RESEARCH

Association of variants in gene coding regions with clinical data in colombian patients using data mining

Jennifer Velez[1*†] and Elizabeth Leon[2]

Abstract

Background:
The need to understand the biological processes that are involved in different diseases, from the available biological data such as genomic sequences, microarrays, protein interactions, biomedical images among others and the rapid adoption of electronic medical records that provides an opportunity for large-scale research . Therefore, data mining techniques for the discovery of knowledge from obtaining information from different sources are increasingly important in biological and medical research.

Results: A group model was implemented for 228 patients, and they were associated with the variants for 4813 genes, obtaining 5 groups with their options available in the rules. As an analysis, an analysis of the CFTR the gene was also carried out by means of association rules and previously obtained groups. This is the search tag the measurements of the frequencies of the population were made in terms of the number of variants present the age, sex, type of variant and the allelic state of the variants. It was found that for the CFTR hay gene without pathogenic variants in the sampled population. A board also created to visualize the groups and the necessary rules for group and a database for variants in exons for Colombian patients.

Conclusions: Data mining techniques applied to disk support allow an inference of genetics structure the Colombian population and the epidemiological follow-up of the variants and their possible effects in patient's phenotypes.

Keywords: Variant; data mining; clustering; association rules

Content

This paper is organized in section Background, Results ,Discussion, Conclusions and Methods.

Background

Biological data mining (seen from bioinformatics) is the process of extracting new knowledge (previously unknown) from biological data, this also allows the use of concepts of data mining and automatic learning in theories and applications in research Biological, by deeding the data that are used to be applied, are genomes that come from DNA sequencing, transcriptome sequences that are RNA or proteins that come from inferences and experimental data from chemistry [1].

*Correspondence: jevelezse@unal.edu.co
[1]School of Engineering, Universidad Nacional de Colombia, Bogotá D.C., Colombia
Full list of author information is available at the end of the article
†Equal contributor

Inferences regarding large amounts of genomic data require analysis of computational tools to interpret data, being one of the most active areas where data mining is used (I understand data mining as the method of extracting information through learning automatic, statistics, artificial intelligence, recognition and visualization patterns) to solve biological problems, some examples where mining techniques have been applied is the classification of genes, the analysis of mutations in cancer and gene expression [2].

Clustering techniques of differentially expressed genes have also been applied, vector support machines have been used to associate the interactions between genes and generate biological networks, as well as traditional methodologies for data mining are not precise or efficient and they require new algorithms to be developed and methodologies that respond in a more precise way to a biological question [3]. Without forgetting that it is necessary to evaluate the available platforms,

the technological tools that allow the implementation of processes that associate data with research and obtain more generalized results. This should apply to the research requirements to ensure successful implementation [3].

Some of the data mining tasks are: 1. Classification: where the data is classified to a predefined class, 2. Association: see elements that are associated by rules, 3. Grouping or grouping: as the definition of a population of data within a subgroup or group [2].

The use of high performance sequencing techniques together with the application of data mining can contribute to the diagnosis of complex diseases such as cancer[4, 5].

Results

Exploratory analysis

The exploratory analysis of the information contained in the database was carried out. A sample of 250 patients donated by the Genetix SA laboratory was taken, of which only 228 had the informed consent to use the information for research purposes.

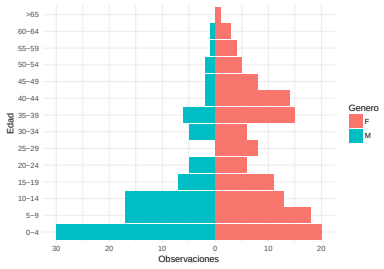

Figure 1 Distribution of age range and gender of patients

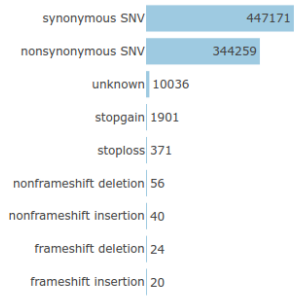

Figure 2 Distribution of age range and gender of patients

Sub-sub heading for section
Text for this sub-sub-heading ...

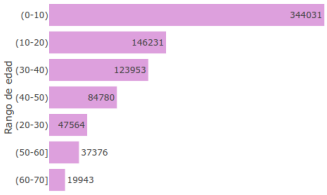

Figure 3 Distribution of age range and gender of patients

Sub-sub-sub heading for section Text for this sub-sub-sub-heading ...In this section we examine the growth rate of the mean of Z_0, Z_1 and Z_2. In addition, we examine a common modeling assumption and note the importance of considering the tails of the extinction time T_x in studies of escape dynamics. We will first consider the expected resistant population at vT_x for some $v > 0$, (and temporarily assume $\alpha = 0$)

$$E\big[Z_1(vT_x)\big] = E\Big[\mu T_x \int_0^{v\wedge 1} Z_0(uT_x)\exp\big(\lambda_1 T_x(v-u)\big)\,du\Big].$$

If we assume that sensitive cells follow a deterministic decay $Z_0(t) = xe^{\lambda_0 t}$ and approximate their extinction time as $T_x \approx -\frac{1}{\lambda_0}\log x$, then we can heuristically estimate the expected value as

$$\begin{aligned} E\big[Z_1(vT_x)\big] &= \frac{\mu}{r}\log x \int_0^{v\wedge 1} x^{1-u}x^{(\lambda_1/r)(v-u)}\,du \\ &= \frac{\mu}{r}x^{1-\lambda_1/\lambda_0 v}\log x \int_0^{v\wedge 1} x^{-u(1+\lambda_1/r)}\,du \\ &= \frac{\mu}{\lambda_1-\lambda_0}x^{1+\lambda_1/rv}\left(1-\exp\left[-(v\wedge 1)\left(1+\frac{\lambda_1}{r}\right)\log\right.\right. \end{aligned}$$

Thus we observe that this expected value is finite for all $v > 0$ (also see [**?**, **?**, **?**, **?**, **?**]).

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section ...

Acknowledgements
Text for this section ...

Author details
[1]School of Engineering, Universidad Nacional de Colombia, Bogotá D.C., Colombia. [2]School of Engineering, Universidad Nacional de Colombia, Bogotá D.C, Colombia.

References

1. Farid, D.M., Al-Mamun, M.A., Manderick, B., Nowe, A.: An adaptive rule-based classifier for mining big biological data. Expert Systems with Applications **64**, 305–316 (2016). doi:10.1016/j.eswa.2016.08.008

2. Littlefield, R.: An introduction into Data Mining in Bioinformatics. https://littlefield.co/an-introduction-into-data-mining-in-bioinformatics-964511e9ea21 Accessed 2017-11-19

3. Zaki, M.J., Karypis, G., Yang, J.: Data mining in bioinformatics (BIOKDD). Algorithms for Molecular Biology **2**(1), 4 (2007). doi:10.1186/1748-7188-2-4. arXiv:1011.1669v3
4. Hannah-Shmouni, F., Seidelmann, S.B., Sirrs, S., Mani, A., Jacoby, D.: The Genetic Challenges and Opportunities in Advanced Heart Failure. Canadian Journal of Cardiology **31**(11), 1338–1350 (2015). doi:10.1016/j.cjca.2015.07.735
5. Kawashima, K.: Text Mining and Pattern Clustering for Relation Extraction of Breast Cancer and Related Genes, 1–5 (2017)

Figures

Figure 4 Sample figure title. A short description of the figure content should go here.

Figure 5 Sample figure title. Figure legend text.

Tables

Table 1 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.