

Systems Biology Warehousing: Challenges and Strategies toward Effective Data Integration

Thomas Triplet

Centre for Structural and Functional Genomics

Department of Computer Science

Concordia University

1455 De Maisonneuve Blvd. West, Montreal, Qc Canada

Email: thomastriplet@gmail.com

Gregory Butler

Centre for Structural and Functional Genomics

Department of Computer Science

Concordia University

1455 De Maisonneuve Blvd. West, Montreal, Qc Canada

Email: gregb@encs.concordia.ca

Abstract—The rapid development of genomics, proteomics, metabolomics and structural genomics techniques have provided an unprecedented amount of data, enabling system-wide biological research. Although information integration has been well investigated in database theory research, biological data present numerous challenges from the lack of standard formats to data inconsistencies resulting from experimental data variations. Satisfying and practical solutions are still lacking and current molecular biology databases serve primarily as simple data repositories with limited query capabilities. They also provide little to no integration with other databases. However, the success of systems biology is contingent on the ability to integrate and utilize a wide variety of types of data. It also relies on computational techniques to automatically predict and assign functional annotations of proteins as effective integration of biological data should enable scientists to perform comparative analyses, modelling and inference of protein functions. Therefore, there is a need for a paradigm shift toward systems biology databases with flexible query systems that focus on answering a diversity of questions from biologists without the need to reconfigure the underlying database architectures.

Keywords-genomics; proteomics; systems biology; data warehousing; data integration

I. INTRODUCTION

Life sciences techniques made significant improvements over the past decades, resulting in huge amounts of data collected over the years by the scientific community. In order to facilitate the organization and the subsequent analyses of this valuable data, databases have been developed very early. Since then, the number of databases has dramatically increased. The 2010 Molecular Biology Database Collection [1] includes well over a thousand databases, each describing millions of biological records.

This unprecedented wealth of information originating from genomic studies represents a tremendous potential in all areas of biological science. The emerging information integrated with existing knowledge bases could lead to an explosive understanding of complex molecular interactions, networks and pathways. Successful data integration is one of the keys to successful bioinformatics research [2]: scientists need an integrated view of these heterogeneous data sources

with advanced data-mining, analysis and visualisation tools. The continuing data growth will lead to an increasing need for large-scale data management as biological discovery depends, to a large extent, on the presence of clean, up-to-date and well-organised datasets.

Unfortunately, the rapidly growing number of different molecular biology databases, created at various places worldwide, serve primarily as data warehouses with simple query interfaces designed for specific tasks. The databases are not readily amenable to complex system-based research that requires the integration of a large number of these disparate databases. There is a need for a paradigm shift toward systems biology databases with flexible query systems that focus on answering a diversity of questions from biologists without the need to reconfigure the underlying database architectures.

In this paper, we present an overview of the problem of the integration of multiple biological databases from the perspective of large-scale analysis of biological systems. Section II overviews some technical aspects of data warehousing. Section III explains the specificities of biological data and some of the challenges they raise when being integrated, from data heterogeneity (Section III-A) to experimental variability (Section III-C). Section IV then describes data warehousing requirements for effective systems biology and reviews key features and limitations of several major data warehousing frameworks.

II. DATA INTEGRATION METHODS

Information integration has been well investigated in database theory. Currently, three main approaches are generally considered when integrating data: the *Extract, Transform and Load* (ETL), the *Local-As-View* (LAV) and *semantic integration* methods.

A. Extract-Transform-Load method

An ETL-based warehouse is constructed by Extracting, Transforming and Loading the data to integrate into a single unified schema. The transformation step allows the data to

be pre-processed before they are integrated in the warehouse, which may be useful to address some of the problems mentioned in Section III, in particular those related to data inaccuracies and inconsistencies. However, ETL methods generally lack flexibility because the warehouse schema is tightly coupled the data sources. As a result, integrating new databases requires considerable effort as the entire warehouse and subsequent queries need to be redefined. The warehouse schema may also have to be redesigned if one of the data sources schema changes after an update.

