Mining_RNA: WEB-Based System Using e-Science for Transcriptomic Data Mining

Carlos Renan Moreira, Christina Pacheco, Marcos Vinícius Pereira Diógenes, Pedro Victor Morais Batista, Pedro Fernandes Ribeiro Neto, Adriano Gomes da Silva, Stela Mirla da Silva Felipe, Vânia Marilande Ceccatto, Raquel Martins de Freitas, Thalia Katiane Sampaio Gurgel, Exlley Clemente dos Santos, Cynthia Moreira Maia, Thiago Alefy Almeida e Sousa, and Cicília Raquel Maia Leite

1 Introduction

Since the conclusion of the Human Genome Project in 2003, it was made possible that several studies could be carried out in the search for the cure of various diseases [1]. The use of bioinformatics contributed to the advance in the field and allowed the production of research in an accelerated pace. Currently, bioinformatics helps in the processing of huge amounts of results [2] and can also be used to perform data mining aiming at the extraction of relevant information that can help in the treatment or diagnosis of diseases [3]. The need to publish the huge amount of raw data from these studies led to the development of repositories where data could be made available in order to confirm the results or for further researches and may also allow for the inference of new information from the compilation of results from more than one study [4].

High-throughput gene expression (RNA) data, from microarray and RNA sequencing (RNA-Seq) studies, are mainly stored in three public databases: Gene

C. Renan Moreira (✉) · C. Pacheco · M. V. Pereira Diógenes · P. V. Morais Batista · P. F. Ribeiro Neto · A. G. da Silva · T. K. Sampaio Gurgel · E. C. dos Santos · C. Moreira Maia · T. A. Almeida e Sousa · C. R. Maia Leite
UERN, Mossoró, RN, Brasil
e-mail: marcosdiogenes@alu.uern.br

S. M. da Silva Felipe · V. Marilande Ceccatto · R. M. de Freitas
UECE, Fortaleza, CE, Brasil
e-mail: vania.ceccatto@uece.br

H. R. Arabnia et al. (eds.), *Advances in Parallel & Distributed Processing, and Applications*, Transactions on Computational Science and Computational Intelligence, https://doi.org/10.1007/978-3-030-69984-0_85

Expression Omnibus[1] (GEO), ArrayExpress[2] (AE), and Genomic Expression Archive[3] (GEA). Although there is a centralization of this information in the abovementioned repositories, there is still some difficulty in conducting studies in the research metadata. This is due to the lack of standardization in the vocabulary employed by the authors of the researches to structure the data that will be stored in the databases [5]. Many researchers use algorithms written in the R programming language for the analysis of this data; however this can be a problematic factor since it is not simple for the life science research community to use or write codes in some programming language. Bioconductor[4] is a repository that provides several tools for this purpose [6]; however the lack of programming skills may drive some researchers away.

To mitigate this situation, the present work proposes, through the application of an e-Science approach, the development of a WEB system is capable of reading a massive amount of data, pre-processing, mining, and displaying it in a user-friendly interface, intending to aid researchers to delve further into the existing research. The architecture and functionalities proposed for this system and the preliminary results of the research will also be demonstrated.

2 e-Science and Biological Data Mining

With the amount of raw data constantly been generated, it would be a complicated task to process this data manually. Being able to combine the raw microarray analysis data from two or more studies would result in a huge effort on the part of the researchers. In order to facilitate such analyses, the scientists can make use of bioinformatics, using e-Science to process a massive amount of data, and can also make use of data mining techniques to be able to extract new information from the pre-existing surveys.

The use of e-Science is justified by offering the researcher a platform capable of storing, interpreting, analyzing, and making available in network this data to enable other working groups to make use of this. With the use of the technique, it will be possible to deal with large volumes of data in a properly scalable system. This will enable the accomplishment of research in a quick and efficient manner, also enabling joint and multidisciplinary studies in the context of bioinformatics.

One of the approaches that can be used in bioinformatics is the mining of biological data [7]. Data mining uses algorithms in the search for valid and understandable patterns within the available data and includes the following steps: associating, grouping, and discovery of classification rules [8]. It is a fundamental

[1] https://www.ncbi.nlm.nih.gov/geo.

