machine learning & knowledge extraction

MDPI

Opinion

Exploiting Genomic Relations in Big Data Repositories by Graph-Based Search Methods

Aliyu Musa [1,2], Matthias Dehmer [3,4,5], Olli Yli-Harja [2,6,7] and Frank Emmert-Streib [1,2,*]

1 Predictive Medicine and Data Analytics Lab, Department of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland; aliyu.musa@tut.fi
2 Institute of Biosciences and Medical Technology, 33520 Tampere, Finland; olli.yli-harja@tut.fi
3 Department of Mechatronics and Biomedical Computer Science, UMIT, 6060 Hall in Tyrol, Austria; matthias.dehmer@umit.at
4 College of Computer and Control Engineering, Nankai University, Tianjin 300071, China
5 Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, 4400 Steyr Campus, Austria
6 Computational Systems Biology Lab, Tampere University of Technology, 33720 Tampere, Finland
7 Institute for Systems Biology, Seattle, WA 98109, USA
* Correspondence: v@bio-complexity.com; Tel.: +358-50-301-5353

Received: 26 September 2018; Accepted: 21 November 2018; Published: 22 November 2018

check for updates

Abstract: We are living at a time that allows the generation of mass data in almost any field of science. For instance, in pharmacogenomics, there exist a number of big data repositories, e.g., the Library of Integrated Network-based Cellular Signatures (LINCS) that provide millions of measurements on the genomics level. However, to translate these data into meaningful information, the data need to be analyzable. The first step for such an analysis is the deliberate selection of subsets of raw data for studying dedicated research questions. Unfortunately, this is a non-trivial problem when millions of individual data files are available with an intricate connection structure induced by experimental dependencies. In this paper, we argue for the need to introduce such search capabilities for big genomics data repositories with a specific discussion about LINCS. Specifically, we suggest the introduction of *smart interfaces* allowing the exploitation of the connections among individual raw data files, giving raise to a network structure, by graph-based searches.

Keywords: knowledge extraction; computational pharmacogenomics; systems pharmacogenomics; network science; computational biology; genomics; big data; databases

1. Introduction

In the last 20 years, technological progress in high-throughput assays, e.g., next-generation sequencing, led to a tremendous increase of our data generation capabilities in genomics. As a result, there are many data collections available providing millions of data points about DNA sequence, gene expression, metabolic, protein structure or protein interaction data [1]. However, to reveal the information buried within these data collections, such data need to be analyzable [2]. The problem is that accessing *selected subsets* of these "big data" for performing a dedicated analysis is non-trivial due to the sheer number of data and, more importantly, the complexity of the connections between different data points. Unfortunately, most data collections do not provide efficient interfaces enabling a direct access to subsets of *raw data*, thus hampering downstream analysis.

For reasons of clarity, we would like to highlight that, here, we are concerned with accessing and selecting *raw data*, not knowledge that has been derived by processing and analyzing raw data and stored in *knowledge databases*. Instead, the data repositories we are concerned with in our paper, store raw data files (see Figure 1 for a brief overview). In the following, we discuss this problem by focusing on the pharmacogenomic data repository LINCS (Library of Integrated Network-based Cellular Signatures) [3–8] and describe how this lack in querying capability could be compensated.

2. Preliminaries

Before we discuss the problem under consideration, we would like to clarify a couple of terms used throughout the paper. We use the term *data repository* for a very general collection and storage of individual data files without providing any dedicated accessing or searching capabilities. Sometimes, this may also be referred to as a data library. Here, by lack of *dedicated* accessing and searching capabilities, we mean that information about data files can in principle be searched but in an inefficient way, which may be as simple as a manual browsing of the data.

In contrast, we use the term database to refer to an *organized* collection and storage of data for which a database management system (DBMS) is available that allows querying the data from the database. The term database system refers to the combination of a database with a DBMS. Here, the term "organized" refers to a specific type of a database, e.g., a relational database or object-oriented database.

It is important to note that each type of data organization (data repository, database system, etc.) comes with its own characteristics. Interestingly, the conceptual idea discussed in the following does not fit nicely into any of these well-known, existing categories, but is situated between them, extending and modifying characteristics thereof.