B. Local-As-View method

Local-As-View (LAV) methods [3], [4] are designed to address the flexibility issues of ETL methods. They are based on functions — or wrappers — that provide an abstraction layer and a simplified view of the integrated data sources. They traditionally rely on dynamic logical views, which are featured by most DBMSs. However, logical views are usually constructed using natural joins to correlate database keys and can therefore provide erroneous or misleading answers (see Section III-C for details). Dynamically generating views that contains millions of records can also be computationally expensive and/or inefficient.

C. Semantic integration

Unlike the ETL and LAV methods, semantic integration methods [5], [6] do not address issues related to the underlying architecture of the DBMS of the warehouse, but focus on the *semantic* integration of related entities or concepts from heterogeneous data sources through the use of ontologies, that is, formal descriptions of the concepts and entities for a domain of interest and the relationships that hold among them. Semantic integration is therefore useful to handle heterogeneous data that lack standardization (see Section III-B) as is often the case in biology. It is however computationally expensive and often requires large data centres to be effective [7].

III. BIOLOGICAL DATA SPECIFICITIES AND CHALLENGES

Biological data present a number of specificities and raise many challenges when being integrated, in particular in the context of large-scale analyses of biological systems as a whole. As a result, satisfying and practical solutions derived from the methods described above have proven to be elusive for these complex data sources.

A. A wealth of heterogeneous data

The 2010 Molecular Biology Database Collection [1] contains 1,230 databases, requiring a database of databases to integrate data and keep track of all the knowledge available to biologists today [8]. Among these databases is the extent of our knowledge related to genomics [9], proteomics [10], metabolomics [11] and structural genomics [12].

As an illustration of the vast amounts of accessible data, consider: (i) the RefSeq [13] database, containing 16 million sequences (152 billion base pairs) from over 10,000 species, (ii) the Joint Genome Institute (JGI) sequencing projects [14], comprising 200 billion base pairs, (iii) the Gene Ontology [15], containing 30,914 terms that describe the biological function of nearly 500,000 gene products (iv) the eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) database [16], describing 224,847 orthologous groups covering 2.5 million proteins (v) the Kyoto Encyclopedia of Genes and Genomes (KEGG) [17], containing 116,962 metabolic pathways comprising 16,291 compounds.

While these five distinct molecular biology databases represent a small fraction of those available, they still contain a wealth of data beneficial to system biology studies. But, conducting searches or identifying correlations across just these five databases ranges from extremely limited to non-existent. Furthermore, the web-based interfaces are generally designed with the expectation of presenting the user with only a few results from a query. The display and manual analysis of queries that generate hundreds or thousands of results is not practical.

B. Lack of unique standards

Integrating data from multiple databases provides an additional challenge because of the different data formats and structures [18]. Those data are very disparate and often stored in database management systems (DBMS) or as simple plain text files in multiple data repositories, which provide very limited compatibility or interoperability between systems.

Describing a chemical structure clearly illustrates this critical problem since there are more than 80 different formats currently in use. There are also numerous ways to name a chemical compound, which includes common names, abbreviations and systematic nomenclatures [19] defined by the International Union of Pure and Applied Chemistry (IUPAC). IUPAC systematic nomenclatures are an evolving process and tend to be cumbersome for relatively complex molecules and difficult to generate even with software that automate the process. Errors are common and there is a relatively high failure rate in actually generating a name. An alternative approach uses a linear character string to represent a 2D structure of the compound. SMILES (Simplified Molecular Input Line Entry System) is the most popular approach of generating a simple text-based representation of a chemical structure [20]. However, numerous distinct and correct SMILES strings can be generated from a single chemical structure. InChI (International Chemical Identifier) — “a new standard for molecular informatics” — is a recent variation on SMILES strings [21]. While InChI generates a unique character string for a given molecule, it is relatively new and not as widely used as SMILES strings.

Another example is DBGET/LinkDB [22], which aims at providing a unified database query mechanism. Yet, only a handful databases are capable of handling DBGET/LinkDB requests. Furthermore, using DBGET, it is currently not possible to query two databases and automatically integrate the results: data from the integrated databases remains largely independent and difficult to combine.

