2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Ontology-based semantic data integration for *Lactococcus Lactis (L. lactis):* Connecting cross-references heterogeneous data sources

Muhammad Akmal Remli [1], Safaai Deris [2] and Afnizanfaizal Abdullah[3 +]

Department of Software Engineering, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Johor, Malaysia

Abstract. This paper presents an approach of semantic web technologies applied in data integration for gram-positive bacteria organism, *Lactococcus lactis subsp. cremoris strain MG1363 (L. lactis)*. *L. lactis* data sources are heterogeneous and not semantically connected among cross-references databases. Researchers continuously study and perform scientific experiments related to modeling and simulation to produce result in order to improve protein and vitamin production using this organism. The goal of this work is to construct an integration approach for bacteria organism *L. lactis* that correctly combines biological databases using semantic web and ontology, thus allows biological question to be answered using queries (e.g. SPARQL). In this paper, we demonstrate how semantic web components such as Ontology Web Language (OWL) and Resource Description Framework (RDF) can be used to represent and integrate these resources. Gene, protein, pathway and ontology are main sources to make this organism semantically integrated. The sources are acquired from Entrez Gene (EG), Universal Protein Resource (UniProt), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology (GO). We illustrate the result of our approach by presenting semantically connected knowledge instances from gene, protein, pathway and ontology data sources by using identifier. In addition, this integrated organism is vital which can lead for further analysis and hypothesis formulation in biological research.

Keywords: semantic data integration, bioinformatics, RDF, GOA, knowledge integration

1. Introduction

The most important phase of research in biological field relies on data gathering and analysis. These processes are performed for the first phase in system biology to enable the biologists to understand the wide range of biological knowledge, for instances; to understand gene function, which gene is related to pathway and what type of biological process is involved. Current collaborating research in this field however, study and perform experiment individually by "*omics"* studies such as genomics, proteomics and metabolomics. In system biology, a study in biological process at system level is significance instead of focusing on molecular level in bioinformatics. Understanding of genes and proteins continues to be important, however sometimes it is also needed to capture the whole biological systems in order to understand how the biological system works [10].

Lactococcus lactis subsp. cremoris strain MG1363 (*L. lactis*), is most efficient cell factory for vitamin and protein production [13]. This organism widely used in food fermentation and dairy product. To date, there is no publicity integrated database available for this organism. Existing data generated from research effort is also not much compared with other organism like *Escherichia coli (E. Coli),* which has almost complete information available. Thus, this leads challenges to build knowledge base repository for the *L. lactis* organism.

In the experimental research using this organism, particularly in vitamin production, firstly biologist needs to find all related term of *riboflavin*. Then, they might choose *riboflavin metabolic process* and want to

+ Corresponding author. Tel.: +60127316310; fax: +6075533033
E-mail address: akmalmuhd@gmail.com

know what gene product annotated with this term at Gene Ontology (GO) [7] website using AMIGO browser. However, the relevant and related information could not be found because GO website is not loaded with this organism. Otherwise, biologists need to build their own local copy of GO database and load the annotation for the organisms to browse it. Finally, each gene found associated with *riboflavin metabolic process* terms in local databases must be searched individually at National Center for Biotechnology Information (NCBI) Entrez Gene (EG) [3] to find related information including which pathway this gene involved and which other genes involved in particular pathway. This is however, not systematic, time consuming and exposed to error prone as technical expert needed to establish local GO database and annotation as web interface must be utilized using AmiGO and Entrez gene browser. In addition, retrieving complex knowledge cannot be done if biologist wants to know; for example trying to connect the gene ontology term and relation in protein and pathway database. Currently, no system supports biological queries such as: "*Which gene participated in Bacterial Secretion System pathway and associated with ATP Binding of molecular function in Lactococcus lactis strain MG1363"*.

2. Biological data integration

Data integration in life science field is the most challenging area to the researchers in order to understand the characteristic of the particular species. Generally, there are two types of data integration approach, namely data warehouse approach and federated approach. These two approaches however, have many limitations and difficulty such as performance issues, heterogeneous nature (e.g. different identifier) and syntactic problem [14]. In order to overcome these limitations, semantic web data integration has emerged. Nowadays, semantic data integration in biological area is already matured. Semantic web which proposed by Tim Berners Lee in 2001 [6] seems widely accepted in the life science research. The Resource Description Framework (RDF) is the standard that used in semantic web, for the purpose to describe resources containing graph triples (resource, property, value). Meanwhile, Ontology Web Language (OWL) is more expressive than RDF. It also supports reasoning in order to infer consistency of particular domain knowledge.

