Pay-as-you-go data integration for bio-informatics (PAYDIBI[1])

SRO NICE project proposal

M. van Keulen (UT/DB), P. van der Vet (UT/HMI)

1 Introduction

Scientific research in bio-informatics is often data-driven and supported by *biological databases*. A biological database contains factual information collected from scientific experiments and computational analyses about areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.[2] Much effort is involved in keeping them up-to-date, extending them and in creating new ones (see for example [1], [2]).

In a growing number of research projects, researchers like to ask combined questions, i.e., questions that require the combination of information from more than one database. [3] estimates that in scientific workflows "in total 30% of all the tasks are data transformation tasks". Combining information from several biological databases can be a painstaking process, because of many reasons such as

- Databases vary in structure, size, and quality. Some contain broad information on a few objects, others contain narrow information on many objects.
- The data largely originates from scientific experiments, hence errors and impreciseness are inevitable, see for example [4].
- Databases often partially overlap. Conflicts and ambiguities may exist concerning the overlapping data.
- The absolute truth is often unknown and is precisely what is researched. Errors, impreciseness, (evolving) opinions and consensus are inevitable.
- There exist many valuable usually smaller specialist databases. The 2010 on-line database collection of the Nucleaic Acids Research journal (NAR) contains "1230 data resources, a growth of 5% over last year" [5]. A scientist may not know how much trust (s)he can place in them.
- One typically does not hold trust in a database, but in committees, groups, and individual persons.
- Biological databases are kept up-to-date for as long as there is funding, hence some required sources may not be 'active' anymore.

---

[1] Pronounced as "pay-dee-bee"

[2] Source Wikipedia: http://en.wikipedia.org/wiki/Biological_database

As a consequence, much effort is necessarily devoted to low-level data integration tasks significantly slowing down the process of scientific discovery (see also [6]).

2 Objective and approach

The objective of this project is to develop data integration technology to support the scientist in the construction of targeted data sets from multiple biological databases and other data sources according to his/her views, opinions, and trust.

The approach will be based on recent ideas for "pay-as-you-go" [7] and "good-is-good-enough" data integration [8] to allow scientists to quickly construct a targeted data set that can be meaningfully used. In practice, only through the use of data does more understanding of that data arise and are irregularities found. We intend to adapt and extend the pay-as-you-go ideas in such a way that this natural process is effectively supported and that all data manipulation as a result of it is properly recorded (data provenance).

As opposed to early pay-as-you-go approaches, we do not intend to replicate data, but to construct a *knowledge base* with data mapping rules, decisions, annotations, assessments (both human assessments and results of statistical analyses), trust information, and other evidence. The knowledge base should be able to contain information about conflicts, ambiguities, missing information, wrong information, and other kinds of information irregularities, as well as knowledge on how to resolve these. A combined question is answered by interpreting and adapting the combined data in the biological databases according to the knowledge in the knowledge base. Since the information about resolving irregularities may not be complete, we apply the ideas of uncertain databases [9] to store and query remaining irregularities as a form of semantic uncertainty.

The construction and evolution of such a knowledge base is usually a team effort, i.e., the knowledge base is a shared resource where people trust each other's input, while at the same time scientists may disagree with each other. We allow a scientist to define trust and disagreement in an effective way, and to browse, query, and compare the information from different viewpoints (e.g., according to his/her personal views and opinions, or according to the generally accepted state of affairs). Corrections or changes to the contents of the knowledge base that are not supported by all users is seen as additional evidence belonging to the scientist's own viewpoint.

In this way, scientists can quickly develop targeted integrated data sets that properly reflect the continuous flow of updates to the sources, adapt and clean them according to their own views and opinions, while making full use of the trusted work of other scientists in a non-interfering manner.

3 Domain and Use Case

We focus on the bio-informatics domain. We established a relationship with Prof.dr. J.A.M. Leunissen, head of the bio-informatics group of Wageningen University (see http://www.bif.wur.nl/UK/). This group supplies the use case and base data concerning the construction of a targeted integrated ~omics data

set on *Brassica* species for the MADMAX project (see below). The use case provides the context and data for validation of the to be developed technology.

