GenotypeAnalytics: a RESTFul platform to mine multiple associations between SNPs and drug response in case-control studies

Giuseppe Agapito[1], **Pietro Hiram Guzzi**[1], **and Mario Cannataro**[1]

[1]**Data Analytics Research Center, Department of Medical and Surgical Sciences, University "Magna Græcia" of Catanzaro, Italy**

Corresponding author:
Mario Cannataro

Email address: cannataro@unicz.it

ABSTRACT

We present GenotypeAnalytics (GA), a RESTFul service that makes it possible to mine association rules from Single Nucleotide Polymorphism (SNP) datasets using standard web browsers. GA can speed up and simplify the analysis of this massive amount of data, highlighting only the SNPs involved, for example, in the development of the disease or responsible for adverse drug reactions to the drug. In this way, the doctor may use this extracted knowledge for a significant improvement in the quality of the treatments.

INTRODUCTION

To understand the complex biology machinery underlying complex diseases, i.e. Cancer, Diabetes, and Alzheimer, it is crucial to improve the capacity to diagnose and treat diseases. High-throughput methodologies (i.e. microarray and mass-spectrometry) are able to produce a huge amount of data for the single experiment, making it possible to study the disease from a broad perspective, Single Nucleotide Polymorphism (SNP) microarray, are considered as promising methods to identify the cause of many complex diseases or to improve the understanding of the basis of variable response to drugs (pharmacogenetics) [1]. The development of novel algorithms able to deal with the huge amount of data generate for a single experiment avoids that researchers are overwhelmed by data as well as, that researchers have to manually analyse these data, that is time-consuming and error prone. Efficient and scalable data analysis tool can simplify the tasks of researchers, speeding-up the analysis and interpretation of complex data, improving the quality of the results as well as the accuracy, by identifying genetic variants that alter susceptibility to complex disease [2], making possible to provide to the single patients the best therapeutic treatments. For these reasons, plenty of software tools able to efficiently deal with SNPs data [3, 4, 5, 6, 7, 8, 9, 10] are available. High-throughput DNA microarrays together with efficient software tools have the potential to provide critical clues for the management of complex diseases such as cancer [11]. Thanks to the continue advanced in the quality, standardization, and ease of use, DNA microarrays are moving toward being a technology useful in clinical investigation and not used exclusively for research activities [12, 13]. On the other hand, to make this moving reality it is needed to provide the software tools able to deal with this huge amount of data as well as easy to use by researchers. To satisfy this necessity, we present GenotypeAnalytics (GA) a RESTFul service, that makes it possible to mine association rules from SNP-dataset by a common web-Browser. Besides, GA can speed up and simplify the analysis of this massive amount of data, highlighting only the SNP markers involved, for example, in the development of the disease or in adverse reaction to the drug. In this way, the doctor can use that knowledge (not known a priori) will allow a significant improvement in the quality of the treatment as the patient will receive the correct dose and correct drug without that the doctor proceeds for attempts.

METHODS

To meet the necessity to simplify the analysis of huge amounts of SNP data, we have developed a novel automatic methodology able to mine association rules from SNP data. To pursuing the easy of use, our analysis methodology is based on the automatic extraction of Association Rules (AR) from SNP dataset. AR mining is a powerful tool to highlight multiple associations at time hidden into the data as well as, are easy to understand, representing the ideal instrument to investigate complex diseases i.e. cancer or diabetes. It is well established that cancer and diabetes are multifactorial diseases and association rules make it possible to highlight variation in more than a single gene related among them at the time. The methodology proposed in this paper consists of the following two main steps: *i)* pre-processing user data; and *ii)* extracting association rules using the previously pre-processed data. This methodology has been implemented in the *GenotypeAnalytics* (GA) software tool. Preprocessing task is performed automatically by the GA and does not require user supervision. GA tries to automatically reduce the search space through a filtering methodology reducing the number of transactions. The filtering methodology is based on the use of the known *Fisher's Test*, that automatically removes all rows (probes) from SNP-dataset for which the null hypothesis can not be accepted (p-value < 0.5). After the filtering step, GA automatically converts the preprocessed data into a format suitable for extracting associative rules. At the core of the rule extraction algorithm, there is the FP-Growth algorithm. We extended the traditional version of FP-Growth to be able to handle SNP-datasets, by defining Frequent Pattern Tree (FP-Tree) data structure, able to store and manage SNP-dataset. FP-Tree is a tree of prefixes that maps together transactions containing the same prefixes. The mapping consists of two main operations: *occurrences-update* and *node-creation*. If during the mapping phase, the current element in the FP-Tree corresponds to the current element in the transaction, the occurrences-update function is invoked, which updates the current node occurrence (increments the frequency). Whereas the current node in the FP-Tree does not match the current node in the transaction, the node creation function is invoked, that adds to the current node a new node, appending it to the FP-Tree as child of the FP-Tree current node. The remaining elements in the current transaction are added as children of the last created FP-Tree node. A generic SNPs dataset is arranged as a huge table $(n \times m)$, wherein the columns there are the subject's labels divided into cases and controls i.e. from column $(1 \dots j)$ are labeled as controls whereas from column $(j+1 \dots m)$ are labeled as cases or vice versa. Furthermore, dataset can be annotated by using clinical data, i.e. the kind of response to drugs, e.g. toxicity or not-toxicity. The rows instead contain the probe's identifiers (i.e. $P_1, P_2, \dots, P_n$). Each cell (i, j) contains the SNP detected in the ith probe and related with the jth subject, such as *G/A*, in which the first letter represents the nucleotide of the first allele, and the second letter represent the nucleotide of the second allele. In such a way a genomic variant in two samples may be expressed like a form: *A/G* for subject S_1, and *G/G* for subject S_2. Table 1 depicts a general cases-controls table obtained by using a SNP microarray. Thus, from the dataset it is possible to extract association rules. In particular, GA provide association rules in which, the consequent part indicate the class where the rule belong g.e. case-control or respond-notrespond, making rules more useful for the researchers.