In the biological field, different organisms have different data in the same structures (also called schema). For instance, the most studies human taxonomy, *Homo Sapiens* (*H. sapiens*) have very wide range of data generated from researchers through experiments performed in laboratory across the world. The gene *H. sapiens in* Entrez Gene (EG), *amyloid beta (A4) precursor protein (APP)* contain details and comprehensive information about gene and other related elements. Annotation data cross-reference database is also provided such as the gene ontology term that is annotated with, pathway which this gene involved and disease information associated with the gene. However, not all species already generated this particular information. There are many other organisms and species which have limitation of such data, especially the new completed sequenced organisms and new isolated strain. This scenario occurs due to the amount of biological research of specific organisms is being published intensively. Thus, information will be generated time by time and leads to a real challenging work to integrate them and make further analysis in biological research.

Semantic web approach cannot be separated with biological field. Current research trend uses semantic web to integrate data from distributed and heterogeneous sources. In addition, these data can be retrieved using query languages (SPARQL and SeRQL). In term of knowledge integration, the more heterogeneous data is connected, the more knowledge can be derived. Semantic System Biology (SSB) [5] for instance, has established over 175 million triples and stored them in one repository. Query interface provided allowing user to get all needed information easily. However, this work is not very comprehensive information due to lack of pathway information and other interested species. BioCyc project [9] also provides their own database containing integration of gene, pathway, protein and other related information. Instead of strain MG1263, other strains of *L. lactis* in BioCyc contain very detail information and user also can visualize the data. There are many other public information of biological database related in this work such BioWarehouse [11], which using relational database management system (RDBMS).

3. Data sources

The Entrez Gene (EG) is the main source for this integration process. EG contains gene related information from a wide range of species at NCBI. Each species contains record such as gene type, reference sequence, maps, pathway, variations and phenotypes. EG data of *L. lactis* was downloaded in *ASN1* format and converted to XML using *gene2xml* tool provided by NCBI. The size of the XML file is 50MB, containing 2,598 elements of genes. The version of downloaded data is date of August 2011.

The Gene Ontology (GO) is the most prominent ontology in bioinformatics, providing shared vocabulary containing terms within gene and gene product across species and databases. GO database has three main upper level components: *biological processes*, *molecular functions* and *cellular component*. AMIGO browser at GO website can query and search gene product or term in limited organisms. In this work, we use daily term database sizing 45.6 MB in RDF format.

The Gene Ontology Annotation (GOA) [1] from EBI contains high-quality GO annotations to protein in UniProt Knowledge Base (UniProtKB). Due to GO information not included in EG, we choose GOA as '*connector'* to integrate these sources. The GOA file format (*goa)* was converted to RDF using *goa2rdf* tool [2], the perl program is used to transform GOA data format to RDF format. The transformation results contain RDF triples of statement and resource of gene ontology and protein identifier.

4. Approach

The propose integration approach is summarized as follows. For constructing RDF knowledge base consisting EG and KEGG database, we follow method in [12]. The original XML file from EG are converted to RDF and make relation semantically between KEGG instance by using EKOM ontology and BIOPAX ontology. EKOM is the ontology model for Entrez Gene data while BIOPAX used to represent knowledge of pathway (applied in KEGG, Reactome and BioCyc). The result of RDF triple then loaded into SESAME OpenRDF [4], the open source RDF management repository, together with GO and GOA annotation of UniProt RDF. With all these sources integrated into RDF repository, queries can be formulated to retrieve complex biological questions. Figure 1 shows the proposed integration approach.

