International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:10, 2009

GeNS: a Biological Data Integration Platform

Joel Arrais, João E. Pereira, João Fernandes and José Luís Oliveira

***Abstract*—**The scientific achievements coming from molecular biology depend greatly on the capability of computational applications to analyze the laboratorial results. A comprehensive analysis of an experiment requires typically the simultaneous study of the obtained dataset with data that is available in several distinct public databases. Nevertheless, developing a centralized access to these distributed databases rises up a set of challenges such as: what is the best integration strategy, how to solve nomenclature clashes, how to solve database overlapping data and how to deal with huge datasets. In this paper we present GeNS, a system that uses a simple and yet innovative approach to address several biological data integration issues. Compared with existing systems, the main advantages of GeNS are related to its maintenance simplicity and to its coverage and scalability, in terms of number of supported databases and data types. To support our claims we present the current use of GeNS in two concrete applications. GeNS currently contains more than 140 million of biological relations and it can be publicly downloaded or remotely access through SOAP web services.

***Keywords*—**Data integration, biological databases

I. Introduction

THE integration of heterogeneous data sources has been a fundamental problem in database research over the last two decades [1-6]. The goal is to achieve better methods to combine data residing at different sources, under different schemas and with different formats in order to provide the user with a unified view of the data. Although simple in principle, due to several constrains, this is a very challenging task where both the academic and the commercial communities have been working and proposing several solutions that span a wide range of fields.

Life sciences are just one of many fields that take advantage from the advances in data integration methods [3, 4, 6]. This is because the information that describes genes, gene products and the biological processes in which they are involved are dispersed over several databases [7]. In addition, due to the advances in some high throughput techniques, such as gene expression, the experimental results obtained in the laboratory only are valuable after being matched with data stored in public databases [8, 9]. Thus, in order to speed up the investigation process, it is very important to have a centralized access to distributed databases.

In this paper, we present GeNS a powerful but easy to use platform that allows the integration of any kind of molecular data. The main advantage of GeNS resides on its schema that has a general organization that supports the addition of new databases and data types without requiring changes in the schema.

University of Aveiro, DETI/IEETA,
3810-193 Aveiro, Portugal

II. Motivation and challenges

According to the last release of the Nucleic Acids Research there are about 1170 databases in the field of molecular biology [7]. Each database corresponds to the output of a specific study or community and represents a huge investment whose potential have not been fully explored.

Being able to integrate data from multiple sources is important for two reasons. First, because data about one biological entity may be dispersed over several databases, for instance, for a gene, the nucleotide sequence is stored in GenBank [10], the pathway in KEGG Pathway [11] and the expression data in ArrayExpress [12]. Obtaining a unified view of this data is therefore crucial to understand the role of the gene. A second reason consists in the fact that many different databases contain redundant or overlapping information [13]. This can be detected by directly comparing databases. Most of the data stored in these databases is publicly available as custom web interfaces, or as text and XML files [14]. To get this data one has to access each database independently, download and parse the files and finally merge all the results in a unified and consistent dataset.

In the last years, several efforts have been made to simplify the process of integrating data from multiple sources. From those we have selected three that seemed the most representative. The first, BioWarehouse [15], contains data from multiple sources including metabolic pathways and enzymes. BioWarehouse uses a database schema oriented to predefined data types, meaning that the addition of new data types implies adding new tables and methods to query them. This database was designed to be more oriented to prokaryotes than for eukaryotes.

A different vision has been applied in BioCoRE [16] that uses a more flexible approach to integrate data. According to the authors the system allows the storage of almost all biochemical process. One drawback is the high complexity of the proposed model that contains more than 200 classes.

A third approach has been applied by Biozon that contains a simple and abstract schema that supports data based on a hierarchical metamodel [17]. Since the schema is general, in Biozon each relation from the metamodel is explicitly stored in the database. As a consequence the current instance contains about 6.5 billion relations, which decrease performance. Biozon is publicly available through an intuitive

and easy to use web interface but is not possible to download the database in order to install a local instance.

The previous databases present different approaches to address the same issue: integrate data from different data sources. The limitations found reflect the difficulty to obtain a simple but comprehensive schema able to accommodate the heterogeneity of the biological domain and maintaining an acceptable level of performance.