3. The Pharmacogenomics Data Repository LINCS

The LINCS data repository is supported by the NIH (National Institute of Health), comprising 5000 genetic perturbagens (e.g., single-gene knockdowns or overexpressions) and 15, 000 perturbagens induced by chemical compounds (e.g., drugs) [9]. To date, almost two million gene expressions have been profiled using the L1000 technology [9]. Specifically, the L1000 technology measures the expression of only 978 so-called landmark genes, and the expression values for the remaining genes are estimated by a computational model using additional data from the Gene Expression Omnibus (GEO) [10]. Access to the raw data is provided by GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742) but, unfortunately, there is no search functionality provided other than to select all raw files for download. For this reason, LINCS is merely a collection of files usually called a data repository.

This particular example of LINCS described above is typical for the current situation of many big *data repositories*. Here, we want to emphasize that a data repository is not a database itself. Instead, it stands for a more generalized term that indicates the lack of basic functionality usually present within a database yet providing data storage capabilities. In our context, the crucial lack of functionality is the limited capability to provide efficient ways to query the data within the data repository for selecting and downloading subsets of the data (files).

To rectify this problem, in our opinion, big data repositories need functionality we summarize by the term *smart interfaces*. We envision a smart interface as a web interface that enables extensive selection capabilities, providing many features for querying, exploration, downloading and analyzing data and related meta information. It would also support programmatic access via API as a search functionality to all the attributes contained within the data repository. Using the API, computational scientists and developers can access the data and build flexible research pipelines. Given the genomic context of LINCS and related data repositories, the data queries can utilize the dependency structure between individual data files as implied by, for instance, experimental or biological conditions (see example below). Hence, queries perform network or graph-based searches within the data repositories exploiting in this way the existing dependency structure between the individual data files.

> Smart interfaces: A web interface enabling extensive selection capabilities of raw data, providing features for graph-based querying, exploration, downloading and analyzing data and related meta information.

In Figure 1, we show a visualization of our idea. A smart interface exploits the connectivity structure among the raw data files (see Figure 1A), which can also include metadata if available, by generating a network representation among the individual data files (see Figure 1B). In the example of the LINCS database, these connections are given by the combination of cell lines, drugs, dosages of drugs, etc. for which gene expression profiles have been generated. In general, these correspond to the attributes of the data files. Importantly, these attributes remain constant and do not change if more data points are added to a database. Once such a network representation among the data files is generated, a user query extracts quickly the desired data files, e.g., that correspond to cell line C2, the drugs D1 and D3 and the dosage Do3 (see Figure 1B and its connection back shown in yellow to the data files), because each search combination connects to a list of associated data files. In this way, a smart interface forms a connection between the data repository and the preprocessing and analysis of the data (see Figure 1C) and its purpose is to provide a graphical-user-interface and query function for an efficient access to selected data files. We want to emphasize that the network representation should be part of the smart interface because, in this way, it would be easily applicable for practitioners such as biologists or clinicians.

