Learning Association Rules for Pharmacogenomic Studies

Giuseppe Agapito, Pietro H. Guzzi, and Mario Cannataro

1Dept of Medical and SurgicalSciences, University of Catanzaro,Italy
2 Data Analytics Research Centre, University of Catanzaro
{agapito,hguzzi,cannataro}@unicz.it,

Abstract. The better understanding of variants of the genomes may improve the knowledge of the causes of different response to drugs of individuals. The Affymetrix DMET (Drug Metabolizing Enzymes and Transporters) microarray platform offers the possibility to determine the gene variants of a patient and correlate them with drug-dependent adverse events. The analysis of DMET data is a growing research area. Existing approaches span from the use of simple statistical tests to more complex strategies based, for instance, on learning association rules. To support the analysis, we developed GenotypeAnalytics, a RESTFul-based software service able to automatically extract association rules from DMET datasets. GenotypeAnalytics is based on an optimised algorithm for learning rules that can outperform general purpose platforms.

Keywords: Association Rules, Genomics, SNP

1 Introduction

One of the problems related to drug-development and clinical practice is the variability of the response to the same drug. Pharmacogenomic is a relatively new discipline based on the rationale that this variability is due to different variants in the genome of patients. [18]. In particular, it has been shown that a set of genes, defined to as drug absorption, distribution, metabolism and excretion genes (ADME-genes) [16] are related to such processes [17, 7]. Such genes present known Single Nucleotide Polymorphisms (SNPs), e.g. variants on the sequence of nucleotides, related to different drug responses [13, 5].

To study genome variants, we need: (i) an experimental platform for investigating the presence of SNPs in the ADME genes (among others we consider the Affymetrix DMET platform) [8, 14], (ii) a computational platform to associate single or multiple SNPs to drug response. Although such analysis is usually performed through statistical analysis, in the following, we will consider analysis approaches based on data mining, i.e. association rule mining. From a computer science point of view, the result of a DMET experiment is a $n \times m$ matrix of alleles, where n is the number of probes ($n = 1936$ in the current DMET plate) and m is the number of samples (patients). Each cell of such table contains a string value including two alleles symbols i.e. a_1/a_2, where

$a_1, a_2 \in A = \{A, C, G, T, -\}$, see for instance Table 1 that reports a fragment of DMET data.

Table 1. A simple DMET SNP microarray data set. S and P respectively refer to sample and probe identifiers.

Probes \ Samples	S_1	...	S_N
P_1	G/A	...	T/T
P_M	G/A	...	T/C

Usually, the algorithms for the analysis of DMET data try to correlate the presence of genomic variants to the phenotype of patients. Early approaches to the analysis were based mainly on statistical approaches, i.e. DMET-Analyzer [12] employed the well-known Fisher test and several statistical corrections such as Bonferroni or False Discovery Rate. Although DMET-Analyzer has demonstrated its validity in several clinical studies [17, 19, 7, 8], DMET-Analyzer is not able to cope with multiple variants. To overcome those limitations, we developed DMET-Miner, a novel methodology for the simultaneous analysis of genomic variants in more than a gene. DMET-Miner employs the association rules mining methodology [4], a well-known method in the data mining field. Despite the innovation introduced by DMET-Miner, it presents some disadvantages due to the Apriori method i.e. the generation of the candidate itemsets could be extremely slow and require a massive amount of main memory [15, 9, 10].

To avoid memory issues and to improve the computation of association rules, we here extended the core of DMET-Miner by implementing a modified FP-Growth algorithm able to deal with SNP data efficiently and we implemented it into a new software named GenotypeAnalytics. The main of FP-Growth concerning Apriori is that FP-Growth does not need to generate candidate set and it needs to read the input data-set only twice, as opposed to Apriori that reads the input data-set on each iteration. GenotypeAnalytics improves the performances of DMET-Miner by using optimised data structures that give good performance results in rule extraction also with massive DMET datasets. Also, GenotypeAnalytics can extract relevant knowledge by computing frequent item-sets efficiently as well as mining association rules [2] that link allelic variants in more than one probe with the health status of patients (e.g. subjects responding or not responding to drugs). Paper is structured as follows: Section 2 discusses the related approaches, Section 3 introduces the problem, Section 4 discusses the proposed algorithm and its implementation, Section 5 presents some experimental results, Section 6 concludes the paper.