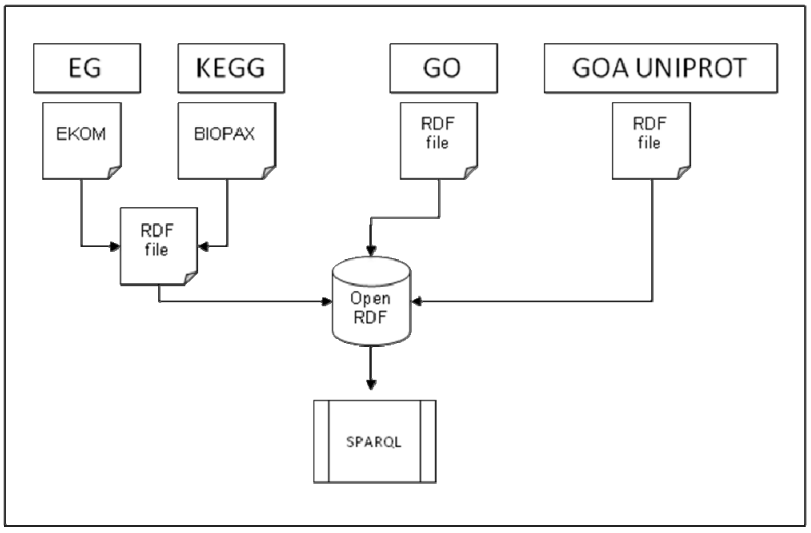

Fig. 1: The proposed semantic data integration approach

4.1. EG and KEGG to RDF

There are many types of format can be used to represent RDF data sources. The work in [12] transform XML to RDF by converting every element in EG as resource by representing them as Universal Resource Identifier (URI). This allows more expressive and meaningful information to each graph in RDF that consistency connected using EKOM and BIOPAX ontology as the schema layer. However, in this work, we transform them into more standardize representation of RDF structure by using every instance of sources as a value string element instead of using URI reference. For example, the basic information about gene like gene identifier and the name of gene, represent as a string value from the element <*ekom:gene*> and

<*ekom:symbol*> except for element of reference from different sources like protein and pathway which are expressed in URI form using URI notation '#' (eg: http://localhost:8080/ontology/lactococcus#protein). This allowed query can be made by connecting GO term more easily in standard way. The Extensible Stylesheet Language Transformation (XSLT) is used in order to convert XML EG to RDF. Java program are made using Java API for XML Processing (JAXP) to transform XML as input file with XSL template which contain the stylesheet rule to express RDF output.

5. Result

Integrated repository for *L. lactis* contains knowledge sources covering gene, protein, pathway and ontology. Figure 2 below shows the example of connected RDF graph using the approach described above. Retrieving knowledge over integrated RDF source can be formulated using SPARQL query. The query "*Which gene participated in Bacterial Secretion System pathway and associated with ATP Binding of molecular function in Lactococcus lactis strain MG1363"* returns one result of gene, which is *secA (preprotein translocase subunit SecA)*. The path in Figure 2 starts with the gene identifier in EG and this gene actually is the output of the query described above. This path involved *L. lactis* gene from EG, *secA* (EG:4798902) contains gene product (protein) *Protein translocase subunit SecA*(UniProt:A2RHJ5) from UniProt. EKOM ontology defines relationship between gene and gene product using *has_product* object properties. The EG data does not contain information about the GO term annotated with this gene product. Thus, GOA source is used to define relationship between gene product and annotated GO term. This protein is annotated using relationship *associated_with* object properties, with GO term *ATP Binding* (GO:0005524) and have relation with top ontology of GO, *molecular function* (GO:00003674) using *is_a* relationship. The protein is involved in Bacterial secretion system pathway from KEGG (KEGG:llmg_0124) and relationship *functionally_related_to* defined in EKOM used to semantically relates EG with BIOPAX ontology of protein for KEGG data. The query easily can be formulated in order to test the biological hypothesis or inference particularly in first phase of system biology research. Moreover, complex relation can be connected which can produce a significance knowledge to biologist.