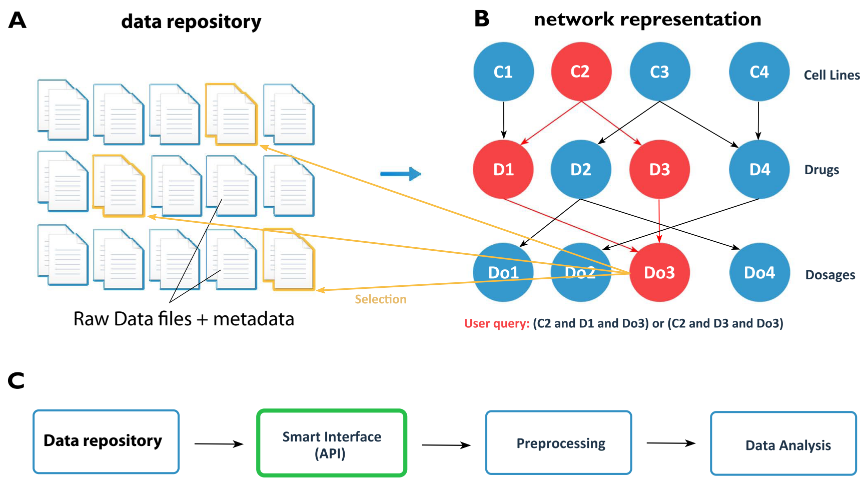

Figure 1. (**A**) A collection of available individual (raw) data files and metadata; (**B**) Network representation of connections between the raw data files. A user query (in red) corresponds to one particular combination of attributes of the data files, which leads to an efficient selection of these (in yellow); (**C**) Conceptual integration of the smart interface, which is an application programming interface (API), into a conventional data analysis pipeline.

4. Technical Considerations

On a technical note, we would like to point out that, here, we focus on a data representation that would allow users to immediately interact with the data. Through the smart interface, users can perform highly specialized queries using attributes that naturally connect the individual data files. The queries can be executed in the web browser or programmatically from the interface. This could be achieved by using modern generalizations of relational databases [11], e.g., NoSQL [12] or graph [13] databases, to efficiently store the data for quick access. Unfortunately, non-relational databases have been naturally fragmented by usage and have drawbacks in scaling, resulting in relatively slow progress in integrating large datasets [14]. However, for genomics problems with a constant number of attributes, e.g., cell lines, drugs, dosages, etc, as is the case for the LINCS data (see below), the known scaling problems of graph databases do not hamper their usage because new data points do not lead to a change in the number of attributes and, hence, the database can grow efficiently in the number of stored data points. An example of this was given by Himmelstein et al. [15] using a graph database for integrating information from 29 public resources to connect compounds, diseases, genes, pharmacologic classes, side effects, etc., which helped to identified network patterns that distinguish treatments from non-treatments drugs [15]. We would like to point out that the result of [15], and similar approaches [16–18], is a knowledge database that operates on processed and analyzed data, not on the raw data files as is our major concern in this paper.

For the LINCS data, one can start from a set of files and select certain attributes to create a network representation by using graph algorithms [19]. This is similar to classical contributions focusing on data and retrieval based on graph theoretical considerations [20,21], which do technically not fall within the strict definition of databases because they are lacking the consideration of database management systems as the most important building block when one refers to the term database. Hence, more research is required to identify if a database structure, an information retrieval system [22,23] or a graph-based file organization system [20,21] provides the most appropriate technical realization for graph-based searches of data repositories in genomics.

5. Conceptual Idea

The general idea of a smart interface is similar to the idea behind Google. If one considers "web sites" as "data files" and realizes that the "connections between web sites" are implicitly provided by the "attributes of data files" (see the example above), then the analogy is apparent. Specifically, Google identifies the connections between web sites by searching the links from and to sites by crawling the web. This establishes a graph structure between the web sites corresponding to a very large network upon which graph-based searches that take user queries into account can be executed.

In the case the world-wide-web (WWW) would consist of only a dozen web sites, there would be no need for a search engine such as Google because a user could quickly go through the list of these web sites manually. However, for billions of web sites, this is no longer feasible (even if such a list would exist) because a linear search would lead to exponential searching times. Interestingly, this is exactly the situation we are facing for data repositories such as LINCS. While the current raw organization of LINCS or similar data repositories is sufficient for certain tasks, it does not favor the selection of complexly determined subsets, such as those required for more advanced or specialized studies. Any additional tool, such as a smart interface, that can be added to facilitate such complex queries, would make these repositories more useful, efficient and popular. This would not only benefit users, but also the repositories themselves by reducing work, reducing download costs and increasing their impact, usage and user satisfaction. This implies also that the true potential of LINCS is currently not yet unlocked due to this limitation. For technical completeness we would like to note that Google uses a NoSQL database of columnar type called BigTable [24].