C. Data inconsistencies and inaccuracies

1) *Experimental variations*: Because of their experimental nature, biological data are also routinely sparse and/or inaccurate. For example, consider the fungus *Aspergillus niger* available from two major repositories for fungal genomics: the *JGI Genomes* [23] database developed by the U.S. Department of Energy and *EnsemblFungi* [24], a fungal database maintained by the European Bioinformatics Institute. GBrowse [25] is the primary means to visualize genomes on JGI whereas EnsemblFungi is powered by BioMart [26].

On JGI Genome, consider the gene `fgenesh1_pg.C_scaffold_6000331` located on Chromosome I:612-2,561, which codes for a protein involved in fungal specific transcription. The sequence of this gene was matched with sequences from EnsemblFungi using BLASTN [27]. The best hit on EnsemblFungi was a perfect match ($E = 0$, 100% sequence identify). However, on EnsemblFungi, the gene is located on Chromosome III:3,638,575-3,640,524. Similarly, gene `e_gwl.12.166.1` at the locus Chr.III:2,768-4,422 on JGI was identified ($E = 1.5 * 10^{-88}$) at Chr.II:2,593,707-2,594,464 on EnsemblFungi. In addition, both repositories use non-standardized mapping systems: coordinates are relative to scaffolds in JGI whereas they are relative to chromosomes in EnsemblFungi.

2) *Data entry errors*: Compounding the above inconsistencies are entry errors in the original data, such as spelling mistakes and the inadvertent addition or deletion of characters or spaces [28]. A number of experimental errors associated with the data are also to be expected [29], where estimates of annotation errors for gene products range from 8% to 49%, depending on the method [30]. Similarly, spelling and typographical errors have been measured to occur at rates of 1.5 to 2.5% and 1 to 3.2%, respectively [31]. Thus, data errors present an inherent challenge in the development of molecular biology databases [32].

3) *Approximate string matching and similarity functions*: 3rd Millennium[®] showed that the number of incorrect entries grows geometrically with the number of joins and reaches nearly 50% by the fourth join, assuming a conservative error rate of 15% per join [2]. Solely relying on database indexes to correlate database keys is generally dangerous and blind data integration can provide erroneous or misleading answers. Similarity functions unique to

molecular biology data are therefore required and numerous similarity algorithms have been developed but these are generally implemented as independent stand-alone programs accessible through web-servers. Nearly 1,200 web links to resources and software accessible to the scientific community have been documented [33]. This provides a variety of valuable tools to improve the quality and flexibility of biological database searches. For instance, BLAST [27] and FASTA [34] are well-known and highly utilized sequence homology approaches that search sequence databases using substitution matrices and string matching heuristics. Alternatively, PSI-BLAST [35] uses a profile-sequence comparison method, and HMMER applies hidden Markov models [36]. Similarly, programs such as Dali [37] and StruCTal [38] align the 3D structures of proteins present in the Protein Data-Bank. The Espresso [39] program combines both sequence and structure alignments to identify similar proteins.

Errors in the data are more likely to be accommodated by the robustness of these similarity searches [40] and the most reliable approach is to incorporate similarity algorithms into the database structure to *simulate* indexes that are normally based on binary search trees or hashes in most DBMS. Similarity measures may also improve semantic integration [41] when combined with ontologies.

Moreover, spelling mistakes may be detected using approximate string matching algorithms [31]. In particular, Damerau [42] and Levenshtein [43] estimated that 80% of human spelling mistakes could be automatically corrected using at most one character insertion, deletion, substitution or transposition.

4) *Data provenance*: To help address data quality issues, tracking the provenance of the data is critical [44]. The provenance typically consists of metadata associated with the data and is helpful for scientists to evaluate their quality and reliability. Its also allows them to examine the lineage of a piece of information, which shows all the steps involved in sourcing, moving and processing the data. Toward that end, all datasets and their transformations must be recorded. Goble [45] suggested that data provenance could be useful to address ownership and copyright issues as well as to record experimental protocols followed to generate the data, effectively ensuring the reproducibility of experimental results. When data is redundantly available from multiple sources, provenance may also be beneficial for automated data curation and arbitration of data inconsistencies.

