BION2SEL: An Ontology-based Approach for the Selection of Molecular Biology Databases

Daniel Lichtnow[1], Ronnie Alves[2,5,6], Oscar Pastor[3], Verónica Burriel[3], and José Palazzo Moreira de Oliveira[4]

[1]Universidade Federal de Santa Maria - UFSM, Santa Maria, RS, Brazil
dlichtnow@politecnico.ufsm.br
[2]Institut de Biologie Computationnelle, Montpellier, France
alvesrco@gmail.com
[3]Universitat Politècnica de València, Valencia, Spain
vburriel@pros.upv.es, opastor@pros.upv.es
[4]Universidade Federal do Rio Grande do Sul - UFRGS, RS, Porto Alegre, Brazil
palazzo@inf.ufrgs.br
[5]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, Université Montpellier 2, Centre National de la Recherche Scientifique, Montpellier, France
[6]PPGCC - Universidade Federal do Pará, Belém, Brazil

Abstract. The catalogs of molecular biology databases does not provide a full description of databases, so the user should select databases using limited information available. Taking into account this fact, in the context of a initiative is called *BioDBCore*, a group of experts proposes core metadata definitions to describe the molecular biology databases. However, how to use these metadata to infer the quality of a database is a clear open issue. In the present work, we propose an ontology-based approach aiming to guide the database selection process from molecular biology database catalogs using these metadata.

Keywords: molecular biology databases, database catalog, data quality, ontology, database selection

1 Introduction

There have been some initiatives to create online molecular biology database collections. Currently the database selection processes from these catalogs is an iterative process, where the users must evaluate manually a set of candidate databases using limited query mechanisms and incomplete information about databases. In the present work, we propose an ontology-based approach to the selection of database catalogs. We start showing examples of molecular biology database catalogs (Section 2). After, an overall perspective of an approach that can help a user in the selection process is presented (Section 3). A preliminary version of the proposed ontology and its data quality rules are presented (Section 4) and an application scenario illustrating the use of the ontology is discussed (Section 5). Final remarks and future works concludes this paper (Section 6).

2 Related Work

The *Nucleic Acids Research Database Collection* is the most well-known molecular biology database catalog. This catalog is published since 1996 and its size has been increasing constantly (in 2001, this catalog included only 96 databases and in 2014 it includes 1,552 databases [1]). In general, the molecular biology database catalogs offer limited information regarding the databases and the querying capabilities. It is not possible for a user to indicate any additional requirement, and there is not any information about the quality of the databases (e.g. reputation, accuracy, etc.). Thus, the database selection processes from these database catalogs use to be an iterative process where, generally, users must carefully evaluate the set of candidate databases using only categories which have been already provided by these catalogs. For these reasons, there are several initiatives to create richer database catalogs and the *BioDBCore* consortium proposed a set of metadata elements for the better description of molecular biology databases [2]. Despite these initiatives, to the best of our knowledge, a little investigation has addressed the use of the database metadata (specially *BioDBCore* attributes) on the identification of the best databases.

3 The Proposed Approach

The background of our approach is an ontology for the evaluation of the overall quality of a molecular biology database. In the approach, a user indicates some quality requirements, and databases of a catalog are evaluated using a set of rules defined in the ontology. The result of the evaluation consists of a set of databases classified according to distinct quality levels (*High*, *Medium*, *Low*) of each data quality dimension (e.g. accuracy, believability, etc.) considered into the evaluation.

The overall evaluation of a database is carried out using metadata defined in the *BioDBCore* initiative that is represented in the proposed ontology. The utilization of the metadata in the data quality evaluation process has been considered in some works where metadata models are matched to data quality model requirements [3]. In order to identify potential quality dimensions to the proposed approach, we started by exploring quality dimensions defined by Wang and Strong, due to the relevance of this work [4] in the data quality community. Next, we adapted these quality dimensions following a process similar to Naumann et al. [5] who selected the quality dimensions taking into account the data integration process of molecular biology information systems. In the quality dimensions selection process the metadata defined in *BioDBCore*, quality dimensions, and quality indicators evaluated on previous works [6, 7] are also considered.