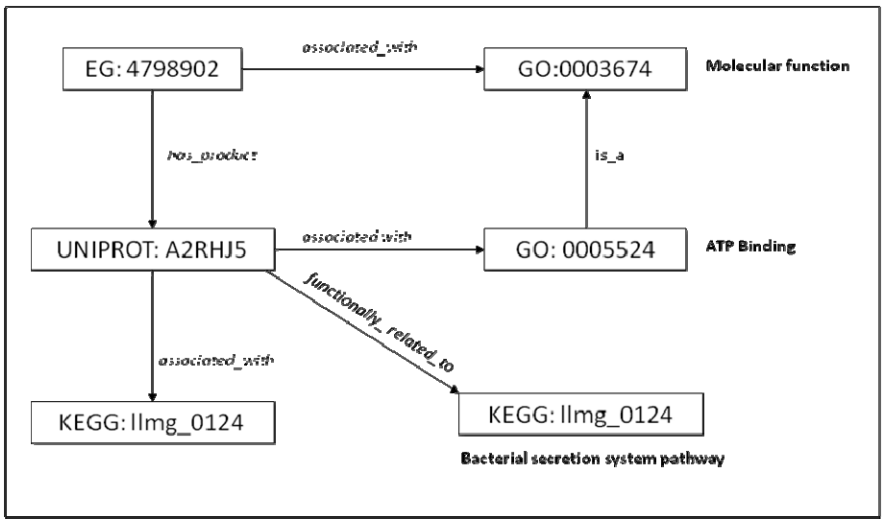

Fig. 2: RDF graph and instantiated

6. Conclusion and future work

There are numerous growing body of research on semantic data integration in the life science including biomedical and bioinformatics area. Generally, most efforts highlight the general use case and perform integration process on very large data sources instead of concerning on specific organism. In this paper, we are focussing the integration data source on one organism, *L. lactis*. Thus, it precisely covered the purpose of integrated data needed by biologist for example, what type of data to integrate, how to representing the data in semantic web and what methodology used to semantically connecting heterogeneous resources. From this experiment, we can conclude that integrating data sources are really depend on specific research area and its

purpose. The approach in this paper generally is very suitable for the first phase of system biology research where biologist need to study new organism that can lead a new knowledge discovery and then hypothesis can be formulated efficiently.

Our approach is flexible and generally can be used to other interested organisms. There are couple of scenarios that this approach can consider to be used for particular purpose. For instances, NCBI EG data doesn't contains related information such as GO annotation; researchers intended to build own local data on specific organisms; no existing publicity available of integrated data source that meet biologist requirements; suitable for collaboration research environment and new completing sequence organism further study. We want to extend this work by modelling more detailed RDF data from EG considering all information, which is not only concern on protein and pathway relation. We also want to investigate more complex biological query using SPARQL to explore potentially new 'connecting dots' which can be inferred from existing resources that can contribute more significance knowledge in biological area.

7. Acknowledgement

This work was supported by Malaysian Genome Institute (MGI) and Ministry of Science, Technology and Innovation for *In Silico Cell Factory* project through Research Management Centre, UTM. Vot 73744.

8. References

[1] Gene Ontology Annotation EBI [cited 11/20/11; Available from:http://www.ebi.ac.uk/GOA/].

[2] GOA2RDF: Generates a simple RDF graph from a given GOA file [cited 11/20/11; Available from:http://search.cpan.org/dist/ONTO-PERL/scripts/goa2rdf.pl].

[3] NCBI Entrez Gene, [cited 11/20/11; Available from: http://www.ncbi.nlm.nih.gov/gene].

[4] SESAME OpenRDF Framework, [cited 11/20/11; Avaialable from: http://www.openrdf.org/].

[5] Antezana, E., W. Blonde, et al. (2009). "*BioGateway: a semantic systems biology tool for the life sciences*." Bmc Bioinformatics **10**.

[6] Berners-Lee, T., J. Hendler, et al. (2001). "*The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.*" Scientific American **284**(5): 34-+.

[7] Harris, M. A., J. I. Clark, et al. (2006). "*The Gene Ontology (GO) project in 2006*." Nucleic Acids Research **34**: D322-D326

[8] Kanehisa, M. and S. Goto (2000). "*KEGG: Kyoto Encyclopedia of Genes and Genomes*." Nucleic Acids Research **28**(1): 27-30.

[9] Karp, P. D. (2005). "*BioCyc pathway database collection and the pathway tools software*." Abstracts of Papers of the American Chemical Society **229**: U1178-U1178.

[10] Kitano, H. (2002). "*Systems biology: A brief overview*." Science **295**(5560): 1662-1664.

[11] Lee, T. J., Y. Pouliot, et al. (2006). "*BioWarehouse: a bioinformatics database warehouse toolkit*." Bmc Bioinformatics **7**: 170.

[12] Sahoo, S. S., O. Bodenreider, et al. (2008). "*An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence*." Journal of Biomedical Informatics **41**(5): 752-765.

[13] Wegmann, U., M. O'Connell-Motherwy, et al. (2007). "*Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp cremoris MG1363*." Journal of Bacteriology **189**(8): 3256-3270.

[14] KH Cheung, AK Smith, KYL Yip, CJO Baker, MB Gerstein (2007). in Semantic Web: *Revolutionizing Knowledge Discovery in the Life Sciences* (eds. C Baker and K Cheung, Springer, NY), pp. 11-30