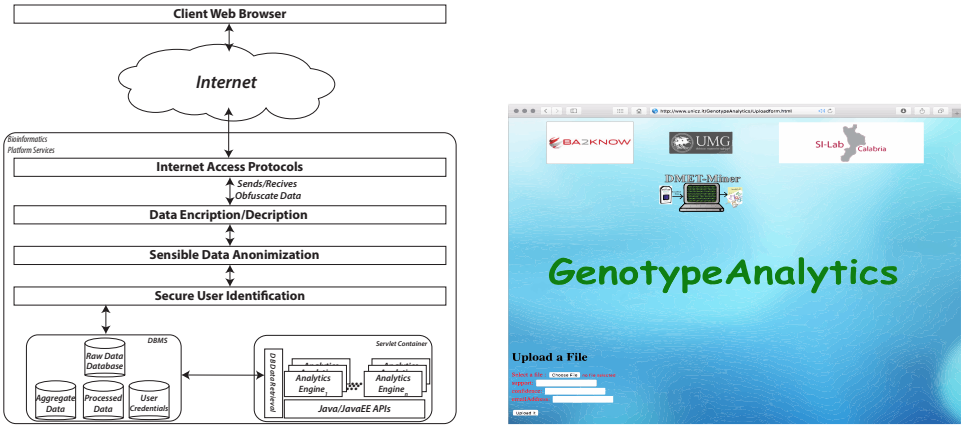
	S_1	S_2	S_3	…	S_m
P_1	A/G	G/G	A/A	…	A/G
P_2	T/T	A/T	T/T	…	A/A
⋮	⋮	⋮	⋮	…	⋮
P_n	A/G	G/G	A/A	…	A/G

Table 1. A simple SNPs dataset obtained from a cases-controls studies. Data are arranged in a tabular format, where S identify the subjects, whereas P indicates the probes. The generic cell (i, j) conveys the detected SNP in the ith probe related to the jth subject.

RESULTS

We have designed and built GA as a web-based platform for the client-server model for the analysis and storage of data produced by using SNP-microarray by implementing the analysis methodology described in the previous section. Using the client-server protocol, it allows the user to access the services provided

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3299v1 | CC BY 4.0 Open Access | rec: 29 Sep 2017, publ: 29 Sep 2017



(a) GenotypeAnalytics Architecture

(b) GenotypeAnalytics SNP-Data Upload and Analysis Starting page

Figure 1. GenotypeAnalytics.

by the platform even remotely through the use of a web browser. Specifically, the client layer provides the user with a simplified platform access interface, while the server layer, via the web, allows the client to communicate and interact with the logic of the application that does not reside locally. In this way, data analysis is transparent for the user who does not have to worry about maintaining and managing high-performance hardware to process these huge data queues. Lastly, the client-server model is efficient and scalable because it allows multiple user data requests to be met at the same time. GA has been developed by using Java v8.0, REST, and HTML5, and it is compatible with all the known operating systems. Figure 1(a) shows the web platform's top architecture that allows remote access to the SNP microarray data analysis service. Whereas Figure 1 (b) shows the GA analysis web page. Below is a description of the main functions of the platform modules as web services.

- *Internet Access Protocols*: provides the tools needed to transfer information from client to server and vice versa. We implemented our software by using REST technology, since REST exploits the potential of the HTTP protocol, such as security and addressing mechanisms.
- *Data Encryption/Decription*: allows researchers to safely transfer sensitive information, avoiding compromising the privacy of patients. To ensure that communications between client servers run safely, communications occur on HTTPS protocol, a variant of the HTTP protocol to which the Secure Socket Layer (SSL, Security Layer) is added.
- *Sensible Data Anonymization*: allows researchers to store data from their products to help shape a digital catalog (for example, the SNP mutation catalog). Anonymization is a process that allows us to modify data so that the subject's identity can not be identified.
- *DBMS*: the DBMS (Database Management System) manages the database and is perfectly integrated with the web server client architecture.
- *Secure User Identification*: allows only registered users to access to their area and manage their data (for example, make them public within the platform).
- *DBDataRetrieval*: allows to submit client and server requests to the DBMS, ensuring scalability and efficiency in data management.
- *AnalyticsEngines*: provides a web interface to the algorithms offered by the platform, enabling the user to locate and use the algorithms using the HTTP protocol.