The Table 1 shows the quality dimensions defined and the metadata element proposed by *BioDBCore*. Because the lack of space, we do not present, quality dimension definitions (see [4]). In Table 1, the metadata element *Scope:*

Element	Quality Dimension
Database name	None
Main resource URL	Believability
Contact information (e-mail; postal mail)	Ease of understanding
Date resource established (year)	Timeliness
Conditions of use (free, or type of license)	License
Scope: Data types	Relevancy
Scope: Curation policy	Accuracy
Scope: Standards Used	Verifiability Completeness
Standards: MIs	Verifiability Completeness
Standards: Data formats	Interpretability
Standards: Terminologies	Representational Consistency
Taxonomic coverage	Relevancy
Data accessibility/output options	Accessibility
Data release frequency	Timeliness
Versioning period and access to historical files	Timeliness
Documentation available	Ease of understanding
User support options	Ease of understanding
Data submission policy	Accuracy
Relevant publications	Believability
Resource's Wikipedia URL	Ease of understanding
Tools available	Accessibility

Table 1. *BioDBCore* metadata elements and quality dimensions

Data type refers to the content of a database (e.g. a database can store microarray experiments, sequence references, protein structure, etc). The metadata elements *Standards: MIs (Minimum Information)*, *Standards: Data formats* and *Standards: Terminologies* refer to standards present in a catalog of standards - *BioSharing*. One example of a standard is *MIAME - Minimum Information About a Microarray Experiment* - whose main goal is to specify all the data necessary to interpret a microarray experiment. Some quality indicators can be derived from some metadata elements in Table 1 (e.g. the element *Main resource URL* can be used to obtain the *PageRank* of the databases homepage and the number of web links pointing to the database's homepage).

Some of these metadata elements were tested in the evaluation process of molecular biology databases in a previous work [6]. In this previous work, a comparation was done between the rankings generated using these indicators and rankings which had been manually generated by three experts using Spearman correlation (all correlations are significant for $p < 0.05$). However, the results also indicate that it is important to take into account aspects related to more specific users' needs. In this sense, the database categories present in the *Nucleic Acid Research* catalog are not enough for guiding the selection process of databases.

This particular observation points out an interesting opportunity for developing an ontology-based approach for the selection problem.

4 An Ontology-based Approach to Guide the Selection of the Best Molecular Biology Databases

Taking into account the quality dimensions, the metadata elements proposed in *BioDBCore* and our previous works, a prototype of an ontology-based approach to guide the selection of the best molecular biology databases, called in short *BION2SEL*, has been defined. The *BION2SEL* is logically separated into a set of distinct components (ontologies). The two main components are (1) The *Molecular Biology Database* ontology where database features are described using the metadata elements proposed in *BioDBCore* and (2) The *Molecular Biology Database Quality* ontology where the quality dimensions are defined.

The *Molecular Biology Database* ontology describes the properties of molecular biology databases. Each instance of this ontology is a molecular biology database. The properties are basically derived from the metadata elements present in *BioDBCore* (see Table 1). Some values assigned to the properties of this ontology are vocabularies defined in other ontologies. Thus, there are ontologies that provide vocabularies (e.g. population class, data format). For example, the range of *Molecular Biology Database* ontology property *data types* refers to classes of *Molecular Biology Summary Data*. The *Molecular Biology Summary Data* ontology was created by us and can be considered as a summarized global schema of all databases present in a database catalog. In the *Molecular Biology Summary Data* ontology, the relationships between classes are also defined (e.g. *Protein hasStructure ProteinConformation*). Using the *Molecular Biology Summary Data* ontology classes, the *OMIM* database is assigned to classes *Gene*, *Allele*, *Allele Variant* and *Disease* classes and the *ALFRED* database is assigned to classes *Gene*, *Allele*, *Allele Variant* and *Population*. Thus, the user should indicate the biological entities (e.g. gene, pathways, etc.) defined in the ontology (this process is similar to adopted in [8]).

The *Molecular Biology Database Quality* ontology defines the quality criteria to evaluate a molecular biology database according to distinct quality dimensions. For each quality dimension, three quality levels are defined in the ontology: *High*, *Medium* and *Low*. A database, an instance of *Molecular Biology Database*, is classified into a quality level in accordance to the quality indicators corresponding to metadata elements (properties). The conditions to classify a database into a specific quality level are defined through *SWRL - Semantic Web Rule Language (SWRL)*, a rule language based on OWL and *SQWRL - Semantic Query-Enhanced Web Rule Language* [9]. Next, examples of rules defined for some quality dimensions are shown. Although the experimental evaluation validates the usefulness of some quality indicators and rules, they must be taken as preliminary heuristics whose the main purpose is demonstrate the utilization of the ontology within the quality evaluation process.

The relevancy ("the extent to which data are applicable and helpful for the task at hand" [4]) is the quality dimension which has been initially considered in the database selection process. As the goal is to select databases from a catalog without access its content, we define relevancy as the percentage of classes of objects (e.g. gene, protein, etc.) required by the user rather than the information stored on a database. Thus, in the proposed ontology, the databases with all required data are classified as having a *High_Relevancy*, databases with 50% or more of required data are classified as a *Medium_Relevancy*, and databases with less than 50% of required data are considered having *Low_Relevancy*. In (1) a rule defined with *SWRL* and *SQWRL* to classify a database as a *High_Relevancy* database is shown.