III. Database integration approaches

Although it is consensual that the use of biological data spread over the web is essential to extract knowledge from local datasets, it isn't always clear what is the best method to access the data [4]. In this section we review a set of integration techniques organized in three main strategies: Mediators, Links and Warehouses.

In the mediator based integration the data is left on its original database being just created a unified view that is provided to the user. Using this approach, the mediator engine reformulates at run time each requested query into a single or multiple queries that are then submitted to the proper databases. The results are then aggregated and processed creating the final result that is returned to the client. Current examples that use this approach are the BioMediator [18] and the SEMEDA [19].

Link-based integration has been the first and yet the most successful approach to data integration. The reason for the success of this approach is that it resembles very closely to the nature of the web. In the context of molecular biology the problem is that an increasing number of sources on the web require users to manually browse through several web pages and data sources in order to obtain the desired information. In addition, since each database has its own interface the user has to learn how to search and navigate in every single database. Examples of this approach are the Entrez [20] and the DiseaseCard [21] databases.

Finally, warehouse integration consists in physically integrate the data from multiple sources into a local database and executing all the queries directly on this repository rather than on the original ones. In order to use data warehouses, it is also required to develop a unified data model that can accommodate all the information that is stored in various source databases. Additionally, it is necessary to have specific applications to fetch the data from the source databases, transform them to match the local unified scheme and, finally, load them into the data warehouse. After this initial setup phase, the warehouse can be used as a single interface to answer any of the questions that the source databases can handle, as well as those that require the interlink of several concepts that are not present in any single database.

Although each of the discussed approaches has its disadvantages we believe that the warehouse approach is more adequate to address this problem mainly because [6]:

- *Performance:* The warehouse is the only approach where the query response time depends only on local factors (CPU, memory, disks, database) rather than remote servers and network delays. This is especially relevant for complex queries that need to be decomposed into several sub-queries.
- *Access restrictions:* Some data sources do not provide query access to their databases (web services, direct URL, or alike) and others includes specific mechanisms in their interfaces (like session management, cookies, etc.) that hinders the use of remote access approaches.
- *Availability:* Other methodologies cannot assure the quality of service due to the total dependency on external factors, namely the data availability. This can happen because the server is down or because it has simply changed the way to access the data. They are vulnerable to name clashes and ambiguities. If the source database change the way the URL is constructed then the database will become unavailable. These problems are usually only solved through human intervention.
- *Data processing:* Another drawback of others than warehousing approach is the impossibility to manipulate directly the source databases. The smallest data element is a web page, while with the warehouse we can have a great granularity and work with, for instance, a gene name or a gene attribute.
- *Versioning:* Warehouses also allows the user to keep track of the version of each single accessed database, a feature that is not very relevant for small projects but crucial to larger ones.

IV. Implementation

A. Requirements

In order for GeNS be usable, one of the main requirements was that its schema should be easy to understand and maintain. To address this issue, we have focused many of our efforts to achieve a comprehensible schema, with a limited number of tables.

Another requirement was that the system should be scalable in size, in order to contain several gigabytes of data and hundreds of millions of biological entities relations.

The system should also be scalable in terms of the number of databases that it stores. This should be obtained without having any changes in the schema.

Even containing a huge leap of data, the system should be efficient in order to give short response times to the most typical queries. This is especially important because we want this tool to be used to answer user-defined queries and also to be a platform that could be used by other software tools. To attain this requirement, we have stored the gene identifiers and the bio entity entries in separated tables and have optimized the database with the addition of indexes.

The data stored in the database should be accessible through the use of several methods. To achieve this we have implemented a set of web services, which can be used to

query and extract data from the database, in addition to SQL queries.

One last requirement was the possibility to track the current version of the inserted data, as well as the possibility to update the existent data without having to change the entire database.

B. Data integration

To construct GeNS database we have selected the most representative databases that cover a wide span of fields. For each database we have identified the most adequate method to obtain the data and have developed a specific loader responsible for converting the data to a format compatible with GeNS schema. We have implemented three distinct database loaders: a Web Services loader, a tabular files loader and a specific XML and text parser loader.

The data integration procedure follows the steps proposed by Davidson [22] in 1995: the data is basically retrieved and transformed to a common data model in order to match the current semantic schema and integrated in the database; Davidson also proposed two further steps that were not followed due to the nature of this platform.

