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.

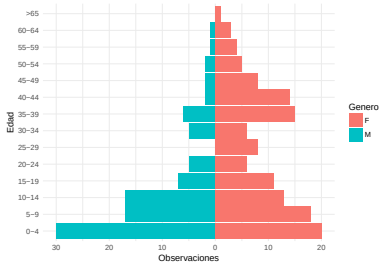

Figure 1 Distribution of age range and gender of patients

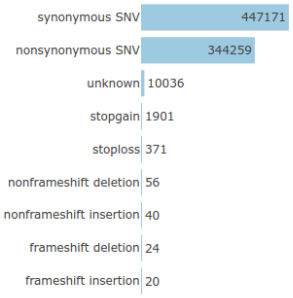

Figure 2 Distribution of age range and gender of patients

The database contains 228 patients of which 133 are female and have a total of 468,485 variants and 95 male patients with 345,239 variants obtaining a total of 803,878 variants. Figure 1 represents the distribution of patients by age range and Figure 2 represents the distribution of variants according to their type. In Figure 2 shows the number of variants that are synonymous and not synonymous, being the most frequent in the population, at a global level it is known that these are the most frequent types of variants[6].

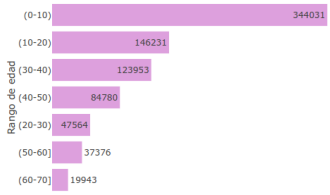

Figure 3 Distribution of age range and gender of patients

The unknown variants are the third type of variant more frequent given that there is still the problem of selecting the transcript to perform the appropriate nomenclature of the variants, so the annotator reports that they are unknown [7]. Figure 3 shows the distribution of the variants identified according to the age range, the range with the highest number of variants being patients between the ages of 0 to 10 years, given that it is the most represented population within the database.

The allelic state of the variants (zygosity) found within the database are divided into 458639 heterozygotes corresponding to 57.05% of the total of the variants and homozygous 345239 corresponding to 42.95%. The distribution of the zygosity of the variants can be explained from the error that can be generated in the identification of the variants given that during the call of variants it is possible that a homozygous variant is classified as heterozygous, if during the sequencing process it is identified wrongly nucleotides [8, 9].

Textual analysis of clinical information.

The results obtained for the frequency of words were breast, cancer, syndrome, suspicion and years. In figure 4 shows the first 30 most frequent words and the word cloud of all the documents.

The frequency of words shows us the main characteristics of the clinical information, the words cancer and breast are the main phenotypes, the word syndrome is also found that can be associated with different diseases and the word suspicion refers to ambiguous diagnoses that patients may have, One of the contributions of sequencing is that based on the phenotype can help a diagnosis, between different symptoms and syndromes that can be applied to rare and complex diseases [10].

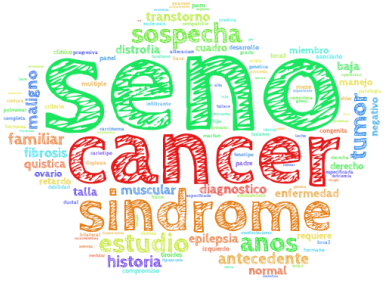

Figure 4 Wordcloud in Spanish

Clusters of clinical characteristics.
Data transformation Calculation of the tf-idf matrix, is done from inverted frequency with the equation:

$$idf_i = \log_2 \frac{|D|}{|\{d \mid t_i \in d\}|}$$

being $|D|$ what denotes the total number of documents and where $|\{d \mid t_i \in d\}|$ in which t_1 appears, the matrix of tf-idf is calculated from the multiplication of the frequency of terms and the inverted frequency $tf_{i,j} \cdot idf_i$ [11].

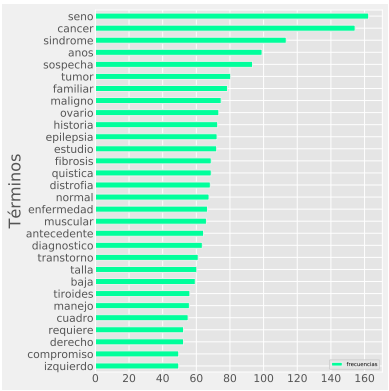

Figure 5 TF-IDF

Figure 5 represents the IDF-TF matrix of the words found within the database.

Clustering model validation
Inertia was computed and the graph of the quadratic error vs the number of clusters was made.Where cluster 5 and 6 can be selected as the optimum of K .For the definition of the optimal number of K, we also calculate the Silhouette coefficient, which is an evaluation of the clusters. Silhouette coefficient was *0.534*.

For 5 clusters obtained, the validation measures that are inside the library scikit-learn were: For homogeneity 0.296, for integrity 1.0, for V-measure 0.457 and the Rand-Index was 0. The perfect homogeneity would be with a value of 1.0, in the present clusters they present a low homogeneity, but an integrity of 1.0 which means that the labels are perfectly complete, this is reflected in the V-measure which is 0.457 where we have clusters with low homogeneity but high integrity. Rand-Index a value of 0.0 was obtained that shows that the classes are separated into different clusters [12].