[2] https://www.ebi.ac.uk/arrayexpress/.

[3] https://www.ddbj.nig.ac.jp/gea/index-e.html.

[4] https://www.bioconductor.org/.

technique in bioinformatics because it allows researchers not to be mere observers of the available platforms (e.g., ENSEMBL and UCSC Genome Browser) and explore the information contained within the data [3]. Data mining can help transform data that is not easily understandable through a purely visual interpretation and can also facilitate the inference of new and useful information for the discovery of new results from new or pre-existing research [9]. The new results from the existing research could be achieved through the researchers' new approach to the data previously analyzed in another study. Often these datasets are deposited in public biological databases and lay there almost forgotten, and new analyses could provide advances in their respective research areas [10].

Currently one of the possible ways to conduct biological data mining is through bioconductor [6], a project based on the premise of free software that aims to promote statistical analysis and understanding of high-throughput data from new and pre-existing biological studies. The project is based on packages written largely using the R programming language; however it may contain contributions from other languages. Bioconductor is one of the most relevant repositories of tools for the study of biological data; the packages made available are destined toward various types of analyses, among them, data mining.

3 Description and Context of the Proposed System

In recent years, data from scientific research in the biological area have been made available in public databases dedicated to the storage and display of this information [11]. At a very similar rate, computational tools capable of interpreting biological data have been emerging in recent decades in order to evaluate the data and produce new results [5].

One of the most popular transcriptomic databases in the scientific community is Gene Expression Omnibus. Its platform has almost 4,500 studies stored in the form of datasets, from researches that analyzed almost 3.5 million samples to generate their results [5]. Currently, the way the platform makes this data available is not easily manageable for users without appropriate technical knowledge for reading information in plain text files, having their data divided online and separated by tabs. Therefore, it is necessary to use tools that can help in this setback so that the researcher can dedicate his effort to the data and not to understanding the storage model.

This research intends to implement a system capable of providing a service that recovers data from the GEO platform; this data will be pre-processed so that the information from the platform is validated in order to pass correctly to the later steps that may include, according to the user's preference, mining of this data or manual processing of information from filters made available in the system being developed.

During the development of the proposed system, as well as in the subsequent stages, after implementation, the objective is to achieve the following results:

1. Implement an Application Programming Interface (API) that allows interaction between the GEO biological database and a local database.
2. Develop an interface for pre-processing obtained data.
3. Implement data mining and machine learning techniques to act upon the pre-processed data.
4. Develop an analytical interface and the interaction between the researcher and the studies' data.
5. Allow that researchers without computer programming knowledge can utilize data from the biological databases.
6. Provide the possibility to extract new information from raw data made publicly available in the GEO biological database.
7. Develop a system that can be used in the daily life of researchers in the area of biological and health sciences.

To achieve the abovementioned results, we intend to develop a WEB system that will be available full-time in a domain on the Internet, thus enabling researchers from all over the world to use it. Initially, it is intended that the system be available in two languages, Portuguese and English. The use of the system should follow the step-by-step style where the researcher should advance between the screens until he reaches the listing of the filtered data during the previous steps.

With the data adequately filtered and the results exposed to the researcher, the system will provide a series of filters aimed at the selection of useful data, to try to identify relevant information aiming for new scientific discoveries. Among the metrics that can be used in the analysis of the results is fold change, this metric evaluates how much higher or lower a given gene was expressed by comparing samples from two individuals or two groups of patients within a given research.

Initially the tests and filters will be applied in studies related to diabetes from the biological database GEO. In this initial stage, compatibility with studies that used the DNA microarray technique will be offered.

4 System Architecture

As can be observed in Fig. 1, the architecture planned for the system will have three distinct APIs, a database, and a WEB interface in which users will be able to query the data made available by the system. This division into distinct APIs and interfaces was motivated by the possibility of allocating the algorithms on different servers if necessary. Some algorithms that are part of the APIs can be computationally complex, and this can lead to a very high processing load. In order to ease the process, aiming for everything to be better tiered, the decentralized architecture presented itself as an efficient solution.