Figure 1 contains a schema with the selected databases and, for each, the method used to extract the data: EMBL-EBI [23], UniProt (SwissProt and TrEMBL) [24], ExPASy (PROSITE and ENZYME), NCBI [20] (Entrez, Taxonomy, Pubmed, RefSeq, GenBank and OMIM), Biomart, ArrayExpress, InterPro, Gene Ontology [25] (GO), KEGG [11] (Genes, Pathway, Orthology and Drug) and PharmGKB (Gene, Drug and Disease).

Altogether these databases represent a very healthy set of data that span over 150 different data types. By merging all of this data, we obtain almost 7 million unique gene entries and over 140 million biological relations.

C. Meta -model design

Providing a correct representation of biological data without sacrificing the system's performance or scalability, among a long list of requirements, is still a challenge in bioinformatics [6]. In order to address this issue, one of two opposing schema design principles is typically applied: generalization or specialization.

A general schema prioritizes flexibility, scalability and the integration of several types and large volumes of data. It uses a large, dynamic set of data sources (which may vary throughout the time) in order to encompass as much diverse data as possible. A database designed according to this principle will allow its users to correlate heterogeneous data and to, eventually, extract conclusions that would otherwise hardly be visible. On the other hand, a specialized schema accommodates only a limited number of datasets. Usually, only a handful of sources of data are used: these sources were chosen from the very beginning and usually remain unaltered for long periods of time. This schema is usually considered as more suited to address more specific issues once that unlike general database schemas, scalability and flexibility are secondary aspects.

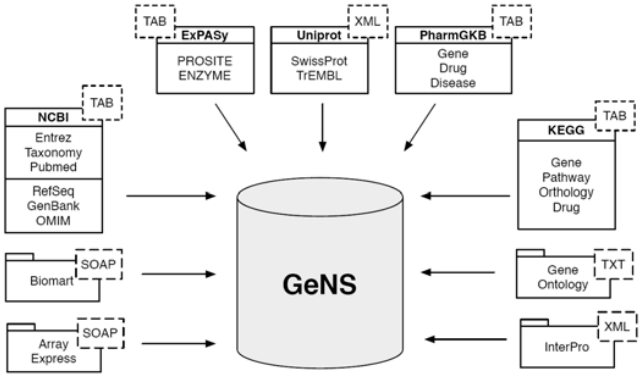

Fig 1. Schematic representation of the dabases integrated in GeNS

In GeNS we take advantage of both methods. To physically store the data we use a general schema that certifies the scalability and flexibility of the database. Then we support this physical schema with a concrete meta-model where all the entities and relations are specified.

Figure 2a contains this meta-model, where the gene plays a central role. Related with each gene there is a network of data types that map to the previously shown databases. The addition of databases that contain new data types only requires changes in the meta-model and not in the physical model.

D. Physical schema design

Figure 2b represents the physical database model. Because we needed to explicitly store all the relations between genes and proteins, due to implementation purposes we have changed the central role from gene in the meta-model to protein in the physical model. Following, we will describe in detail the concepts applied in the design of the database.

Organism: Stores taxonomic information; each entry corresponds to an organism with any given number of associated proteins. This table is the root of the hierarchical model. For each organism, we store organism detailed information such as its scientific names and reference sequence.

Protein: This table stores information regarding each protein entry. This information includes gene *locus*, gene and protein sequence and the relations to two distinct tables: *Identifier* and *BioEntity*.

Identifier: Contains all the synonyms, alternative names and identifiers for each entry.

DataType: Contains a list with all the types of data retrieved from external databases, encompassing both identifiers and biological entities. Every entry in either *Identifier* and/or *BioEntity* tables references this table, so that we may easily determine the type of the data, thus preventing semantic related errors (e.g. comparing two completely unrelated objects).