Basically, the rule defines that a database - *Molecular_Biology_Database(?x)* - must have required data. The metadata element *data_types* contains the names of the classes of *Molecular Biology Summary Data* ontology present in a database. Thus using *SQWRL* two sets are compared: one set contains the classes of *Molecular Biology Summary Data* required by a user *(?ds)*, and another set contains the classes of *Molecular Biology Summary Data* that a database has *(?s)*.

$$
\begin{aligned}
& Molecular\_Biology\_Database(?x) \wedge \\
& data\_types(?x, ?dt) \wedge \\
& sqwrl : makeSet(?s, ?dt) \wedge \\
& sqwrl : groupBy(?s, ?x) \wedge \\
& sqwrl : makeSet(?ds, Gene) \wedge \\
& sqwrl : makeSet(?ds, Disease) \wedge \\
& sqwrl : difference(?dif, ?ds, ?s) \wedge \\
& sqwrl : size(?sdif, ?dif) \wedge \\
& swrlb : equal(?sdif, 0) \rightarrow sqwrl : select(?x)
\end{aligned}
\tag{1}
$$

The believability ("The extent to which data are accepted or regarded as true, real and credible" [4]) is related to the content creator and the explicit or implicit users' ratings [3]. Implicit users ratings can be the number of references in the web to a particular database. Some previous experiments have demonstrated that the number of paper citations related to databases are interesting quality indicators [7]. Thus, rules based on the number of citations of these papers to classify a database as a database with high, medium, and low believability are defined. At the present moment, the defined rules take into account the median of the number of citations. Thus, databases where papers are cited higher than the median are considered databases with *High_Believability*, databases where the number of citations is equal to the median are considered databases with *Medium_Believability*, and databases where the number of citations is lower than the median are considered databases with *Low_Believability*.

Regarding to accessibility ("The extents to which data are available or easily and quickly retrievable" [4]), one possibility is to measure the accessibility degree of a database considering all access mechanisms available. Thus, we defined that databases are considered to have *High_Accessibility*, if they allow to inform an specific argument for a query and returns an specific data.

5 Ranking Candidate Gene Markers in Alzheimer: An Ontology-based Quality Model Experience

In this section, some aspects of the utilization of the *BION2SEL* to guide researchers who needs to retrieve data from a set of molecular biology databases are illustrated. The quality dimensions considered here are accessibility and relevancy. Believability is not considered because, in the example, we are considering a small number of databases (if two or more databases have the same relevancy degree, databases with higher believability must be selected first. The case of use consists of identifying the most differentially expressed genes (*DEG*) in brains affected by Alzheimer's disease (documented in [10]). In the present case of use, the *DEG* task in brains affected by Alzheimer's disease can be subdivided in a set of subtasks:

1. Obtaining transcriptomic (*Microarrays*) data related to Alzheimer;
   (a) Selecting a set of transcriptomic studies;
   (b) Selecting the microarrays experiments related to hippocampus;
2. Conducting the differential expression analysis i.e. identifying *DEG*;
   (a) Ranking probes according to a two-group (normal *vs.* affected tissue)
   (b) Mapping probes to genes. This is an annotation process (genes names are assigned to probes);
3. Exploring functional enrichment analysis using the *DEG* set;
4. Identifying and discarding genes whose codified proteins are not known;
5. Verifying whether the selected candidate gene set is related to the Alzheimer.

The database selection process starts by exploring aspects related to relevancy. After other quality dimensions can be evaluated (e.g. believability). Thus, for each task, when it is necessary, the user indicates the required data, i.e. the biological entities of *Molecular Biology Summary Data*. Table 2 presents a set of databases with a list of biological entities of *Molecular Biology Summary Data* present in each database (this list is not exhaustive).

The requirement related to task 1 can be expressed by a query like "retrieve all gene expression studies related to Alzheimer". Thus, this query must be executed in a database which contains data about gene expression studies and diseases. In this sense, the relevant classes in the *Molecular Biology Summary Data* are *Assay* and *Disease*. Thus, the most relevant database is *ArrayExpress Archive*. The *ArrayExpress Archive* has an access mechanism (an *HTML form*) which allows to retrieve gene expression experiments using as argument the name of a disease.