2 Related Work

Existing approaches of analysis of DMET data, span from preprocessing of raw data, e.g. Affymetrix-power-tools, Affymetrix-DMET-Console, to the correlation of variants of different patient conditions, e.g. DMET-Analyzer [12], Cloud4snp [1], coreSNP [15], and DMET-Miner [3].

The `apt-dmet-genotype` software of the Affymetrix Power Tools suite, or the DMET Console platform [20], generally allows only the sequential preprocessing of binary data and simple data analysis operations. DMET-Analyzer [12] is a software platform for the automatic statistical analysis of DMET data that employs the well-known Fisher test and several statistical corrections such as Bonferroni or False Discovery Rate. Although DMET-Analyzer has demonstrated its validity in several clinical studies [17, 19, 7, 8], DMET-Analyzer is not able to cope with multiple variants, and it is not able to group all of them in a single, easy to understand, and biologically relevant information. *Cloud4SNP* is the Cloud-based version of *DMET-Analyzer*. *Cloud4SNP* allows to statistically test the significance of the presence of SNPs in two classes of samples using the well known Fisher test. To cope with high dimensional dataset deriving from the screening of population, we developed coreSNP, a parallel version of DMET-Analyzer.

To overcome limitations of the analysis, we developed DMET-Miner, a novel methodology for the simultaneous analysis of genomic variants in more than one gene. DMET-Miner uses on the association rules mining methodology [**?**], a well-known method in the data mining field. Despite the innovation introduced by DMET-Miner, it presents some disadvantages due to the Apriori method.

3 Problem Statement

DMET datasets are represented as a $m \times n$ SNP DMET table. In particular, m is the number of probes (in the current version of the DMET chip is equals to 1936), whereas n is the number of subjects (patients) gathered for the class of membership i.e. Healthy and Diseased or responding and not responding to the drug . Each element (i, j) of the table contains the allele recognised on the ith probe and at the jth sample. An example of synthetic SNP DMET dataset randomly generate is reported in Table 2.

To extract relevant rules, we need to convert the input DMET data set into a transaction database. The conversion of DMET data set includes the following steps:

- loading and transposing of the input DMET dataset and (let see e.g. Table 2) obtaining a $n \times m$ table of alleles named AllelesTable AT (see Table 3). In this way, each row of the AT contains a transaction and related items. Table 3 shows the transformed matrix AT for the input dataset of Table 2.
- choose of desired support and confidence
- ATTable 3 is then used to extract frequent itemsets.

Table 2. A simple DMET SNP microarray data set. S and P respectively refer to sample and probe identifiers.

Probes \ Samples	S_1	S_2	S_3	$\cdots$	S_N
P_1	G/A	A/G	A/G	$\cdots$	T/T
P_2	G/A	A/G	A/G	$\cdots$	T/C
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
P_M	G/A	A/G	A/G	$\cdots$	T/C

Table 3. The AllelesTable AT obtained transposing the input DMET microarray dataset. S and P respectively refer to sample and probe identifiers.

Samples \ Probes	P_1	P_2	$\cdots$	P_M
S_1	G/A	G/A	$\cdots$	G/A
S_2	A/G	A/G	$\cdots$	A/G
S_3	A/G	A/G	$\cdots$	A/G
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_N	T/T	T/C	$\cdots$	T/C

– Biological interpretation of extracted rules

To explain the overall process, we here recall main concepts.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items (alleles), where an item is identified by a specific SNP into a cell (i, j) of AT. Let T the set of transactions, formally a transaction over I is a couple $T = (tid, I)$, where tid is the transaction identifier, and I is an item or item set. The number of items present in a transaction is defined as *transaction width*. A transaction T_j contains an itemset J, if J is a subset of T_j, this is $J \subset T$. Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions, called *DMET-Dataset* D hereafter. Each transaction in D is identified by an unique *ID* of the corresponding sample or patient.