D. Non-textual data

Biological data are not always textual. This adds to the difficulties of indexing data for effective data-mining. This is most notably the case for high throughput microscopy imaging and enzymatic activity experimental characterization. For example, microplate assays are widely used in research and drug discovery to detect biological or chemical events of samples. Those events are typically detected

by measuring the fluorescence intensity of samples from each of the ninety-six wells that compose a plate and are usually stored as greyscale pictures. SDS-PAGE (Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis) gels are broadly used for protein separation and analysis. Gels are usually stored as images and analysed using imaging tools such as ImageJ [46] and to date, indexing and mining those files is not possible using current databases.

E. Data versioning

Data versioning offers many benefits to end users. First, it enables data recovery in case of an entry mistake or system corruption. More importantly, it facilitates the analysis of historical changes of data, which can be helpful to enhance predictions and automatic data classification [47]. Barkstrom presented a formal structure for keeping track of files, source code, scripts, and related material for large-scale Earth science data production [48]. However, biological data present a number of unique challenges and effective solutions for biological data versioning are still lacking.

One of the main issues in biological data versioning is the variability of the data and the lack of well-defined protocol to compare the numerous formats. Biological experiments are conducted using some biological data as input and yield results. These results may be used as data sources in other experiments to yield further sets of results. This cycle of experimentation continues, presumably until the required set of results are achieved. Research questions and interests may also vary over time, thereby altering the nature of the data. This is often the case for emerging high throughput technologies, where refinement of experimental protocols can significantly change the types of data to be collected [2].

IV. DATA WAREHOUSING FOR SYSTEMS BIOLOGY

To date the National Center for Biotechnology Information lists over 2,500 whole-genome sequencing projects, including 833 in progress. JGI lists nearly 1,500 whole-genome in addition to over 850 sequencing projects. Obtaining protein functional assignments is a necessary first step to enable progress in research areas associated with development, evolution and physiology. Solving the challenging problem of assigning a function to proteins requires multiple approaches because sequence similarity techniques may provide functional information for at most 50% of these proteins [49], [50]. Addressing this complex biological issue necessitates a *Systems Biology* [51], [52] approach to data analysis that requires identifying relationships hidden within multiple databases [53]. An important mechanism to achieve this goal will require the development of next-generation databases that enable sophisticated queries beyond simple text-based searches.

A. Systems biology: a new scientific perspective

Biological systems are more than a set of independent components working together. Although they are composed

of a limited number of elements, these elements — proteins in particular — usually have multiple functions and interact very tightly to form complex pathways. Historically, scientists have focused their research on isolating those elements to understand their individual functions and activities. The knowledge of their function is indeed critical to comprehend the intricate biological machinery of the whole system.

However, the success of the Human Genome Project has ushered in a new scientific perspective, a system-wide view of protein function and biological activity: while traditional methods focus on the detailed analysis of isolated proteins to understand its cellular function, systems biology applies a holistic approach to understand the details of cell biology and evolution as a whole. This relatively new concept in biology, which is expected to yield more realistic models of a complete biological system, requires the integration of data from the individual subsystems. Biological models are usually constructed — and validated — using high-throughput quantitative data including, but not limited to, genome sequencing, gene expression, proteomics, metabolomics and high-resolution microscopy imaging.

Effective integration of biological data should enable scientists to perform comparative analyses, modelling and inference of protein functions. The success of systems biology is therefore contingent on the ability to integrate and utilize a wide variety of types of data and computational techniques to automatically predict and assign functional annotations of proteins. Systems biology databases are expected to expand upon the traditional sequence and structure approach because the primary method to assign a function to a protein of unknown function is to identify a relationship with a protein of known function. Additional protein associations may also be made through protein interaction networks, metabolic pathways, protein expression patterns or any number of relationships envisioned by a biologist. The key to this approach is moving the design focus from a fixed database structure defining precomputed relationships between elements to a fluid and flexible relational model that can be adapted to the biologist questions [54] without re-designing the underlying data structure.

