Integration of data in biosciences

Vesna Pajić[1], Gordana Pavlović Lažetić[2], Dragana Dudić[1], Dragica Radovanović[1] and Jelena Kozoderović[1]

[1]University of Belgrade, Faculty of Agriculture, Center for Data Mining and Bioinformatics, Nemanjina 6, 11080 Belgrade, Serbia

{svesna, ddragana, macura, jelenak}@agrif.bg.ac.rs

[2]University of Belgrade, Faculty of Mathematics, Studentski trg 16, 11000 Belgrade, Serbia

gordana@matf.bg.ac.rs

Abstract. Experimental data in life sciences, created with today's sequencing techniques, accumulate rapidly in different formats and data sources, leaving facts and knowledge on observed phenomena hidden and unreachable to researchers. The "big data" problem can be seen from various aspects. In this paper, we pointed out to those that can be solved with some sort of data integration. Combining data from different sources and analyzing them jointly can give a complex insight of observed biological phenomena. The problem of designing data integration systems is important in current real world applications. This document presents several levels of data integration, with some examples of good practice.

Keywords: big data, data integration, data mining, data curation, text mining

1 Introduction

With the rise of biotechnology and bioinformatics, a number of problems regarding storing, searching and using biological data occurs. The data are scattered over a large number of repositories (public or private) and stored in different formats. Furthermore, some formats are not adequate for automatic computer processing (such as text documents) and require some kind of preprocessing before they can be input into computer algorithms. This situation makes searching and analyzing the data very difficult, leaving facts and knowledge on observed biological phenomena hidden in databases and digital collections.

The aforementioned and other similar problems can be overcome only through an integrative bioinformatics approach. The integration can be done across different aspects and on various levels: database integration techniques, data acquisition from heterogeneous sources, genotype-phenotype associations researches, integrative modeling and analysis of processes and systems in biosciences, tool integration and workflow systems, computational infrastructure for biological re-

adfa, p. 1, 2011.

searches, biological ontologies and metadata, integrative data and text mining approaches.

While attempting to analyze and interpret the raw biological data, bioinformaticians and computer scientists have developed a huge number of software tools with intention to carry out some tasks in the research process. In this context, hundreds of new tools are published every year.

The tools are more or less specific, intended for one group of analyses or being more general. They also differ in quality, usability, adaptability and robustness.

Sometimes, it is very hard to find an appropriate tool for some research. When searching for a tool, search engines usually search in software documentation. The fact is that majority of software tools is poorly documented, without bug reports and clear directions how to use it. In that case, it is easier to write a new tool than to search for and use an existing one, and this is how yet another specific bioinformatics tool gets developed and added to the bag of tools. Needless to say, many of them will become redundant.

2 The Data Explosion

As a consequence of the revolution in bioinformatics and computational biology, the amount of biological data has increased rapidly. The term "*big data*" is used to express this phenomenon, and although it is not specific just for biology and biosciences, it is a major challenge in bioinformatics and computational biology.

Nowadays, researchers are faced with numerous resources of biological data. It is usually very hard to decide which one to use for a particular task, where to find some information or how to interpret them. For example, the number of publicly available databases, listed at NAR online Molecular Biology Database Collection, reached 1552 [1]. As of January 2014, NCBI Genbank provides access to almost 170 million sequences, with 260000 formally described species [2]. PubMed includes over 23 million citations for biomedical literature from MEDLINE, scientific journals, and online books. The amount of data residing in different laboratories, private computer networks or data warehouses is much greater. Here, we will inspect some of the big data characteristics that turn data processing into quite a challenge. The data and tools integrations can be the way to solve these problems.