Association of clusters with variants.
Results of applying association rules to the variants along with their clusters were:

Cluster 1

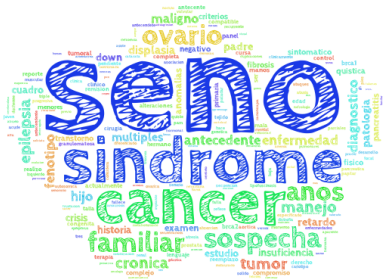

Figure 6 Wordcloud cluster 1

Figure 7 Cluster 1

The figure 6 represents the cluster 1 with the frequency of words that were grouped for this cluster and frequency of words is shown, being breast, syndrome and cancer are the most frequent words, along with ovarian, familial suspicion and epilepsy. Figure 7 represents the distribution of patients by age and gender within the group by age range over a 10-year interval. The first 10 rules obtained are represented by the figure 8 that shows the association of variants with clinical information of cluster 1. The results are without removing the synonymous variants. For the present cluster there are two types of variants distributed by the male gender with nonsynonymous variants and that are patients aged between 10 and 20 years, the allelic

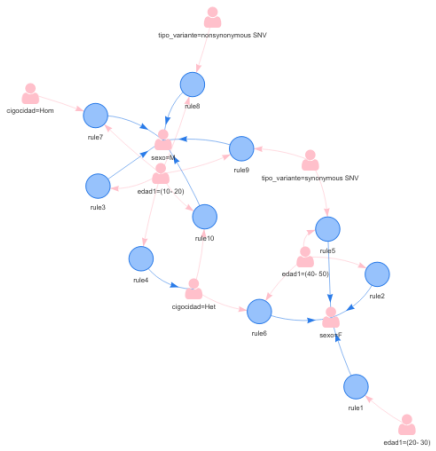

Figure 8 Cluster 1 association rules with synonymous variants.

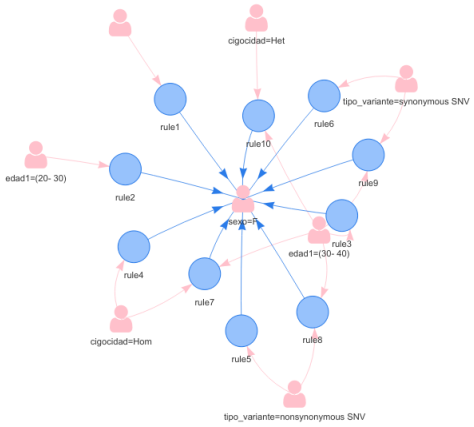

Figure 11 Cluster 2 association rules with synonymous variants.

state of the variants is homozygous, for this group a high difference is observed in the rules both genders, where female patients have synonymous variants with heterozygous status.

Cluster 2

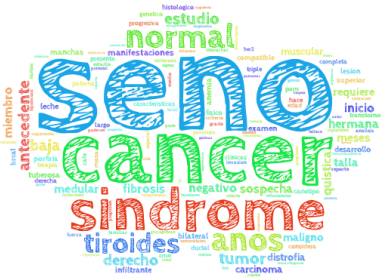

Figure 9 Wordcloud cluster 2

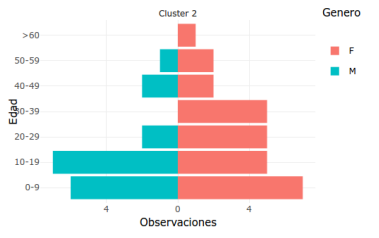

Figure 10 Cluster 2

In the figure 9 is observed that like cluster 1 the most frequent words are cancer, breast and syndrome, but words like thyroid and sister antecedents appear, according to the figure 10 is observed that age ranges between 20 and 30 years, and over 40 there are no male patients, this being a cluster represented mainly by female patients.

The figure **??**, shows us only an association of variants to the female gender, which corresponds to the low representation of male patients.The distribution of the homozygous allelic state occurs more frequently with patients of the age of 30 and 40 years, and they are variants of non-synonymous type, although the age range is not the most representative of the cluster, it is the one with the highest frequency of variants.

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)
6. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D.A., Bamshad, M.J., Akey, J.M.: Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature **493**(7431), 216–220 (2013). doi:10.1038/nature11690
7. McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.B., Donnelly, P.: Choice of transcripts and software has a large effect on variant annotation. Genome Medicine **6**(3) (2014). doi:10.1186/gm543
8. Babraham Bioinformatics: FASTQC manual (2016). http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/Help/3 Analysis Modules/ Accessed 2016-06-25
9. Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W.R., Zandi, P.P.: Validation and assessment of variant calling pipelines for next-generation sequencing. Human genomics **8**(1), 14 (2014). doi:10.1186/1479-7364-8-14
10. Tetreault, M., Bareke, E., Nadaf, J., Alirezaie, N., Majewski, J.: Whole-exome sequencing as a diagnostic tool: Current challenges and future opportunities. Informa Healthcare (2015). doi:10.1586/14737159.2015.1039516.
11. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management **24**(5), 513–523 (1988)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning **12**, 2825–2830 (2011). doi:10.1007/s13398-014-0173-7.2. arXiv:1201.0490v2

Figures

Figure 12 Sample figure title. A short description of the figure content should go here.

Figure 13 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.