B. Overview of Existing Frameworks and their Limitations

Over the past few years, a number of specialized data warehouses have been developed to accommodate the specificities of biological data and to address specific needs. For example, the e-Fungi database [55] integrates data from 36 fungal genomes and aims at facilitating the systematic comparative study of those genomes. GeWare [56] is a laboratory information management system, featuring tools for the integrated analysis of clinical data from large biomedical research studies.

In this section, we briefly describe the main capabilities and limitations of a few general data warehousing frameworks although it is beyond the scope of this paper to

provide a formal and comprehensive benchmark.

1) *BioMart Central Portal*: BioMart Central Portal [26] is a complete framework that provides tools to federate a variety of biological databases. These include major biomolecular sequence, pathway and annotation databases. Moreover, the web server features a unified user-friendly interface for mining data from various datasets. The web server also supports programmatic access through a Perl API as well as RESTful web services. Queries are defined as a set of successive filters to be applied to data. Queries are however limited to two datasets at once, hence limiting the scaling capabilities of BioMart. It is also not possible to edit or create new filters, restricting possible queries to the otherwise comprehensive set of filters defined by the developers of the framework.

2) *BioXRT*: The BioXRT framework [57] is designed to allow biologists to publish their data on the Internet with only minimal knowledge of database design and usage. It provides a highly flexible and extensible database structure as well as tools to import spreadsheets. However, flexibility is achieved to the detriment of data types as all pieces of data are defined as strings of characters. Consequently, queries are constrained to string matching and BioXRT does not support advanced data-mining tools for complex biological questions.

3) *InterMine*: FlyMine [58] is a data warehouse built upon InterMine that integrates and utilizes numerous biological data sources. It features parsers for integrating data from numerous biological formats and facilities for adding one's own data. It provides a web access to integrated data at a number of different levels, from simple browsing to construction of complex queries and it includes a user-friendly web interface that can be easily customised for the user's needs. The provenance of the integrated datasets is also tracked. However, InterMine does not provide tools for classifying or clustering data and queries may not include similarity functions to address annotation errors as discussed in Section III-C.

4) *Open Genome Resource (OGeR)*: Strepto-DB [59] and SYSTOMONAS [60] are databases for the comparative genome analysis of streptococci and pseudomonas respectively, and rely on the Open Genome Resource (OGeR) to achieve data integration of external resources. OGeR is an open source system for the storage, visualization and analysis of prokaryotic genome data. Genome sequences and annotations can be automatically downloaded from relevant databases and features cross-references to external databases. However, like other frameworks, OGeR does not provide tools for clustering and statistical analysis, nor does it provide advanced mining tools besides pairwise and multiple sequence alignment tools.

5) *PROFESS*: The PROtein Function, Evolution, Structure and Sequence (PROFESS) database [61] integrates numerous biological databases and aims at giving an overview

of biological systems by integrating protein annotations at different levels: function, evolution, structure and sequence. The primary means to query the database is the "PROFESSor", a unified text field that mines data from any integrated database. PROFESS also provides clustering and aggregation tools for statistical analysis of large datasets and features a user-friendly modular web interface. However, like other systems, its query system is based on a predefined non-customizable set of filters and does not yet support similarity functions besides standard BLAST searches.

V. CONCLUSION

Although biological data present a number of unique specificities making them challenging to integrate, there is a growing need for effective integration of biological datasets to enable large scale and comparative analysis of the numerous genomes being sequenced. To date, no biological data warehouse meets all the requirements for effective integration of system-wide data. BioXRT offers a flexible and extensible database structure, BioMart provides advanced data-mining tools although they may not be extended by users. PROFESS features a flexible and modular user interface and tools for clustering and statistical analysis of large datasets. InterMine also features a customizable user interface and is helpful to track the provenance of data.

However, there now exists a variety of resources that may be helpful in accommodating data inaccuracies, such as approximate string matching or similarity-based algorithms that may be implemented within database management systems for the next generation of biological data warehouses.

FUNDING

This work was supported by the Cellulosic Biofuel Network, funded by Agriculture and Agri-Food Canada.