3 Big Data Characteristics

3.1 Rapid Data Production

High throughput experiments in life sciences are able to uncover characteristics of thousands of entities in a single experiment. Although they are valuable and inevitable to obtain comprehensive insight into the different biological phenomena, they are typically followed by high levels of noise. Additionally, they

produce large amounts of data. It is estimated that in under a year, genomics technologies will enable individual laboratories to generate terabyte or even petabyte scales of data at a reasonable cost [3]. It would be a big challenge to process and maintain such datasets, and to integrate them with other large-scale data.

3.2 Data Curation

Data curation is a process of selecting, organizing, assessing quality, describing, and updating data, usually done in order to enhance quality of data. It is indispensable in the maintenance of biological databases and an essential task in data integration process. Usually, a number of different strategies to curation are used, including computational or manual curation, or their combination. Appropriate curation, including semantic mark-up, would enable easier finding, maintaining and usage of data [4].

The data curation importance in life sciences is evident from the fact that an entirely new concept, called *biocuration*, has been introduced. Biocuration involves the translation and integration of information relevant to biology into a database or resource that enables integration of the scientific literature as well as large data sets. Accurate and comprehensive representations of biological knowledge, as well as easy access to this data are primary goals of biocuration. In this complex process, it is important to determine answers to key questions, including: For whom the curated resources are? Which kinds of resources are there? Where are the resources? What metadata to curate? How and when to curate?

3.3 Hidden Knowledge

Having a lot of data does not necessarily mean having a lot of knowledge. The data themselves are not useful at all if we cannot interpret them adequately. The development of different data mining techniques is important, in order to discover some facts and knowledge from the data [5]. Classic data mining techniques used for bioinformatics are classification, clustering and association rules. Classification consists of finding patterns that can classify future data into predefined classes. Clustering is employed to partition data into relatively homogenous groups based on their properties from which interesting information may be discovered. Association rules are used to reveal relationships or associations among sets of data items.

3.4 Data Storage

Previously mentioned characteristics of big data cause the problem of data storage. Today's information systems can store a lot of bytes, but the main question is: what data need to be stored, in what formats and for how long. Should we store the raw or processed data? Should we archive the data? There are no clear and definite answers to those questions; they depend on the society (researchers

and users) needs for that particular data. We should keep in mind, though, that today's data mining techniques produce a lot of false positives when applied on large-scale datasets. As they are expected to develop and to decrease the error rate, it would be useful to have raw data, so the experiments can be repeated in the future with more accurate and efficient methods.

3.5 Data Redundancy

What seems to be another problem regarding data storage is the redundancy of the data. Redundancies refer to data which are recorded in more than one database entries due to different data sources, varying views of the proteins (PDB protein structures versus Swiss-Prot protein annotations), or repeated submissions of the sequence by the same or different annotators. Methods for detecting redundant data in biological databases have been proposed in [6-8]. One way to lower the level of redundancy in a database is to use summarization techniques [9].

3.6 Data Heterogeneity

Due to the difference in the nature of observed biological phenomena, methods and tools used in a research, the data produced are very heterogenic. They must be combined so as to obtain a full picture and to build new knowledge. However, current databases do not use a uniform way to name biological entities. There is no unique standard for describing entities, so a same resource is frequently identified with different names. For example, the gene BRCA1 is identified by number "672" in the GDB Human Genome Database and by number "1100" in the HUGO Gene Nomenclature Committee (HGNC) [10]. There were initiatives and proposals of some naming standards (such as [11]) and standardizing data formats, that could solve this problem, but unfortunately they are still not widely adopted by biological data providers.

Moreover, data are heterogeneous due to various biological phenomena they represent. Integration of related data, i.e. data representing related phenomena is a very important part of biological research. For example, the integration of large genomic data sets with other types of information, like gene function, gene interaction or phenotype information can give an insight into more complex biological systems.

3.7 Processing Unstructured Data

Biological data are dispersed across thousands of biological databases and hundreds of scientific journals and encyclopedias. Information located in scientific texts is unstructured and therefore hard to process. On the other hand, scientific journals are the main resources for biological data acquisition and are widely used in the process of biocuration. The acquisition is usually done by human experts, and it is very time-consuming and expensive. Text mining and information

extraction techniques are developed to ease acquisition of data and biocuration and to make them semi or fully automated [12].