API 01 will be responsible for the tasks related to mining the obtained data, as well as the machine learning tasks. It is intended that this API will support the implementation of several mining algorithms that will be useful for analyzing the

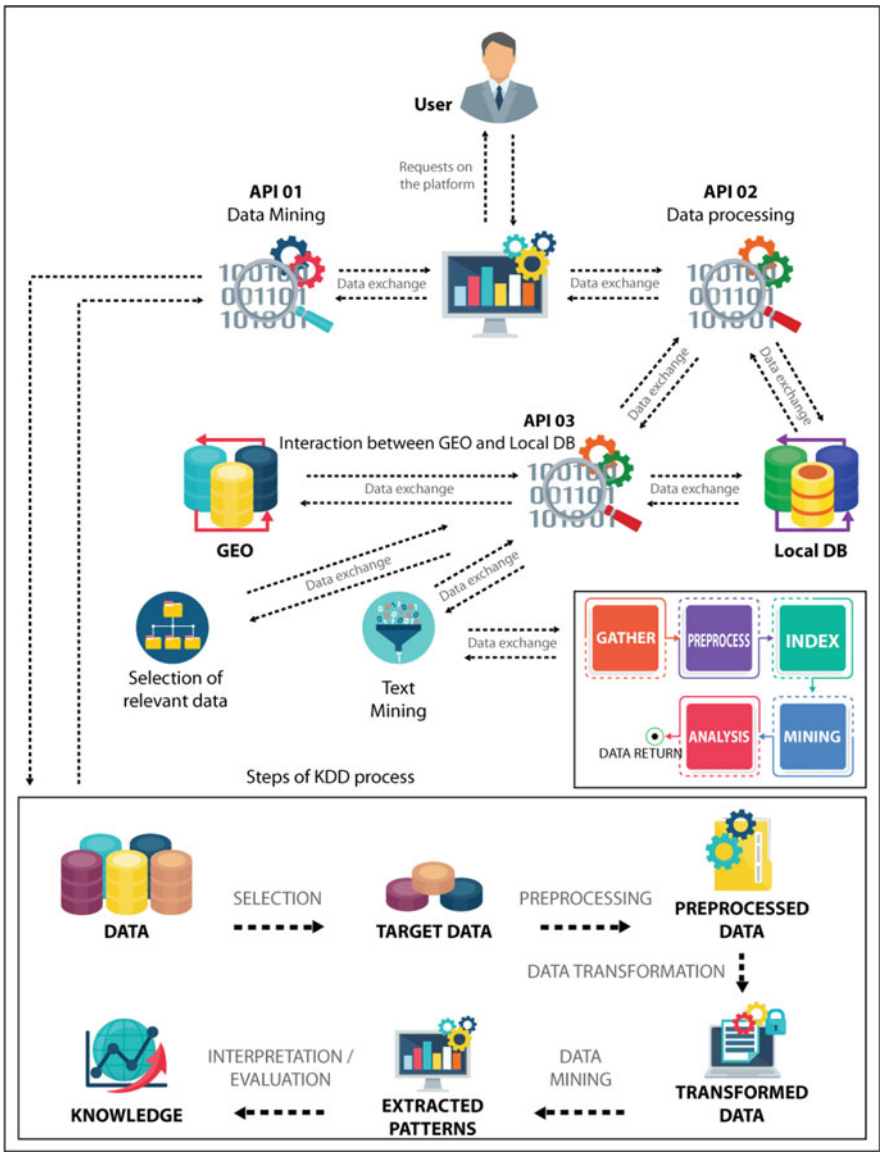

Fig. 1 System architecture overview. Source: The authors

data of the study explored by the scientist. The algorithms should be implemented gradually, thus supporting improvements even after the completion of the initial phase of the project.

API 02 will be responsible for pre-processing the data that will be uploaded to the WEB system; the API in question can also load data directly from the local data pool, as well as request API 03 to search for new data in the GEO biological database. The development of this API must strictly respect principles related to reuse so that new filters can be inserted into future implementations of the project.