BioEntity: This table stores unique identifiers belonging to the biological entities associated with a given protein; this includes, among other things, pathway, gene ontology and

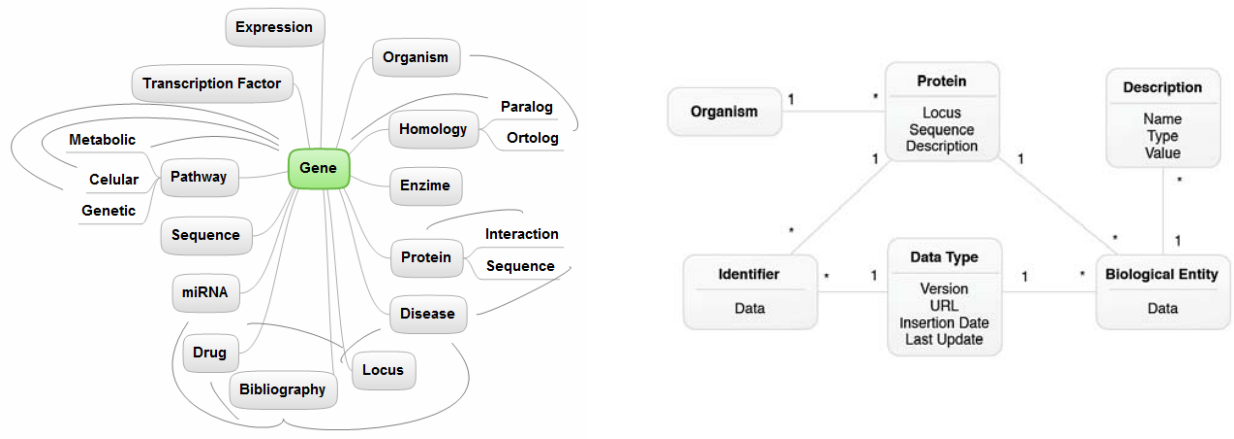

Fig. 2. GeNS database schema; a) meta-model centred in the gene and related concepts; b) the protein centric database physical schema

gene expression identifiers. Detailed data regarding a specific entry in this table will reside in the *Description* table.

Description: The description table stores structured data related to a specific biological entity. Examples of usage include the detailed description of a pathway or the mutation of a genetic disease.

This hierarchical organization not only simplifies the database schema (hence making it easier to understand and maintain) but also greatly improves the system's performance upon queries by simplifying access to the data that really matters. The system is also very flexible due to the way it maps data to proteins: each biological entity in the *BioEntity* table is unique and, thanks to an association table, multiple to multiple connections between proteins and biological entities keep data replication down to a bare minimum, while ensuring that the system's scalability and performance remain unaffected, along with all the benefits provided by the hierarchical model.

V. Results

A. Manual utilization

The following example (Figure 3) demonstrates one of many possible scenarios in GeNS: a researcher wants to obtain the network of concepts related with the following gene: 'sce:Q0085'. The system starts by determining the internal protein identifier through the *Identifier* table. With this identifier, we can now determine the alternative gene ids (still within the *Identifier* table).

Subsequently, the system will ascertain the corresponding organism; in this particular case, we already know the answer due to the first three letters of the identifier (*sce*, the short name for *Saccharomyces cerevisiae*) but this fact will not affect the process. In order to do so, GeNS looks up the *Protein* table and uses the taxonomic id to identify the organism in the *Organism* table. In the *Protein* table it is also possible to find the gene locus, it sequence and a general description.

Following this procedure, GeNS maps every biological entity associated to our pre-determined protein identifier by looking up the *ProteinBioEntity* table (that contains all the relations between the two). This allows GeNS to retrieve the biological entities in the *BioEntity* table which, in turn, contain homology, bibliography, expression, ontology, pathway and enzyme related data, among others.

Finally, more details about each biological entity can be obtained by looking up its description in the *BioEntityDescription* table.

Extending this example, the researcher wants to obtain all others genes related with the KEGG pathway 'sce00190' where the gene 'sce:Q0085' was initially present. To do so he searches the *Protein* table for all the entries that contain a relation to the table *BiologicalEntity* that matches the required pathway.

B. Programmatic utilization

1) A text mining example

QuExT (*Query Expansion Tool)* is a web application designed to search the biomedical literature in order to find relationships among sets of genes [26]. For a given list of genes, it expands the initial search in several biological domains using a mesh of co-related terms, extracts the most relevant document from the literature, and organizes them according to domain weighted factors. The role of GeNS database is to retrieve the network of concepts related with each gene entry in order to perform the query expansion.