3.8 Raw Data and Metadata

The raw data or sets of data should be annotated and described with metadata, in order to preserve their semantic and meaning. Today's annotations are based on different ontologies. The problem with the ontologies is that there are too many of them created and published, but too few widely used. In order to use ontologies (and other annotation schema) at their full potential, concepts, relations and axioms must be shared when possible. Domain ontologies must also be anchored to an upper ontology in order to enable the sharing and the reuse of knowledge. Unfortunately, each bio-ontology seems to be built as an independent piece of information in which every piece of knowledge is completely defined [10].

4 Examples of good practice in data integration

The first level of data integration includes finding, extracting, merging and synthesizing information from multiple sources. With no intention of recommending or advertising any of the researches or tools, we are providing some good examples of data integration that faces the challenges brought by the big biological data.

Since genome sequences and annotations come from various data sources, the several genome browsers were developed with an intention to integrate data from different sources into one common graphical interface. Initially, genome browsers were displaying assemblies of smaller genomes of specific organisms, but today they provide navigation through sequences and simultaneous browsing for genomic annotations and other information on sequences. The most widely used are CGView [13], Combo [14], Ensembl [15,16], Integrated Genome Browser [17] and UCSC Genome Browser [18].

The access to biological data integrated from multiple heterogeneous sources is enabled by web applications and services such as EMBL-EBI web services [19] or NCBI Entrez database [20].

There are several studies that respond to the challenge of associating genotype to phenotype characteristics [21-27]. In [23], cross-organism distribution of phenotypic traits is used for gene function prediction. The association rule mining algorithm netCAR has been developed and applied in order to extract sets of COGs (clusters of orthologous groups of proteins) associated with a phenotype, based on co-occurrence between sets of genes and the phenotype [27]. In [22], genotype-phenotype associations have been systematically discovered by combining information from a biomedical database GIDEON with the molecular information from NCBI COGs database.

The biological information found in the literature is very valuable, especially from the perspective of biocuration. Techniques used for literature mining are providing access and retrieval of the most up-to-date biological data from scientific articles and reports on ongoing researches. They are usually based on text mining and information extraction. Korbel et al. [25] reported on systematic association of genes to phenotypes by literature mining and comparative genome analysis. In [12] we can see a database of genotype and phenotype data on microbes, created completely with text mining and information extraction techniques. Furthermore, various tools for assisting curation of biological databases from biomedical literature have been developed. PubTator [28] features a PubMed-like interface and uses multiple text mining algorithms to ensure the quality of its automatic results. In [29] the list of dictionary-based text mining tools that could be used for biocuration tasks is given, with the evaluation metrics for each tool.

Bio-ontologies can be used for providing the semantic service in data integration from different sources. Gene Ontology [30] is the most widely used ontology in bioinformatics with main goal to standardize gene representation among species and data repositories. GOA project [31] is one of the first examples of biological data integration. This project aims to provide Gene Ontology annotations to proteins in the UniProt [32] knowledgebase. Kumar et al. [33] emphasized need for integration and enrichment of biomedical ontologies with different biological and medical information sources and showed an example of connecting colorectal carcinoma data to the molecular level. The paper by Sahoo and coworkers [34] considered how information integration can be supported with Semantic Web technologies.

Finally, the data warehouse approach is very successful in providing a robust information infrastructure for biological data. MoDa (Molecular Data warehouse) [35] provides a unified framework for finding and visualizing results of various experimental techniques of molecular biology, with warehouse architecture optimized for various types of filtering and querying annotations of samples, experimental results and properties of genes and other molecular entities. Atlas [36] is a biological data warehouse that stores data of similar types using common data models, enforcing the relationships between data types. It integrates and locally stores biological sequences, molecular interactions, homology information, functional annotations of genes, and biological ontologies, providing data and a software infrastructure for bioinformatics research and development.