The requirement related to task 3 can be expressed by a query like "returns the function of a set of genes". Thus, the query must be executed in a database that contains data about gene function. The most appropriate database is *Gene Ontology*. The next step verifies which identified genes code to known proteins (task 4). In this case, the user need can be expressed by "identifying which genes are assigned to known proteins". This query must be executed in a database containing data about genes and proteins. The *Pfam* (see Table 2) is the database

Database	URL	Content
Array Express Archive	http://www.ebi.ac.uk/arrayexpress	Gene Disease Assay Microarray
Pfam database	http://pfam.sanger.ac.uk	Gene Protein Protein Function Protein Conformation
Gene Ontology	http://www.geneontology.org/	Gene Gene Function
BioMart	http://www.biomart.org	Gene Gene Function Protein Protein Function Disease
Phenopedia	http://www.hugenavigator.net	Gene Disease
OMIM	http://www.ncbi.nlm.nih.gov/omim	Gene Disease

Table 2. Examples of molecular biology databases

that provides the support for this query. Thus, by *R* programming, 45 genes are selected. Finally, each one of these 45 genes is verified with respect to its association to the Alzheimers disease. The biological entities which are relevant to this query are *Gene* and *Disease* (Task 5). Take into account only relevancy quality dimension, four databases provides proper support: *OMIM*, *Phenopedia*, *BioMart* and *ArrayExpress Archive*. The *OMIM* does not have an access mechanism, where the user informs the name of a disease and receives genes related to it. The same situation is observed for the *BioMart*. The *ArrayExpress* database returns raw Assays. The *Phenopedia* provides several meta-analysis studies in which genes are ranked properly according to a disease phenotype, i.e. Alzheimers disease. Therefore, the *Phenopedia* is the best database for this particular task. This last task demonstrates that a richer description of access mechanism of databases can help users in the database selection process.

6 Final remarks

In the present work, the *BioDBCore* metadata has been analyzed and a prototype of an ontology has been defined, where some quality rules are defined for different quality dimensions using database metadata. The aim is to facilitate the searching process in a database catalog. In the future, we intent to improve this initial ontology. In this sense, the *Molecular Biology Summary Data* ontology, for example, has been defined to illustrate the approach, but it is necessary to improve this ontology and to verify, by experiments, the usefulness of it.

Thinking in a scenario where *BioDBCore* data will be available in a near future (nowadays some of these data are available [1]), it will be possible to implement a real application using the proposed ontology.

7 Acknowledgments

This work is partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil. The work of V. Burriel, and O. Pastor has been developed with the support of MICINN and GVA under the projects PROS-Req TIN2010-19130-C02-02 and ORCA PROMETEO/2009/015, and co-financed with ERDF and Cátedra Tecnologías para la Salud of Universitat Politècnica de València financed by INDRA Systems.

References

1. Galperin, M.Y., Rigden, D.J., Fernndez-Surez, X.M.: The 2014 nucleic acids research database issue and an updated nar online molecular biology database collection. Nucleic Acids Research **42**(D1) (2014) D1–D6
2. Gaudet, P., Bairoch, A., Field, D., Sansone, S., Taylor, C., Attwood, T., Bateman, A., Blake, J., Bult, C., Cherry, J., et al.: Towards BioDBcore: a community-defined information specification for biological databases. Nucleic acids research **39**(suppl 1) (2011) D7
3. Naumann, F., Rolker, C.: Do metadata models meet iq requirements. In: Proceedings of the International Conference on Information Quality (IQ). (1999) 99–114
4. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. **12**(4) (1996) 5–33
5. Naumann, F., Leser, U., Freytag, J.C.: Quality-driven integration of heterogenous information systems. In: VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1999) 447–458
6. Lichtnow, D., Levin, A., Alves, R., Castello, I.M., Dopazo, J., Pastor, O., de Oliveira, J.P.M.: Using metadata and web metrics to create a ranking of genomic databases. In White, B., Isaas, P., Santoro, F.M., eds.: WWW/Internet 2011 conference, IADIS, IADIS Press (2011) 253–260
7. Lichtnow, D., Alves, R., Levin, A., Castello, I.M., Dopazo, J., Pastor, O., de Oliveira, J.P.M.: Using papers citations for selecting the best genomic databases. In: SCCC 2011 conference. (2011) 1–10
8. Cohen-Boulakia, S., Lair, S., Stransky, N., Graziani, S., Radvanyi, F., Barillot, E., Froidevaux, C.: Selecting biomedical data sources according to user preferences. Bioinformatics **20**(1) (2004) 86–93
9. Oconnor, M., Knublauch, H., Tu, S., Grosof, B., Dean, M., Grosso, W., Musen, M.: Supporting rule system interoperability on the semantic web with swrl. The Semantic Web–ISWC 2005 (2005) 974–986
10. Martello, C.L., Alves, R.: Explorando marcadores gênicos em estudos transcriptomicos relacionados ao alzheimer. Technical report, Instituto de Biociências - UFRGS (2011)

[1] http://www.biosharing.org/biodbcore