2) Studing common characteristics in a set of genes

GeneBrowser is a web-based application that offers to the user several interpretation perspectives to help giving biological significance to the result coming from a DNA-microarray experiment. A previous version was presented in

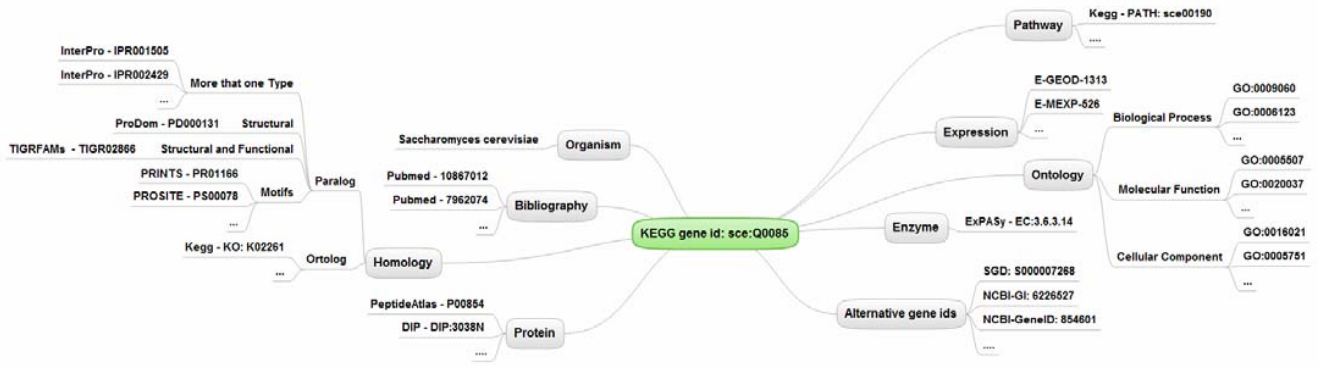

Fig. 3. Example that demonstrates the usage of GeNS to obtain the network of concepts related with the gene ‘sce:Q0085’.

[27] and an improved version is expected soon. For a given set of genes the system obtains and shows to the user relevant information extracted from external databases. Other features of the system include the possibility to see the accumulation of genes into several categories (Pathways, Gene Ontology terms and KEGG Orthology terms).

The possibility of using the GeNS system enables a fast and easier development of an application because the development team only has to concentrate in the visualization and analysis of data due to the easiness of use of the database.

VI. Availability

Both the schema and the data are available to be downloaded at http://bioinformatics.ua.pt/applications/gens. We provide a full copy compatible with SQL Server 2008 representing approximately 20GB. For other DBMS, one can download Tab delimited files that mirror the database schema.

As a convenience to users who do not want to maintain a local instance of the database, we also provide a public web services interface. To download the database, or to obtain more information regarding the web services interface access http://bioinformatics.ua.pt/applications/gens.

VII. Conclusions

In this paper we have presented the schema and the implementation of a platform for the integration of biological data. The main contributions of this tool are its easiness to use and maintain, while offering great performance, its coverage and scalability, attested by the number of data sources already integrated and by the simple procedure to augment these sources.

The current instance already integrates the most relevant molecular biology databases having a total of 140 million biological relations. Despite that we are still working on feeding GeNS with other databases and data types. To show the functionality of GeNS we have presented two applications that we have been developing and that are mainly supported by GeNS services. The first is a tool that performs the functional analysis of microarray data and the second uses GeNS data to improve text mining results over PubMed.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 215847- the EU-ADR project. J. Arrais is funded by FCT grant SFRH/BD/23837/2005.

References

[1] W. Zhong and P. W. Sternberg, "Automated data integration for developmental biological research," *Development*, vol. 134, pp. 3227-38, Sep 2007.

[2] Z. Lacroix, "Biological data integration: wrapping data and tools," *IEEE Trans Inf Technol Biomed*, vol. 6, pp. 123-8, Jun 2002.

[3] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, "Data integration and genomic medicine," *J Biomed Inform*, vol. 40, pp. 5-16, Feb 2007.

[4] L. D. Stein, "Integrating biological databases," *Nat Rev Genet*, vol. 4, pp. 337-45, May 2003.

[5] T. Topaloglou, A. Kosky, and V. Markowitz, "Seamless integration of biological applications within a database framework," *Proc Int Conf Intell Syst Mol Biol*, pp. 272-81, 1999.