REFERENCES

- [1] G. R. Cochrane and M. Y. Galperin, "The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources." *Nucleic acids research*, vol. 38, no. Database issue, pp. D1–4, Jan. 2010.
- [2] 3rd Millennium Inc., "Practical Data Integration in Biopharmaceutical R&D: Strategies and Technologies," 2002.
- [3] R. Pottinger and A. Halevy, "MiniCon: A scalable algorithm for answering queries using views," *The VLDB Journal*, vol. 10, no. 2-3, pp. 182–198, 2001.
- [4] A. Halevy, "Answering queries using views: A survey," *VLDB Journal*, vol. 10, no. 4, pp. 270–294, 2001.
- [5] F. Giunchiglia, M. Yatskevich, and P. Shvaiko, "Semantic Matching: Algorithms and Implementation," *Journal on Data Semantics*, vol. 9, pp. 1–38, 2007.
- [6] S. M. Falconer, N. F. Noy, and M.-A. Storey, "Ontology Mapping - a User Survey," in *Proceedings of the 2nd International Workshop on Ontology Matching*. CEUR-WS.org, 2007, pp. 113–125.

- [7] C. Hewitt, "ORGs for Scalable, Robust, Privacy-Friendly Client Cloud Computing," *IEEE Internet Computing*, vol. 12, no. 5, pp. 96–99, Sep. 2008.
- [8] P. A. Babu, J. Udyama, R. K. Kumar, R. Boddepalli, D. S. Mangala, and G. N. Rao, "DoD2007: 1082 molecular biology databases." *Bioinformatics*, vol. 2, no. 2, pp. 64–67, 2007.
- [9] H. Camacho, A. Cintado, and M. Duenas, "Technology evolution for genomic revolution," *Biotechnologia Aplicada*, vol. 22, no. 2, pp. 83–90, 2005.
- [10] W. Yang, H. Steen, and M. R. Freeman, "Proteomic approaches to the analysis of multiprotein signaling complexes." *Proteomics*, vol. 8, no. 4, pp. 832–851, Feb. 2008.
- [11] E. M. Lenz and I. D. Wilson, "Analytical strategies in metabonomics." *Journal of proteome research*, vol. 6, no. 2, pp. 443–458, Feb. 2007.
- [12] F. von Delft, D. McRee, and C. Kang, *Prospects for high-throughput structure determination by X-ray crystallography*. CRC Press, 2003, pp. 55–94.
- [13] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott, "NCBI Reference Sequences: current status, policy and new initiatives." *Nucleic acids research*, vol. 37, no. Database issue, pp. D32–36, Jan. 2009.
- [14] U.S. Department of Energy Joint Genome Institute, "Sequencing project," Nov. 2010. [Online]. Available: <http://www.jgi.doe.gov/sequencing/>
- [15] The Gene Ontology Consortium, "The Gene Ontology in 2010: extensions and refinements." *Nucleic acids research*, vol. 38, no. Database issue, pp. D331–335, Jan. 2010.
- [16] J. Muller, D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks, L. J. Jensen, and P. Bork, "eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations." *Nucleic acids research*, vol. 38, no. Database issue, pp. D190–195, Jan. 2010.
- [17] 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*, vol. 36, no. Database issue, pp. D480–484, Jan. 2008.
- [18] L. Wong, "Technologies for integrating biological data," *Brief Bioinform*, vol. 3, no. 4, pp. 389–404, Dec. 2002.