5 Conclusion

In this paper we presented some of the successful work on data integration. This list is far from being complete. A lot of problem remains, with new ones being made continuously.

Bioinformatics itself is a multidisciplinary scientific field, based on the integration of data, tools and techniques from other sciences. Although the mixture of

data and processes in bioinformatics are often complex, it is the only way to give answers to the raising questions and to understand better the life itself. For successful integration of data, more strict standards of data annotations, ontologies, database architectures and similar data-related issues should be adopted from a broader research community.

References

1. Fernández-Suárez, X.M., Rigden, D.J., Galperin, M.Y.: The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection, Nucl. Acids Res. 42 (D1): D1-D6 (2014)
2. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/genbank/statistics
3. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., Nolan. G. P.: Computational solutions to large-scale data management and analysis. Nat Rev Genet, 11(9):647–657, (2010)
4. Goble, C., Stevens, R., Hull, D., Wolstencroft, K., Lopez, R: Data curation + process curation = data integration + science, Briefings in Bioinformatics, 9(6):506-517, (2008)
5. Wang, J.T.L., Zaki, M.J., Toivonen,, H.T.T., Shasha, D.: Data Mining in Bioinformatics, (Advanced Information and Knowledge Processing) . Springer, pp. 3-8 (2005)
6. Chellamuthu, S., Punithavalli, D. M.: Detecting Redundancy in Biological Databases? An Efficient Approach. Global Journal of Computer Science and Technology, 9(4) (2009)
7. Apiletti, D., Bruno, G., Ficarra, E., and Baralis, E.: Data cleaning and semantic improvement in biological databases. Journal of Integrative Bioinformatics, 3(2):40 (2006)
8. Koh, J. L. Y., Lee, M. L., Khan, A. M., Tan, P. T., Brusic, V.: Duplicate detection in biological data using association rule mining. Locus, 501(P34180), S22388 (2004)
9. Sorani, M. D., Ortmann, W. A., Bierwagen, E. P., Behrens, T. W: Clinical and biological data integration for biomarker discovery. Drug discovery today, 15(17):741-8 (2010).
10. Pasquier, C. : Biological data integration using Semantic Web technologies. Biochimie 2, 90 (4): 584–94 (2008)
11. Clark, T., Martin, S., Liefeld, T.: Globally distributed object identification for biological knowledgebases. Briefings in Bioinformatics, 5: 59-70, (2004)
12. Pajić, V.S., Pavlović-Lažetić, G.M., Brandt, B.W., Pajić, M.B.: Towards a database for genotype-phenotype association research: mining data from encyclopaedia. Int. J. Data Mining and Bioinformatics, 7(2): 196-213 (2013)
13. Grant J.R., Stothard P : The CGView server: a comparative genomics tool for circular genomes. Nucleic Acids Res , 36 (Web Server issue):W181-184 (2008)
14. Engels R., Yu T., Burge C., Mesirov J.P., DeCaprio D., Galagan J.E.: Combo: a whole genome comparative browser. Bioinformatics , 22(14):1782-1783 (2006)
15. Flicek P., Amode M.R., Barrell D., Beal K., Brent S., Carvalho-Silva D., Clapham P., Coates G,. Fairley S., Fitzgerald S., et al.: Ensembl 2012. Nucleic Acids Res 2012, 40(Database issue):D84-90 (2012)
16. Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., et al.: The ensembl genome database project. Nucleic Acids Res, 30(1): 38-41 (2002)
17. Nicol J.W., Helt G.A., Blanchard S.G. Jr, Raja A., Loraine A.E.: The integrated genome browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics, 25(20):2730-2731 (2009)