[6] L. Wong, "Technologies for integrating biological data," *Brief Bioinform*, vol. 3, pp. 389-404, Dec 2002.

[7] M. Y. Galperin, "The Molecular Biology Database Collection: 2008 update," *Nucleic Acids Res*, Nov 19 2007.

[8] F. Al-Shahrour, L. Arbiza, H. Dopazo, J. Huerta-Cepas, P. Minguez, D. Montaner, and J. Dopazo, "From genes to functional classes in the study of biological systems," *BMC Bioinformatics*, vol. 8, p. 114, 2007.

[9] Z. Fang, J. Yang, Y. Li, Q. Luo, and L. Liu, "Knowledge guided analysis of microarray data," *J Biomed Inform*, vol. 39, pp. 401-11, Aug 2006.

[10] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Res*, vol. 35, pp. D21-5, Jan 2007.

[11] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "KEGG for linking genomes to life and the environment," *Nucleic Acids Res*, Dec 12 2007.

[12] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma, "ArrayExpress--a public database of microarray experiments and gene

expression profiles," *Nucleic Acids Res*, vol. 35, pp. D747-50, Jan 2007.

[13] V. Detours, J. E. Dumont, H. Bersini, and C. Maenhaut, "Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets," *FEBS Lett*, vol. 546, pp. 98-102, Jul 3 2003.

[14] F. Achard, G. Vaysseix, and E. Barillot, "XML, bioinformatics and data integration," *Bioinformatics*, vol. 17, pp. 115-25, Feb 2001.

[15] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. W. Stringer-Calvert, J. D. Tenenbaum, and P. D. Karp, "BioWarehouse: a bioinformatics database warehouse toolkit," *BMC Bioinformatics*, vol. 7, p. 170, 2006.

[16] J. Kuntzer, C. Backes, T. Blum, A. Gerasch, M. Kaufmann, O. Kohlbacher, and H. P. Lenhof, "BNDB - the Biochemical Network Database," *BMC Bioinformatics*, vol. 8, p. 367, 2007.

[17] A. Birkland and G. Yona, "BIOZON: a hub of heterogeneous biological data," *Nucleic Acids Res*, vol. 34, pp. D235-42, Jan 1 2006.

[18] E. Cadag, B. Louie, P. J. Myler, and P. Tarczy-Hornoch, "Biomediator data integration and inference for functional annotation of anonymous sequences," *Pac Symp Biocomput*, pp. 343-54, 2007.

[19] J. Kohler, S. Philippi, and M. Lange, "SEMEDA: ontology based semantic integration of biological databases," *Bioinformatics*, vol. 19, pp. 2420-7, Dec 12 2003.

[20] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res*, vol. 35, pp. D26-31, Jan 2007.

[21] J. L. Oliveira, G. Dias, I. Oliveira, P. Rocha, I. Hermosilla, J. Vicente, I. Spiteri, F. Martin-Sánchez, and A. S. Pereira, "DiseaseCard: A Web-Based Tool for the Collaborative Integration of Genetic and Medical Information," in *Biological And Medical Data Analysis: 5th International Symposium*, Springer, Ed., 2004, pp. 409-417.

[22] S. B. Davidson, C. Overton, and P. Buneman, "Challenges in integrating biological data sources," *Journal of Computational Biology*, vol. 2, pp. 557-572, 1995.

[23] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, and S. Searle, "Ensembl 2008," *Nucleic Acids Res*, vol. 36, pp. D707-14, Jan 2008.

[24] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Res*, vol. 34, pp. D187-91, Jan 1 2006.

[25] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, May 2000.

[26] J. Arrais, J. G. L. M. Rodrigues, and J. L. Oliveira, "Improving Literature Searches in Gene Expression Studies," in *Advances in Intelligent and Soft Computing : 2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics*, J. M. Corchado, J. F. De Paz, M. P. Rocha, and F. Fernandez Riverola, Eds. Berlin, DE: Springer Berlin / Heidelberg, 2009, pp. Capt. 10, p. 74 - 82.

[27] J. Arrais, B. Santos, J. Fernandes, L. Carreto, M. A. S. Santos, and J. L. Oliveira, "GeneBrowser: an approach for integration and functional classification of genomic data," in *Journal of Integrative Bioinformatics*. vol. 4, 2007.