- [19] D. I. Cooke-Fox, G. H. Kirby, and J. D. Rayner, "Computer translation of IUPAC systematic organic chemical nomenclature. 1. Introduction and background to a grammar-based approach," *Journal of Chemical Information and Modeling*, vol. 29, no. 2, pp. 101–105, May 1989.
- [20] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Modeling*, vol. 28, no. 1, pp. 31–36, Feb. 1988.
- [21] A. McNaught, "The IUPAC international chemical identifier : InChI-A new standard for molecular informatics," *Chemistry international*, vol. 28, no. 6, pp. 12–14, 2006.
- [22] W. Fujibuchi, S. Goto, H. Migimatsu, I. Uchiyama, A. Ogiwara, Y. Akiyama, and M. Kanehisa, "DBGET/LinkDB: an integrated database retrieval system." in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, Jan. 1998, pp. 683–694.
- [23] U.S. Department of Energy Joint Genome Institute, "Genome database," Nov. 2010. [Online]. Available: <http://genome.jgi-psf.org/programs/fungi/>
- [24] P. J. Kersey, D. Lawson, E. Birney, P. S. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kähäri, R. J. Kinsella, E. Kulesha, U. Maheswari, K. Megy, M. Nuhn, G. Proctor, D. Staines, F. Valentin, A. J. Vilella, and A. Yates, "Ensembl Genomes: extending Ensembl across the taxonomic space." *Nucleic acids research*, vol. 38, no. Database issue, pp. D563–569, Jan. 2010.
- [25] L. D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva, and S. Lewis, "The generic genome browser: a building block for a model organism system database." *Genome research*, vol. 12, no. 10, pp. 1599–1610, Oct. 2002.
- [26] S. Haider, B. Ballester, D. Smedley, J. Zhang, P. Rice, and A. Kasprzyk, "BioMart Central Portal—unified access to biological data." *Nucleic acids research*, vol. 37, no. Web Server issue, pp. W23–27, Jul. 2009.
- [27] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J Mol Biol*, vol. 215, no. 3, pp. 403–410, Oct. 1990.
- [28] C. Hadley, "Righting the wrongs," *EMBO Reports*, vol. 4, no. 9, pp. 829–831, 2003.
- [29] CODATA Task Group on biological macromolecules and colleagues, "Quality control in databanks for molecular biology," *BioEssays*, vol. 22, no. 11, pp. 1024–1034, Oct. 2000.
- [30] C. Jones, A. Brown, and U. Baumann, "Estimating the annotation error rate of curated GO database sequence annotations," *BMC Bioinformatics*, vol. 8, no. 1, 2007.
- [31] G. Navarro, "A Guided Tour to Approximate String Matching," *ACM Computing Surveys*, vol. 33, 1999.
- [32] J. B. Cushing, "Metadata and semantics: a computational challenge for molecular biology." *Omics : a journal of integrative biology*, vol. 7, no. 1, pp. 23–24, Jan. 2003.
- [33] J. A. Fox, S. McMillan, and B. F. F. Ouellette, "Conducting research on the web: 2007 update for the bioinformatics links directory." *Nucleic acids research*, vol. 35, no. Web Server issue, pp. W3–5, Jul. 2007.
- [34] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proc Natl Acad Sci U S A*, vol. 85, no. 8, pp. 2444–2448, Apr. 1988.