API 03 will capture the data from the dataset requested by the user and store the relevant information in the local database. Although it is a simple task, sometimes considerable processing power is required because the data from some searches is too complex, with a lot of information. The API will interpret all the data and store it so that later processing steps are faster.

The WEB system will include the interface that will receive all user requests. The procedures to be performed should be moderately facilitated, but important filters will not be left aside because they are complex; however their use should be intuitive and whenever possible, with mechanisms to help the user. It is intended to be an adaptive system, but at the moment, a specific interface for devices with small resolution is not part of the scope. This choice is justified by the density of the data to be displayed, thus requiring a higher resolution for an intelligible visualization.

The system's users need to be registered so that in this way, there can be adequate customization, as well as enable relevant analyses for decision-making related to resource allocation and future implementations. This authentication is also intended to prevent user abuse in order to preserve the availability and integrity of the developed system.

Finally, the local database will be relational, and it will store all the data that is relevant to the queries. It is intended that some processed data is stored in a way that allows the user to retrieve the results processed by him, he should also store presets of filters defined by the user so that he can reapply the same filters in future situations.

5 Preliminary Results

During the studies for the development of this project, it was necessary to carry out the implementation of some functionalities that will be part of the system. In order to perform the initial analyses of the information available on the GEO platform, a preliminary study was carried out in order to map the format and patterns of the data contained in the files made available by the GEO database. From this starting point, it was possible to create an intelligent search algorithm capable of reading and interpreting datasets made available by GEO and store the relevant data collected. A relational database was designed for storing information.

In order to visualize the results, a WEB interface was created, in which researchers can define which data within the datasets they wish to use in their

searches. As an output, the system will show relevant data in a user-friendly interface. Within each dataset, it is possible to visualize information about how the original gene expression study was conducted, how the study divided the subjects into subgroups, as well as the gene expression readings of the subjects studied.

In Fig. 2, it is possible to visualize some data obtained from a study on diabetes. The first table, denominated dataset, contains some basic information about the research such as the search code on the GEO platform, the title, a description, and the organism in which that research was conducted. It is also possible that, through the "See Subsets" link, the groups of patients who were part of the research are viewed. In the second table it is possible to identify some data obtained in the original microarray study. The columns initiated with "GSM" show the gene expression of individuals participating in the research for different genes, these are identified in the "Identifier" column.

Since the presented data are divided into research groups, it is possible to perform the comparison between the groups so that the fold change calculations are performed, allowing to evaluate which gene set has a significant change when comparing the control group and a group of patients with a certain disease. Also in this study, it was possible to obtain a series of other numerical or categorical data related to the research; these data can be mined in search of relevant information. It was also perceived that the texts described by the authors of the researches do

DATASET

DATA	VALUE
CODE	GDS3875
TITLE	Diabetic children: peripheral blood mononuclear cells (U133B)
DESCRIPTION	Analysis of PBMCs from children (2-18 years) with newly diagnosed type 1 or 2 diabetes (T1D or T2D). One and 4 month samples were obtained from 20 of the T1D patients. Results provide insight into molecular mechanisms that distinguish T1D and T2D and those that are common to both forms of diabetes.
ORGANISM	Homo sapiens

SEE SUBSETS

1 2 3 4 5 Next » Last »

ID_REF	IDENTIFIER	GSM254177	GSM254179	GSM254180	GSM254182	GSM254183	GSM254277	GSM254278	GSM254281	GSM254282
200000_s_at	PRPF8	8016	8032.8	7955	3256.4	4021	6612.3	6286.4	6967.1	6676.4
200001_at	CAPNS1	12936.5	12661.6	19142.4	4731.1	6213.5	16610.2	12436.3	12676.4	14712.3
200002_at	RPL35	66093.2	74160.5	63443.7	35064.3	42399	54791.6	44498.2	58658.8	61862.2
200003_s_at	MIR6805	153186	150770	117061	104832	120795	151946	104801	139336	129766
200004_at	EIF4G2	25483.4	21681.9	26903.1	23736.1	25317.3	28751.5	26696.7	26480.8	24446
200005_at	EIF3D	20952.3	22003.2	16172.9	16634.2	16788.3	18192.8	19572.6	21380.3	16769.3
200006_at	PARK7	23918.8	22589.2	25744.1	24559.8	27586.2	21332.9	24587.6	21702	22594.2
200007_at	SRP14	24583.6	26654.8	30690	23587.3	24459.4	19803.1	25721.4	32627.7	24050.2
200008_s_at	GDI2	12830.2	13772.1	12613.5	12039.6	12462.8	9891.1	10958.7	13669.9	13238.1

Fig. 2 Partial visualization interface for obtained data. Source: The authors

not have a well-defined structure and, with this it was perceived that an automated analysis of this information could only be made from the use of text mining.