CONCLUSIONS

GA is a platform for services whose goal is to provide support in experimental and clinical medicine, due to the huge amount of raw data produced daily. Data analysis software tools for Microarray, Mass

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3299v1 | CC BY 4.0 Open Access | rec: 29 Sep 2017, publ: 29 Sep 2017

Spectrometry, and other high-throughput analysis technologies will allow researchers to discover new aspects of hidden biological phenomena in data and can not be accessed explicitly at very fast times with respect to interpretation and Manual analysis (which is impractical). GA could support researcher and medical doctors to identify new molecular markers that can be used clinically. Possible new molecular markers can be used to understand which SNPs are responsible for a particular pathology, or what are the SNPs that negatively or positively influence drug responses in subjects with the same pathology and subjected to the same therapy. GA will make data analysis simple and effective through the implementation of methods ranging from statistics to data mining analysis of data generated by microarrays. Thanks to an automated system tuning process, the users will only have to send the data they intends to analyze by specifying the level of significance and await the results, allowing the researchers to identify multiple genetic variants that affect complex pathologies in humans.

ACKNOWLEDGMENTS

This work has been partially funded by the "BA2Know-Business Analytics to Know" (PON03PE_00001_1) project funded by the MIUR.

REFERENCES

1. Arbitrio, Mariamena, Maria Teresa Di Martino, Francesca Scionti, Giuseppe Agapito, Pietro Hiram Guzzi, Mario Cannataro, Pierfrancesco Tassone, and Pierosandro Tagliaferri. "DMET^{TM}(Drug Metabolism Enzymes and Transporters): a pharmacogenomic platform for precision medicine." Oncotarget 7, no. 33 (2016): 54028

2. Risch, N., and Kathleen M. "The future of genetic studies of complex human diseases." Science 273.5281 (1996): 1516-1517

3. Guzzi, P. H., et al. "DMET-Analyzer: automatic analysis of Affymetrix DMET Data." BMC bioinformatics 13.1 (2012): 258

4. Guzzi, Pietro Hiram, Giuseppe Agapito, & Mario Cannataro. "coresnp: Parallel processing of microarray data." IEEE Transactions on Computers 63.12 (2014): 2961-2974

5. Agapito, G., Guzzi, P. H., & Cannataro, M. (2015). DMET-Miner: Efficient discovery of association rules from pharmacogenomic data. Journal of biomedical informatics, 56, 273-283.

6. Agapito, G., Cannataro, M., Guzzi, P. H., Marozzo, F., Talia, D., & Trunfio, P. (2013, September). Cloud4snp: distributed analysis of snp microarray data on the cloud. In Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (p. 468). ACM.

7. Agapito, G., Botta, C., Guzzi, P.H., Arbitrio, M., Di Martino, M.T., Tassone, P., Tagliaferri, P. and Cannataro, M., 2016. OSAnalyzer: A Bioinformatics Tool for the Analysis of Gene Polymorphisms Enriched with Clinical Outcomes. Microarrays, 5(4), p.24.

8. Cannataro, M., Talia, D., Tradigo, G., Trunfio, P., & Veltri, P. (2008). SIGMCC: A system for sharing meta patient records in a Peer-to-Peer environment. Future Generation Computer Systems, 24(3), 222-234.

9. Gamazon, Eric R., Heather E. Wheeler, Kaanan Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler et al. "PrediXcan: Trait Mapping Using Human Transcriptome Regulation." bioRxiv (2015): 020164.

10. Yang, Tsun-Po, Claude Beazley, Stephen B. Montgomery, Antigone S. Dimas, Maria Gutierrez-Arcelus, Barbara E. Stranger, Panos Deloukas, and Emmanouil T. Dermitzakis. "Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies." Bioinformatics 26, no. 19 (2010): 2474-2476.

11. Agapito, Giuseppe, Pietro Hiram Guzzi, and Mario Cannataro. "Visualization of protein interaction networks: problems and solutions." BMC bioinformatics, 14.1 (2013): S1.

12. Di Martino, M. T., Arbitrio, M., Guzzi, P. H., Leone, E., Baudi, F., Piro, E., & Veltri, P. (2011). A peroxisome proliferator-activated receptor gamma(PPARG)polymorphism is associated with zoledronic acid-related osteonecrosis of the jaw in multiple myeloma patients: analysis by DMET microarray profiling. British journal of haematology, 154(4), 529-533.

13. Guzzi, P. H., Agapito, G., Milano, M., & Cannataro, M. (2016). Methodologies and experimental platforms for generating and analysing microarray and mass spectrometry-based omics data to support P4 medicine. Briefings in bioinformatics, 17(4), 553-561.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3299v1 | CC BY 4.0 Open Access | rec: 29 Sep 2017, publ: 29 Sep 2017