Now we may start the mining phase by performing the following steps:

1. prune all the items that present a support value lower than the specified minimum frequency threshold.
2. add all the frequent items to the FP-Tree.
3. mine association rules from the FP-Tree.

The power of frequent item sets extraction concerns with the ability to discover interesting relationships hidden in large data sets. This feature relies on a fundamental property of the itemset also known as *Support*. Support refers to the number of transactions that contain a particular item or item

set. Formally, the Support $S(\cdot)$ of an item X, $S(X)$ can be defined as follows: $S(X) = |\{\forall t_i \subset X \wedge t_i \in T\}|$, where $|\cdot|$ denotes the cardinality of the set. In other words, $S(X)$ is the fraction of transactions in T containing the item/item-set X.

Association models extract rules that express the relationships among items into frequent itemsets. For example, a rule belonging to the frequent itemsets composed by the following elements $\{A/A, G/C, C/T\}$, and might be stated as: **IF** $(A/A \wedge G/C)$ **THEN** C/T and, can be read as: if A/A and G/C are included in the transaction, then C/T likely should also be included.

4 The optimised FP-Growth algorithm of GenotypeAnalytics

This section illustrates the core algorithm of GenotypeAnalytics and its optimisations used to reduce the space search and to minimise the number of mined association rules. The goal of GenotypeAnalytics given a SNPs dataset D is to discover all the frequent patterns above a user support threshold named $min_{support}$ (Minimum Support).

Before to convert the input dataset in a transaction database, GenotypeAnalytics tries to reduce the search space through a suitable preprocessing methodology able to decrease the number of possible transactions. The preprocessing method is based on the use of the well known *Fisher's Test* as a filter, which allows removing all the rows from the original DMET dataset for which it is not possible to accept the *null hypothesis*. The discharging of this rows does not lead to lost useful information, rather allow to improve the mining of the association rules (see for a better explanation [3]). After the filtering step, the resulting table is transformed into a transaction database. Such transformation is necessary since the extraction of frequent itemsets is more efficient with this data format. Alleles of different probes have some time the same name. Therefore, we modified the variables adding information related to the probe to which each allele belongs using the following notations (i.e. X.A/A and Y.A/A).

The resulting table is stored in a data-structure called *Transaction DB* (*TDB*). In the TDB the transaction id (*TID*) is the entry of the table and the matching items set (value) are encoded into hash-set using a hash-function. Thus, it is possible to compress the items and to ensure constant time for standard operations such as: inserting, deleting and searching items in the hash-set.

Despite the preprocessing phase of the input dataset, the number of items that compose the *TDB* is huge enough. Thus, a further compression step is necessary to manage the enumeration and generation of frequent itemsets better. For this reason, we decided to implement a customized version of the FP-Growth algorithm, able to deal with SNPs data. The FP-Tree, allows storing in a compressed way the TDB into the main memory named *FP-Tree*. The *FP-Tree* keeps track of the same item contained in different transactions by connecting the prefix tree nodes indicating the same item into a *frequent items list* The mining of the associative rules is done using a *Depth-First-Search*, (DFS, in short), sorting in descending order the items in each transaction. The reason behind this choice

is that the average size of the conditional *TDB* tends to be smaller if the items are processed in this order. Moreover, the order of the items influences only the search time, not the result of the algorithm.

The GenotypeAnalytics core algorithm needs to scan twice the *TDB*. The first pass is necessary to discover the frequency of each item I into the *TDB* for which $S(I) \geq min_{sup}$ and sorting the items according to their descending frequency. The second *TDB* scan is necessary to delete the items for which their support is $S(I) < min_{sup}$. Sorting the items in descending order of frequency allows to further compress the FP-Tree, by limiting the number of different possible prefixes. Now all the items into the *TDB* can be mapped on the FP-Tree. The mapping is performed by means the *support-update* and *node-creation* functions. If during the mapping, the current element in the transaction matches the current element in the FP-Tree, the function *support-update*, which updates the support of the current node, is invoked. Whereas, if the current node in the FP-Tree and the current node in the transaction do not match the function *node-creation* is called. The *node-creation* function starting from the current item creates a new node, adding it as children of the current FP-Tree node. The other items in the current transaction are appended as children of the last created FP-Tree node.