The data obtained from the metadata of the captured GEO datasets demonstrated that the objectives foreseen for this study are achievable and also shows that other functionalities can be implemented in future versions of the project, thus ensuring that a more complete system can be made available to the target audience.

6 Conclusion

This paper describes a system under development capable of assisting biological and life sciences researchers in conducting new studies from public data from various scientific researches. Initiatives like this enable computing to support new discoveries without the need to conduct new laboratory tests. The results already obtained prove that with due treatment the data can be easily interpreted and thus easing the difficulty of comprehension of these data on the part of researchers.

In future developments, we intend to adapt the system in order to include research data that use the RNA-Seq technique, so that more studies can be analyzed. It is also part of future implementations to enable the system to obtain data from other biological databases.

Acknowledgments The authors would like to thank *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) and *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) for funding researchers' scholarships.

References

1. K.-C. Wong, Big data challenges in genome informatics. Biophys. Rev. **11**(1), 51–54 2019. https://doi.org/10.1007/s12551-018-0493-5
2. M. Gasparovica-Asīte, L. Aleksejeva, Classification methodology for bioinformatics data analysis. Autom. Control. Comput. Sci. **53**(1), 28–38 (2019)
3. X.M. Fernández-Suárez, E. Birney, Advanced genomic data mining. PLoS Comput. Biol. **4**(9), e1000121 (2008), 18818719[pmid]. https://www.ncbi.nlm.nih.gov/pubmed/18818719
4. A. Brazma, Minimum information about a microarray experiment (miame)—successes, failures, challenges. TheScientificWorldJournal **9**, 420–3 (2009)
5. Z. Wang, A. Lachmann, A. Ma'ayan, Mining data and metadata from the gene expression omnibus. Biophys. Rev. **11**(1), 103–110 (2019). https://doi.org/10.1007/s12551-018-0490-8
6. H. Nie, P. Neerincx, J. van der Poel, F. Ferrari, S. Bicciato, J. Leunissen, M. Groenen, Microarray data mining using bioconductor packages. BMC Proc. **3**(Suppl. 4), S9 (2009)
7. V. Gangwar, U. Ghose, Y. Singh, Data mining of biological data in bioinformatics using transcription, translation algorithm and pattern matching of protein sequences. Int. J. Adv. Res. Comput. Sci. **3**(3) (2012)
8. F.S. Espindola, L.K. Calábria, A. Azenha Alves de Rezende, B. Barbosa Pereira, F. Assumpção Santana, I. Marques Rodrigues Amaral, J. Lobato, J. Luzia França, J. Luiz Mario, L. Bruno Figueiredo, L. Pereira dos Santos, N. Moura de Gouveia, R. Nasci-

mento, R. Roland Teixeira, T. Alves dos Reis, T. Gonçalves de Araújo, Recursos de bioinformática aplicados às ciências ômicas como genômica, transcriptômica, proteômica, interatômica e metabolômica. Biosci. J. **26**(3) (2010). http://www.seer.ufu.br/index.php/biosciencejournal/article/view/7146
9. S.B. Garg, A.K. Mahajan, T. Kamal, An approach for diabetes detection using data mining classification techniques. J. Eng. Sci. **26** (2017)
10. K. Lan, D.-T. Wang, S. Fong, L.-S. Liu, K.K. Wong, N. Dey, A survey of data mining and deep learning in bioinformatics. J. Med. Syst. **42**(8), 139 (2018)
11. H. Bono, All of gene expression (AOE): an integrated index for public gene expression databases. PLoS One **15**(1), e0227076 (2020)