- [35] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.
- [36] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999.
- [37] S. Dietmann, J. Park, C. Notredame, A. Heger, M. Lappe, and L. Holm, "A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3." *Nucleic acids research*, vol. 29, no. 1, pp. 55–57, Jan. 2001.
- [38] M. Levitt, "A unified statistical framework for sequence comparison and structure comparison," *Proceedings of the National Academy of Sciences*, vol. 95, no. 11, pp. 5913–5920, May 1998.
- [39] F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaepli, V. Keduas, and C. Notredame, "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee." *Nucleic acids research*, vol. 34, no. Web Server issue, pp. W604–608, Jul. 2006.
- [40] D. J. States and D. Botstein, "Molecular sequence accuracy and the analysis of protein coding regions." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 13, pp. 5518–5522, Jul. 1991.
- [41] Q. Ji, P. Haase, and G. Qi, "Combination of Similarity Measures in Ontology Matching using the OWA Operator," in *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Base Systems*. In: Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Base Systems, 2008.
- [42] F. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.
- [43] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, 1966.
- [44] P. Buneman, S. Khanna, and W. C. Tan, "Data Provenance: Some Basic Issues," *Lecture Notes In Computer Science; Vol. 1974*, p. 87, 2000.
- [45] C. Goble, "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics," in *Workshop on Data Derivation and Provenance*, Chicago, USA, 2002.
- [46] W. Rasband, "ImageJ," Nov. 2010. [Online]. Available: <http://imagej.nih.gov/ij/>
- [47] P. Revesz and T. Triplet, "Temporal Data Classification Using Linear Classifiers," *Information Systems*, vol. 36, no. 1, pp. 30–41, 2011.
- [48] B. Barkstrom, *Data Product Configuration Management and Versioning in Large-Scale Production of Satellite Scientific Data*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2003, vol. 2649, pp. 118–133.
- [49] Y. Pouliot, J. Gao, Q. J. Su, G. G. Liu, and X. B. Ling, "DIAN: a novel algorithm for genome ontological classification." *Genome research*, vol. 11, no. 10, pp. 1766–1779, Oct. 2001.
- [50] L. J. Jensen, R. Gupta, H.-H. Staerfeldt, and S. Brunak, "Prediction of human protein function according to Gene Ontology categories." *Bioinformatics (Oxford, England)*, vol. 19, no. 5, pp. 635–642, Mar. 2003.
- [51] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: systems biology." *Annual review of genomics and human genetics*, vol. 2, pp. 343–372, Jan. 2001.
- [52] H. Kitano, "Systems biology: a brief overview." *Science*, vol. 295, no. 5560, pp. 1662–1664, Mar. 2002.
- [53] A. R. Joyce and B. O. Palsson, "The model organism as a system: integrating 'omics' data sets," *Nat Rev Mol Cell Biol*, vol. 7, no. 3, pp. 198–210, Mar. 2006.
- [54] R. Stevens, "A classification of tasks in bioinformatics," *Bioinformatics*, vol. 17, no. 2, pp. 180–188, Feb. 2001.
- [55] C. Hedeler, H. M. Wong, M. J. Cornell, I. Alam, D. M. Soanes, M. Rattray, S. J. Hubbard, N. J. Talbot, S. G. Oliver, and N. W. Paton, "e-Fungi: a data resource for comparative analysis of fungal genomes." *BMC genomics*, vol. 8, no. 1, p. 426, Jan. 2007.
- [56] E. Rahm, T. Kirsten, and J. Lange, "The GeWare data warehouse platform for the analysis of molecular-biological and clinical data," *Journal of Integrative Bioinformatics*, vol. 4, no. 1, 2007.
- [57] J. Zhang, G. E. Duggan, R. Khaja, and S. W. Scherer, "BioXRT: a novel platform for developing online biological databases based on the Cross-Referenced Tables model," in *3rd Canadian Working Conference on Computational Biology*, Markham, Canada, 2004.
- [58] R. Lyne, R. Smith, K. Rutherford, M. Wakeling, A. Varley, F. Guillier, H. Janssens, W. Ji, P. McLaren, P. North, D. Rana, T. Riley, J. Sullivan, X. Watkins, M. Woodbridge, K. Lilley, S. Russell, M. Ashburner, K. Mizuguchi, and G. Micklem, "FlyMine: an integrated database for Drosophila and Anopheles genomics." *Genome biology*, vol. 8, no. 7, p. R129, Jan. 2007.
- [59] J. Klein, R. Münch, I. Biegler, I. Haddad, I. Retter, and D. Jahn, "Strepto-DB, a database for comparative genomics of group A (GAS) and B (GBS) streptococci, implemented with the novel database platform 'Open Genome Resource' (OGeR)." *Nucleic acids research*, vol. 37, no. Database issue, pp. D494–8, Jan. 2009.
- [60] C. Choi, R. Munch, B. Bunk, J. Barthelmes, C. Ebeling, D. Schomburg, M. Schobert, and D. Jahn, "Combination of a data warehouse concept with web services for the establishment of the Pseudomonas systems biology database SYS-TOMONAS," *Journal of Integrative Bioinformatics*, vol. 4, no. 1, 2007.
- [61] T. Triplet, M. Shortridge, M. Griep, J. Stark, R. Powers, and P. Revesz, "PROFESS: a PROtein Function, Evolution, Structure and Sequence database." *Database : the journal of biological databases and curation*, p. baq011, Jan. 2010.