5 Performance Evaluation

In this section, we present the performance evaluation of our version of the FP-Growth algorithm with respect to the FP-Growth algorithm available in *SPMF* an Open-Source Data Mining Library [11], and the version proposed in [6]. Experiments have been ran on the same data sets, namely *"Vote.arff"*, *"Supermarket.arff"* and a synthetic *"DMET-SNP"* data set. As proof-of-principle, we report the performance evaluation results of all the FP-Growth implementations. All the experiments have been executed on a machine equipped with a Pentium i7 2.5 GHz CPU, 16 GB RAM and a 512 GB SSD disk. The reported execution times refer to average times; each value has been computed repeating 10 times the experiments with the same settings. In this way, it is possible to ensure that the results are comparable.

Figures 1,2, and 3 convey, the execution times obtained analyzing the data sets by the three different implementation of the FP-Growth algorithms. All the execution time are obtained by varying the minimum support values. The solid black line refers to our implementation of the FP-Growth algorithm, the dashed green line refers to the FP-Growth version available in *SPMF*, and finally the dash-dot red line refers to the FP-Growth version proposed in [6].

Among these implementations of FP-Growth, all show good performance on the classical Vote and Supermarket data sets. Our implementation of FP-Growth does not present so much differences with the other two methods, showing very similar performance to those of other tools. With the exception of the synthetic SNP data set, on which they are bet with an appreciable margin by our imple-

mentation of FP-Growth, because our version of FP-Growth is highly optimized to dig with SNPs data, thus clearly performs best.

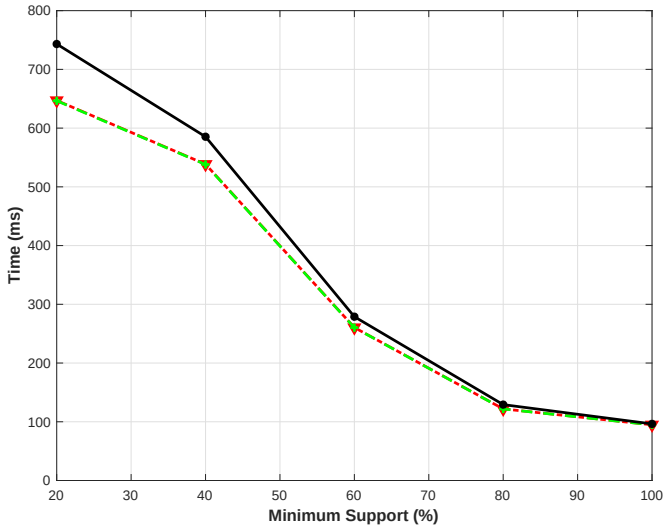

Fig. 1. Execution times of the different FP-Growth algorithm on the Vote data set. The execution times are obtained varying the value of minimum support.

6 Conclusion

Analysing genotyping datasets presents various challenges due to the huge volumes of data and due to the specific characteristics of SNPs data. Thus, using general purpose data mining implementation is not feasible and for this reason we implemented GenotypeAnalytics, a specialised association rule mining system system to mine association rules from DMET genotype data. It includes an optimised version of the implementation of the FP-Growth algorithm. Preliminary experiments show how our solution outperformes off the shelf implementation of FP-Growth. As future work we will investigate automatic methods to rank the extracted rules on the basis of their biological significance and memory by using real DMET dataset in the oncology domain.

