Journal of Biomedical Informatics 43 (2010) 752–761

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

OpenFlyData: An exemplar data web integrating gene expression data on the fruit fly *Drosophila melanogaster*

Alistair Miles [a], Jun Zhao [a], Graham Klyne [a], Helen White-Cooper [b], David Shotton [a,*]

[a] Image Bioinformatics Research Group, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK
[b] School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK

ARTICLE INFO

Article history:
Received 12 November 2009

Keywords:
Chado
Data integration
Data web
Drosophila
Gene expression
Performance
RDF
SPARQL
Triple store
User interface

ABSTRACT

Motivation: Integrating heterogeneous data across distributed sources is a major requirement for *in silico* bioinformatics supporting translational research. For example, genome-scale data on patterns of gene expression in the fruit fly *Drosophila melanogaster* are widely used in functional genomic studies in many organisms to inform candidate gene selection and validate experimental results. However, current data integration solutions tend to be heavy weight, and require significant initial and ongoing investment of effort. Development of a common Web-based data integration infrastructure (a.k.a. data web), using Semantic Web standards, promises to alleviate these difficulties, but little is known about the feasibility, costs, risks or practical means of migrating to such an infrastructure. *Results:* We describe the development of OpenFlyData, a proof-of-concept system integrating gene expression data on *D. melanogaster,* combining Semantic Web standards with light-weight approaches to Web programming based on Web 2.0 design patterns. To support researchers designing and validating functional genomic studies, OpenFlyData includes user-facing search applications providing intuitive access to and comparison of gene expression data from FlyAtlas, the BDGP *in situ* database, and FlyTED, using data from FlyBase to expand and disambiguate gene names. OpenFlyData's services are also openly accessible, and are available for reuse by other bioinformaticians and application developers. Semi-automated methods and tools were developed to support labour- and knowledge-intensive tasks involved in deploying SPARQL services. These include methods for generating ontologies and relational-to-RDF mappings for relational databases, which we illustrate using the FlyBase Chado database schema; and methods for mapping gene identifiers between databases. The advantages of using Semantic Web standards for biomedical data integration are discussed, as are open issues. In particular, although the performance of open source SPARQL implementations is sufficient to query gene expression data directly from user-facing applications such as Web-based data fusions (a.k.a. mashups), we found open SPARQL endpoints to be vulnerable to denial-of-service-type problems, which must be mitigated to ensure reliability of services based on this standard. These results are relevant to data integration activities in translational bioinformatics. Availability: The gene expression search applications and SPARQL endpoints developed for OpenFlyData are deployed at http://openflydata.org. FlyUI, a library of JavaScript widgets providing re-usable user-interface components for *Drosophila* gene expression data, is available at http://flyui.googlecode.com. Software and ontologies to support transformation of data from FlyBase, FlyAtlas, BDGP and FlyTED to RDF are available at http://openflydata.googlecode.com. SPARQLite, an implementation of the SPARQL protocol, is available at http://sparqlite.googlecode.com. All software is provided under the GPL version 3 open source license.

1. Introduction

1.1. Drosophila gene expression data

Genetic insights from model organisms such as the fruit fly (*Drosophila* spp.) have translated into benefits for human health [1,2]. However, fundamental questions remain, and work is ongoing to characterize the function of every gene in the sequenced *Drosophila* genomes. For *Drosophila melanogaster*, genome-scale data on patterns of gene expression in time and space are publicly available. FlyAtlas,[1] for example, publishes quantitative data on tissue-specific mRNA expression levels for 26 different tissues and approximately 18,770 genes [3]. The Berkeley Drosophila Genome Project (BDGP)[2] publishes image data from *in situ* mRNA

* Corresponding author. Fax: +44 (0)1865 310447.
E-mail address: david.shotton@zoo.ox.ac.uk (D. Shotton).

[1] http://flyatlas.org/.
[2] http://fruitfly.org/cgi-bin/ex/insitu.pl.

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.04.004

hybridization experiments and DNA time-course microarray assays at different stages of embryogenesis covering 6138 genes [4,5]. FlyTED, the *Drosophila* Testis Gene Expression Database,[3] provides similar data for the adult *Drosophila* testis, currently containing 2762 mRNA *in situ* hybridization images and ancillary data revealing the patterns of gene expression of 817 genes in testes of wild type flies and seven meiotic arrest mutant strains in which spermatogenesis is defective [6]. These data are regularly consulted by research groups focused on specific aspects of organism biology, such as male infertility, to inform decisions about resource allocation and experimental design, particularly which genes or mutations to study in detail. These data are also used to validate personal experimental results, where a discrepancy, perhaps due to sample contamination, typically indicates that an assay needs to be repeated.

1.2. The need for Drosophila data integration

For any one investigation, data from several different sources, on thousands of genes, must be reviewed and compared with locally derived results. However, data repositories exhibit varying degrees of syntactic and semantic incompatibility, making it difficult to find related information scattered across several databases without searching each source individually. This inability to bring together relevant data quickly and interpret them accurately can impact a researcher's productivity and success.