> *(excerpts from the MADMAX project description)*
> Rapid increase of "~omics"datasets generated by microarray (transcriptomics), mass spectrometry (proteomics and metabolomics) and next generation sequencing (genomics) for non-model organisms require an integrated platform to combine all related studies under certain criteria (same tissues, same species, similar treatments, same pathways, etc.). In this way, biological hypotheses can be tested by using as many as possible/available different ~omics datasets, and hence a better understanding of the questions can be gained in more accurate way. Such a hypothesis-driven analysis platform can broaden the scale of answered questions easily with additional designed experimental validation, giving genetically tractable organisms (in particular crops) higher utility.
>
> Besides rapidly available genetic information on *Brassica* species, current transcriptomics, proteomics and metabolomics data are useful to unravel genes that involved in plant developmental and nutrition related traits. However, a specialized *Brassica* web-based database / service to interpret and analyse all these sequence and "~omics" data is lacking.
>
> Through the integration of ~omics datasets one can produce different views on biological questions, and gain insight in the underlying biological processes by interrogating the data from different perspectives. For example, the integration of genomic and transcriptomic data in different species can address fundamental evolutionary questions. Genomic and metabolomic data combination can result in discovery of previously uncharacterized metabolic reactions, both reactants and enzymes. Transcriptomic and metabolic data can be utilized to assess the influence of exogenous pathways/genes. To establish a platform able to list all available information from these views related to gene of interest and further evaluation for current gene functional annotation.

4 Scientific challenges

The research project focus lies in the following main scientific challenges:

- How to adapt and extend the ideas for pay-as-you-go and good-is-good-enough data integration to the bio-informatics domain.
- How to define a method and tool-support for the quick construction and enrichment of a targeted data set from a set of external data sources.
- How to define a method and tool-support for effective continuous improvement of the data quality.

The intended approach means that the following secondary challenges need to be addressed within the limited scope of the project domain:

- How to combine knowledge rules, evidence and data to represent and query information according to the desired viewpoint.
- How to model data mapping and data cleaning as knowledge rules.
- How to model information on trust, assessments, decisions, etc. as evidence.

- How to define a method for effective trust management.
- How to define a method for opinion and consensus management.
- How to scale to data volumes of biological databases and their numbers of users.

5 Consortium and expertise

M. van Keulen (UT/DB): Database interoperability.

P. van der Vet (UT/HMI): Knowledge representation; "liaison" between computer science and bio-informatics.

J. Leunissen (WUR/BI): bio-informatics; provider of validation use case and data.

6 Request

The consortium asks CTIT SRO NICE for the financing of one PhD student for this project that is intended to strengthen the cooperation between the DB and HMI groups. The involvement of the WUR/BI-group is to support the PhD student with data and use cases.

7 Bibliography

1. Ott, M., Vriend, G.: Correcting ligands, metabolites, and pathways. BMC Bioinformatics 7(517) (November 2006)
2. Spronk, C., Nabuurs, S., Krieger, E., Vriend, G., Vuister, G.: Validation of protein structures derived by NMR spectroscopy. Progress in Magnetic Resonance Spectroscopy 45(3-4), 315-337 (2004)
3. Wassink, I.: Work flows in life science. PhD thesis. University of Twente, Enschede, The Netherlands. CTIT Ph.-D-thesis series No. 09-157. ISBN 978-90-365-2932-7 (2010)
4. Joosten, R., Vriend, G.: PDB Improvement Starts with Data Deposition. Science 317(5835), 195-196 (July 2007)
5. Database issue. Nucleic Acids Research 38(1) (January 2010)
6. Stein, L.: Creating a bioinformatics nation. Nature 417(6885), 119-120 (May 2002)
7. Halevy, A., Franklin, M., Maier, D.: Principles of dataspace systems. Proceedings of the Twenty-Fifth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Chicago, Illinois, USA, 1-9 (June 2006)
8. Van Keulen, M., De Keijzer, A.: Qualitative effects of knowledge rules and user feedback in probabilistic data integration. The VLDB Journal 18(5), 1191-1217 (2009)
9. Huang, J., Antova, L., Koch, C., Olteanu, D.: MayBMS: a probabilistic database management system. Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, 1071-1074 (June 2009)