7 Acknowledgements

This work has been partially funded by the following research projects:

–

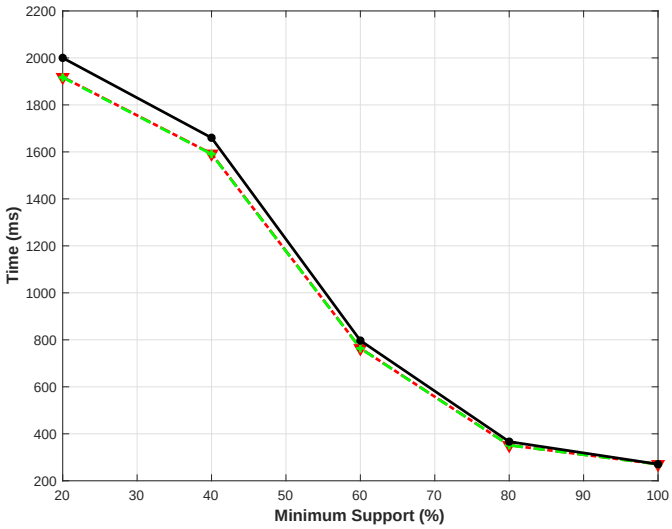

Fig. 2. Execution times of the different FP-Growth algorithm on the Supermarket data set. The execution times are obtained varying the value of minimum support.

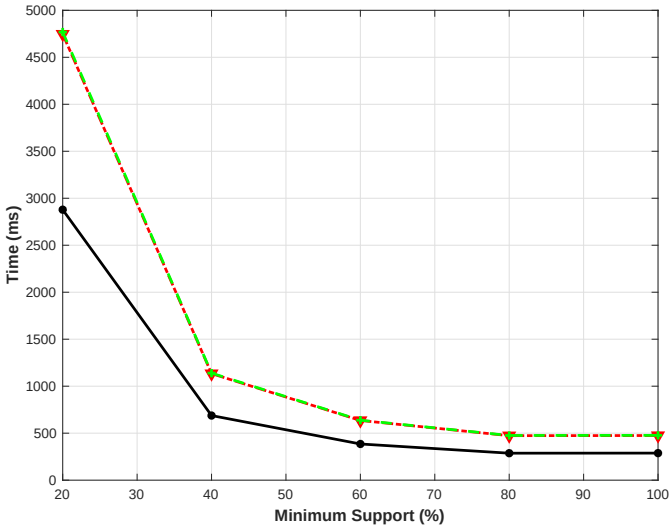

Fig. 3. Execution times of the different FP-Growth algorithm on the Vote data set. The execution times are obtained varying the value of minimum support.

– "BA2Know-Business Analytics to Know" (PON03PE_00001_1), funded by the Italian Ministry of Education and Research (MIUR)

– INdAM - GNCS Project 2017: "Efficient Algorithms and Techniques for the Organization, Management and Analysis of Biological Big Data".