6. Further Applications

We would like to note that our idea extends beyond the LINCS data repository. Other examples of raw data repositories that would benefit from a similar approach are:

- Gene Expression Omnibus (GEO) [1]
- NCI60 human tumour cell line anticancer drug screen [25]
- ArrayExpress [26]
- Cancer Cell Line Encyclopedia [27]

However, the largest benefit for the community would result from the integration of some (or all) such data repositories to address the problems by a systems biology approach taking holistically all aspects into account. We expect that the smart interface needs to be adapted to the specific characteristics of the data types in the corresponding data repositories but the conceptual core idea would be generic to all these different repositories.

In our opinion, the implementation costs would be rather limited because it only requires a software solution. However, the intellectual costs are considerable because the creation of graph-based relations among the individual data files requires familiarity with basic graph-theoretical concepts and graph-search methods [19,28].

7. Conclusions

The transition from simple data repositories to big pharmacological warehouses requires new forms of data accessing strategies and we think that smart interfaces, enabling graph-based querying capabilities, provide the needed functionalities. While current repositories offer the possibility to mirror data to access it in a local implementation, it would carry unduly efforts and costs for most users, many of whom would not be able to do it (and hence to benefit from the data), and this cost would be best addressed if data repositories offering advanced search technologies were available, whether at the primary curation site, or at a separate publicly accessible resource, or both. Otherwise, the opportunities offered by these big data cannot be translated into new knowledge by means of modern data science [29]. We discussed our idea for the LINCS data repository and provided a specific outline of the graph structure induced by the available data files. However, our idea is not limited to LINCS, but we selected this data repository because of its popularity to emphasize the need for such search capabilities.

Finally, we would like to note that, in our presentation, we focused on the network-based search capabilities and neglected many other data aspects of practical relevance, e.g., data privacy, data quality, etc, to convey a clear message. However, we do not want to miss emphasizing that these aspects are also part of the data analysis pipeline (see Figure 1C) that needs to be integrated into our framework to obtain a functional implementation.

Author Contributions: F.E.S. and A.M. conceived the study. All authors wrote the paper and approved the final version.

Funding: AM thanks the CIMO foundation of Finland for a scholarship. MD thanks the Austrian Science Funds for supporting this work (project P30031).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210. [CrossRef] [PubMed]
2. Holzinger, A.; Jurisica, I. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*; Springer: Berlin, Germany, 2014; pp. 1–18.
3. Lamb, J.; Crawford, E.D.; Peck, D.; Modell, J.W.; Blat, I.C.; Wrobel, M.J.; Lerner, J.; Brunet, J.P.; Subramanian, A.; Ross, K.N.; et al. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313*, 1929–1935. [CrossRef] [PubMed]
4. Ma'ayan, A.; Rouillard, A.; Clark, N.; Wang, Z.; Duan, Q.; Kou, Y. Lean Big Data integration in systems biology and systems pharmacology. *Trends Pharmacol. Sci.* **2014**, *35*, 450–460. [CrossRef] [PubMed]
5. Campillos, M.; Kuhn, M.; Gavin, A.C.; Jensen, L.J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266. [CrossRef] [PubMed]
6. Subramanian, A.; Narayan, R.; Corsello, S.M.; Peck, D.D.; Natoli, T.E.; Lu, X.; Gould, J.; Davis, J.F.; Tubelli, A.A.; Asiedu, J.K.; et al. A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. *BioRxiv* **2017**. [CrossRef] [PubMed]
7. Musa, A.; Ghoraie, L.; Zhang, S.D.; Glazko, G.; Yli-Harja, O.; Dehmer, M.; Haibe-Kains, B.; Emmert-Streib, F. A Review of Connectivity Mapping and Computational Approaches in Pharmacogenomics. *Brief. Bioinform.* **2017**, *19*, 506–523.

