found that Semantic Web technologies are well-suited for data integration and implementation of a faceted search interface. This means that the method and the interface we developed can be extended for any other biomedical domains which requires integration of heterogeneous data where interactive search can assist a data-driven approach.

References

1. Katayama T, Wilkinson MD, Micklem G, *et al*.: **The 3rd DBCLS BioHackathon: improving life science data integration with semantic Web technologies.** *Journal of Biomedical Semantics* 2013, **4**:6.

2. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE: **Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database.** *Nature genetics* 2007, **39**:17–23.

3. Mizuno S, Iijima R, Ogishima S, *et al*.: **AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease.** *BMC systems biology* 2012, **6**:52.