References

1. Agapito, G., Cannataro, M., Guzzi, P.H., Marozzo, F., Talia, D., Trunfio, P.: 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 (2013)
2. Agapito, G., Cannataro, M., Guzzi, P.H., Milano, M.: Using go-war for mining cross-ontology weighted association rules. Computer methods and programs in biomedicine 120(2), 113–122 (2015)
3. Agapito, G., Guzzi, P.H., Cannataro, M.: Dmet-miner: Efficient discovery of association rules from pharmacogenomic data. Journal of biomedical informatics 56, 273–283 (2015)
4. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Acm sigmod record. vol. 22, pp. 207–216. ACM (1993)
5. Arbitrio, M., Di Martino, M.T., Barbieri, V., Agapito, G., Guzzi, P.H., Botta, C., Iuliano, E., Scionti, F., Altomare, E., Codispoti, S., et al.: Identification of polymorphic variants associated with erlotinib-related skin toxicity in advanced non-small cell lung cancer patients by dmet microarray analysis. Cancer chemotherapy and pharmacology 77(1), 205–209 (2016)
6. Borgelt, C.: Frequent item set mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(6), 437–456 (2012)
7. Di Martino, M.T., Arbitrio, M., Guzzi, P.H., Leone, E., Baudi, F., Piro, E., Prantera, T., Cucinotto, I., Calimeri, T., Rossi, M., Veltri, P., Cannataro, M., Tagliaferri, P., Tassone, P.: 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 pp. 529–533 (2011), http://dx.doi.org/10.1111/j.1365-2141.2011.08622.x
8. Di Martino, M.T., Arbitrio, M., Leone, E., Guzzi, P.H., Saveria Rotundo, M., Ciliberto, D., Tomaino, V., Fabiani, F., Talarico, D., Sperlongano, P., Doldo, P., Cannataro, M., Caraglia, M., Tassone, P., Tagliaferri, P.: Single nucleotide polymorphisms of ABCC5 and ABCG1 transporter genes correlate to irinotecan-associated gastrointestinal toxicity in colorectal cancer patients: A DMET microarray profiling study. Cancer Biology and Therapy 12(9), 780–787 (November 1 2011)
9. Di Martino, M.T., Guzzi, P.H., Caracciolo, D., Agnelli, L., Neri, A., Walker, B.A., Morgan, G.J., Cannataro, M., Tassone, P., Tagliaferri, P.: Integrated analysis of microrna, transcription factors and target genes expression discloses a specific molecular architecture of hyperdiploid multiple myeloma. Oncotarget 6(22), 19132 (2015)
10. Di Martino, M.T., Scionti, F., Sestito, S., Nicoletti, A., Arbitrio, M., Guzzi, P.H., Talarico, V., Altomare, F., Sanseviero, M.T., Agapito, G., et al.: Genetic variants associated with gastrointestinal symptoms in fabry disease. Oncotarget 7(52), 85895 (2016)

11. Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.W., Tseng, V.S.: Spmf: a java open-source pattern mining library. The Journal of Machine Learning Research 15(1), 3389–3393 (2014)
12. Guzzi, P., Agapito, G., Di Martino, M., Arbitrio, M., Tassone, P., Tagliaferri, P., Cannataro, M.: Dmet-analyzer: automatic analysis of affymetrix dmet data. BMC Bioinformatics 13(1), 258 (2012), http://www.biomedcentral.com/1471-2105/13/258
13. Guzzi, P.H., Agapito, G., Milano, M., Cannataro, M.: Methodologies and experimental platforms for generating and analysing microarray and mass spectrometry-based omics data to support p4 medicine. Briefings in bioinformatics p. bbv076 (2015)
14. Guzzi, P.H., Cannataro, M.: μ-cs: An extension of the tm4 platform to manage affymetrix binary data. BMC bioinformatics 11(1), 315 (2010)
15. Guzzi, P.H., Agapito, G., Cannataro, M.: coresnp: Parallel processing of microarray data. IEEE Transactions on Computers 63(12), 2961–2974 (2014)
16. Li, J., Zhang, L., Zhou, H., Stoneking, M., Tang, K.: Global Patterns of Genetic Diversity and Signals of Natural Selection for Human ADME Genes. Human Molecular Genetics 20(3), 528–540 (2011)
17. Lombardi, G., Rumiato, E., Bertorelle, R., Saggioro, D., Farina, P., Della Puppa, A., Zustovich, F., Berti, F., Sacchetto, V., Marcato, R., et al.: Clinical and genetic factors associated with severe hematological toxicity in glioblastoma patients during radiation plus temozolomide treatment: A prospective study. Am J Clin Oncol. doi 10, 1097 (2013)
18. Meyer, U.A.: Pharmacogenetics and adverse drug reactions. The Lancet 356(9242), 1667–1671 (2000)
19. Rumiato, E., Boldrin, E., Amadori, A., Saggioro, D.: Dmet (drug-metabolizing enzymes and transporters) microarray analysis of colorectal cancer patients with severe 5-fluorouraci-induced toxicity. Cancer Chemotherapy and Pharmacology 72(2), 483–488 (2013), http://dx.doi.org/10.1007/s00280-013-2210-1
20. Sissung, T., English, B., Venzon, D., Figg, W., Deeken, J.: Clinical pharmacology and pharmacogenetics in a genomics era: the dmet platform. Pharmacogenomics 11, 89–103 (2010)