8. Musa, A.; Tripathi, S.; Kandhavelu, M.; Dehmer, M.; Emmert-Streib, F. Harnessing the biological complexity of Big Data from LINCS gene expression signatures. *PLoS ONE* **2018**, *13*, e0201937. [CrossRef] [PubMed]
9. Vidovic, D.; A, K.; Schurer, S. Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. *Front. Genet.* **2014**, *5*, 342. [PubMed]
10. Barrett, T.; Troup, D.B.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; et al. NCBI GEO: Archive for functional genomics data sets -10 years on. *Nucleic Acids Res.* **2011**, *39*, D1005–D1010. [CrossRef] [PubMed]
11. Codd, E.F. A Relational Model of Data for Large Shared Data Banks. *Commun. ACM* **1970**, *13*, 377–387. [CrossRef]
12. Wiese, L. *Advanced Data Management: For SQL, NoSQL, Cloud and Distributed Databases*; De Gruyter: Berlin, Germany, 2015.
13. Angles, R.; Gutierrez, C. Survey of Graph Database Models. *ACM Comput. Surv.* **2008**, *40*, 1–39. [CrossRef]
14. Zou, L.; Chen, L.; Özsu, M.T. Distance-join: Pattern match query in a large graph database. *Proc. VLDB Endowment* **2009**, *2*, 886–897. [CrossRef]
15. Himmelstein, D.S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S.L.; Hadley, D.; Green, A.; Khankhanian, P.; Baranzini, S.E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **2017**, *6*, e26726. [CrossRef] [PubMed]
16. Matthews, L.; Gopinath, G.; Gillespie, M.; Caudy, M.; Croft, D.; de Bono, B.; Garapati, P.; Hemish, J.; Hermjakob, H.; Jassal, B.; et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **2009**, *37*, D619–D622. [CrossRef] [PubMed]
17. Swainston, N.; Batista-Navarro, R.; Carbonell, P.; Dobson, P.D.; Dunstan, M.; Jervis, A.J.; Vinaixa, M.; Williams, A.R.; Ananiadou, S.; Faulon, J.L.; et al. biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PLoS ONE* **2017**, *12*, 1–14. [CrossRef] [PubMed]
18. Touré, V.; Mazein, A.; Waltemath, D.; Balaur, I.; Saqi, M.; Henkel, R.; Pellet, J.; Auffray, C. STON: Exploring biological pathways using the SBGN standard and graph databases. *BMC Bioinform.* **2016**, *17*, 494. [CrossRef] [PubMed]
19. Cormen, T.; Leiserson, C.; Rivest, R.; Stein, C. *Introduction to Algorithms*; MIT Press: Cambridge, MA, USA, 2001.
20. Lipski, W.; Marek, W., File organization, an application of graph theory. In *Automata, Languages and Programming: 2nd Colloquium, University of Saarbrücken 29 July– 2 August 1974*; Loeckx, J., Ed.; Springer: Berlin/Heidelberg, Germany, 1974; pp. 270–279.
21. Lipski, W. Information storage and retrieval ? mathematical foundations II (combinatorial problems). *Theor. Comput. Sci.* **1976**, *3*, 183 – 211. [CrossRef]
22. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*; ACM Press: New York, NU, USA, 1999; Volume 463.
23. Chowdhury, G.G. *Introduction to Modern Information Retrieval*; Facet Publishing: London, UK, 2010.
24. Chang, F.; Dean, J.; Ghemawat, S.; Hsieh, W.C.; Wallach, D.A.; Burrows, M.; Chandra, T.; Fikes, A.; Gruber, R.E. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst.* **2008**, *26*, 4. [CrossRef]
25. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **2006**, *6*, 813–823. [CrossRef] [PubMed]
26. Brazma, A.; Parkinson, H.; Sarkans, U.; Shojatalab, M.; Vilo, J.; Abeygunawardena, N.; Holloway, E.; Kapushesky, M.; Kemmeren, P.; Lara, G.G.; et al.. ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **2003**, *31*, 68–71. [CrossRef] [PubMed]
27. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehár, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–607. [CrossRef] [PubMed]
28. Dehmer, M.; Emmert-Streib, F., Eds. *Analysis of Complex Networks: From Biology to Linguistics*; Wiley-VCH: Weinheim, Germany, 2009.
29. Emmert-Streib, F.; Moutari, S.; Dehmer, M. The process of analyzing data is the emergent feature of data science. *Front. Genet.* **2016**, *7*, 12. [CrossRef] [PubMed]