Data from different sources and derived by different experimental methods reveal complimentary aspects of the underlying biology. For example, *in situ* data from BDGP show transcription of the gene *schuy* during late embryogenesis to be clearly localized in the gonad. Microarray data from FlyAtlas show *schuy* strongly transcribed in the adult testis, but not in any other tissue. In this case, the transcriptional profile established in embryogenesis appears similar to that of the adult, and the combined data strongly suggest a role for *schuy* in sperm development. These data influenced the choice of *schuy* as a candidate gene for testis *in situ* studies relating to *Drosophila* male infertility, leading to the discovery of post-meiotic transcription (previously thought not to occur in *Drosophila* spermatogenesis) and of two new classes of sub-cellularly localized transcripts dubbed "comets" and "cups" [7].

The publication in the Web, without access restriction, of these and other gene expression data, such as [8], Fly-FISH[4] [9], gene expression annotations curated by FlyBase[5] [10], and microarray studies available via NCBI GEO[6] or ArrayExpress,[7] has been highly beneficial for *Drosophila* functional genomics. Nevertheless, these data are published at separate locations and use dissimilar access methods and user interfaces, such that there is no easy way to ask questions that span the data sources.

1.3. Related work

There have been notable attempts to provide an integrated view of gene expression data for *D. melanogaster*. FlyMine[8] is an instance of the generic data warehouse platform InterMine customised for data on *Drosophila*, *Anopheles* and *Caenorhabditis* spp. [11]. FlyMine currently stores copies of embryo *in situ* data from BDGP and Fly-FISH, and of microarray data from FlyAtlas and from [8]. While it is possible to use FlyMine to view anatomy ontology annotations from BDGP side-by-side with FlyAtlas microarray data for one or more genes, this involves constructing a new query template using the generic query builder interface. The user interface is powerful, because it can be used to construct arbitrary queries over the FlyMine data model, but it can be hard to understand without an informatics background and requires time to master. Additionally, FlyMine does not provide any way to view the *in situ* images from BDGP, only their annotations. The images themselves provide more detail than the 145 anatomy ontology terms used by BDGP to annotated them [5], and are important sources of information for decision making and validation.

Approaches to biomedical data integration are reviewed by Goble and Stevens [12] and Stein [13]. These range from distribution of SQL queries over relational databases, for example the OGSA-DAI Project,[9] to data warehousing activities such as FlyMine described above. Many authors now favour Web standards, but tend to emphasize either a *service oriented* or a *data-oriented* perspective. The service-oriented perspective focuses on standards for Web service description, invocation and coordination, to enable at least semi-automated assembly and execution of computational workflows, e.g. [14–16]. The data-oriented perspective focuses on the syntax and semantics of data, to enable automated crawling, merging and reasoning over data published in the Web, e.g. [17–21]. These two perspectives on the Web as an infrastructure for sharing biomedical data are influenced by top-down work on Web standards, led by the World Wide Web Consortium (W3C). In particular, the service-oriented perspective is influenced by the Web Services Activity,[10] while the data-oriented perspective is influenced by the Semantic Web Activity. Both are also influenced by bottom-up trends in pragmatic Web developer communities, often grouped under the banner of 'Web 2.0', e.g. [22–24].

These perspectives are converging: ontologies are being used to describe the inputs, outputs and capabilities of Web services, and data from the 'deep Web' of analytical services are being exposed via Semantic Web standards [25–27]. SPARQL sits at the crux of this convergence. Although it will not satisfy all of the life science's bioinformatics requirements, SPARQL provides a standard means of making research data available to systems that need to query and analyse data from multiple sources, including both *in silico* experimental work flows and Web-based mashups. It can remove some of the "shim" software currently needed to cope with syntax and protocol differences between services, and thus provides a higher point of departure from which to tackle the challenging issues of semantic interoperability. It can also reduce the burden on data providers, because its expressiveness means that many queries can be answered 'out of the box', removing the need to design, implement or maintain complicated Web service interfaces. We thus chose to evaluate SPARQL for programmatic access to *D. melanogaster* gene expression data, and to determine whether such an infrastructure could provide the functionality, performance, reliability and scalability to underpin tools for candidate gene selection and experimental data validation in *Drosophila* functional genomics.

1.4. OpenFlyData, an exemplar data web for Drosophila gene expression data

In this paper, we describe the development of OpenFlyData [28], a proof-of-concept data web [29–31] for *D. melanogaster* gene expression data. OpenFlyData uses the Web as a data-sharing platform, employing Semantic Web standards[11] and simple HTTP protocols in the Representational State Transfer (REST) style [32], and is built from loosely coupled Open Source software components. OpenFlyData also makes use of Web 2.0 design patterns to accelerate

[3] http://www.fly-ted.org.
[4] http://fly-fish.ccbr.utoronto.ca/.
[5] http://flybase.org/.
[6] http://www.ncbi.nlm.nih.gov/geo/.
[7] http://www.ebi.ac.uk/microarray-as/ae/.
[8] http://www.flymine.org/.
[9] http://www.ogsadai.org.uk/.
[10] http://www.w3.org/2002/ws/.
[11] http://www.w3.org/2001/sw/.