18. Karolchik D., Barber G.P., Casper J., Clawson H., Cline M.S., Diekhans M., Dreszer T.R., Fujita P.A., Guruvadoo L., Haeussler M., Harte R.A., Heitner S., Hinrichs A.S., Learned K., Lee B.T., Li C.H., Raney B.J., Rhead B., Rosenbloom K.R., Sloan C.A., Speir M.L., Zweig A.S., Haussler D., Kuhn R.M., Kent W.J.: The UCSC Genome Browser database: 2014 update. Nucleic Acids Res., 42(1):D764-70. Epub 2013 Nov 21 (2013)
19. The European Bioinformatics Institute, http://www.ebi.ac.uk/services
20. National Centre for Biotechnology Information, http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html
21. Chang, R., Shoemaker, R. and Wang, W.: A novel knowledge-driven systems biology approach for phenotype prediction upon genetic Intervention. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(5):1170–1182 (2011)
22. Goh, C.S., Gianoulis, T.A., Liu, Y., Li, J., Paccanaro, A., Lussier, Y.A. and Gerstein, M.: Integration of curated databases to identify genotype-phenotype associations. BMC Genomics, 7:.257–257 (2006)
23. Jim K. Parmr K., Singh M., Tavazoie S. : A Cross-Genomic Approch for Systematic Mapping of Phenotypic Traits to Genes. Genome Research, 14(1):109-115 (2014)
24. Jimeno-Yepes, A., Berlanga-Llavori, R., Rebholz-Schuhmann, D.: Exploitation of ontological resources for scientific literature analysis: Searching genes and related diseases. In: Engineering in Medicine and Biology Society, EMBC 2009, Annual International Conference of the IEEE, pp. 7073–7078 (2009)
25. Korbel, J. , Doerks, T., Jensen, L. J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S. D., Andrade, M. A., Bork, P.: Systematic association of genes to phenotypes by genome and literature mining. PLoS Biol, 3:134-134 (2005)
26. MacDonald, N. J., Beiko, R. G.: Efficient learning of microbial genotype-phenotype association rules. Bioinformatics, 26(15):1834-1840 (2010)
27. Tamura, M., D'haeseleer P. : Microbial genotype-phenotype mapping by class association rule mining. Bioinformatics, 24 (13): 1523 -1529 (2008)
28. Wei C.H., Kao H.Y., Lu Z.: PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W518-22. doi: 10.1093/nar/gkt441. (2013)
29. Dowell, K. G., McAndrews-Hill, M. S., Hill, D. P., Drabkin, H. J., Blake, J. A.: Integrating text mining into the MGI biocuration workflow. Database: The Journal of bi-ological databases and curation, (2009).
30. Gene Ontology, http://www.geneontology.org/
31. GOA project, http://www.ebi.ac.uk/GOA
32. Apweiler R., Bairoch A., Wu C.H., Barker W.C., et al.: UniProt: the Uni-versal Protein knowledgebase, Nucleic Acids Res. 32: D115–D119, (2004).
33. Kumar, A., Yip, Y.L., Smith, B., Grenon P.: Bridging the gap between medical and bioinformatics: an ontological case study in colon carcinoma. Comput Biol Med. 36(7-8):694-711, (2006).
34. Sahoo, S.S., Bodenreider, O., Rutter, J.L., Skinner, K.J., Sheth, A.P., An ontol-ogy-driven semantic mash-up of gene and biological pathway information: applica-tion to the domain of nicotine dependence, J Biomed Informatics, 41:752–76, (2008).
35. Neogi, S. G., Krestyaninova, M., Kapushesky, M., Emam, I., & Brazma, A.: MoDa-A Data Warehouse for Multi-"Omics" Data. J Data Mining Genomics Proteomics, 4(145), 2153-0602. (2013).
36. Shah, S. P., Huang, Y., Xu, T., Yuen, M. M., Ling, J., Ouellette, B. F.: Atlas–a data warehouse for integrative bioinformatics. BMC bioinformatics, 6(1), 34. (2005).