Algorithm for classification of biological data based on data mining.

Eduardo Moniz Garcia
Technological Research Department
University of Mogi das Cruzes – UMC
São Paulo, Brasil
moniz@umc.br

Simone A. S. Fonseca
University of Mogi das Cruzes – UMC
São Paulo, Brasil
simoneap@umc.br

Jorge R. Beingolea
Technological Research Department
University of Mogi das Cruzes – UMC
São Paulo, Brasil
jorgegaray@umc.br

***Abstract*— - The study of genetic changes is regarded as being of paramount importance, since it can yield a greater understanding of the genetic expression and its consequences, such as: the anticipated forecast of certain types of diseases. The task of identifying changes in the DNA sequence (deoxyribonucleic acid), hitherto not described after next generation sequencing analysis has become one of the main activities of bioinformatics due to the capacity to analyze and interpret a wide range of genetic data. Numerous software applications were designed for purposes of sequence aligning, and subsequently identifying genetic changes. This study aims to establish a method that prepares genomic data and the discovery of existing correlations between changes in DNA sequence and other nitrogen bases, with the use of association rule algorithm using data mining, aiming to identify correlations between nucleotides of a DNA sequence, the correlation is made between nucleotides that significantly alter the DNA sequence and the other nucleotides of the analyzed DNA sequence. The purpose of this study is to identify nucleotide correlations of DNA sequences still unknown and to acquire a better understanding of the DNA structure.**

Keywords*— *DNA, Bioinformatics, algorithm.

I. INTRODUCTION

Bioinformatics has been extensively used when comparing genomes, intended to achieve a better understanding of the expression of gene characteristics present in living beings [1][2][3][4]. DNA is a chemical substance having molecules with genetic instructions responsible for the development and functioning of living beings. Phosphate, sugar called deoxyribose and the four nitrogenous bases: adenine, guanine, cytosine and thymine form the structure of the DNA molecule. The structure of the DNA molecule is composed of two sugar and phosphate bands and the sequence of nitrogenous base pairs. Every strand contains four types of nucleotide subunits that are joined by hydrogen bonds between the bases of the nucleotides. Nucleotides are molecules formed by nitrogenous bases, phosphate and pentose.

Changes in the DNA sequence comprise a basis of huge information about genetic variation; amid the changes, the change of a single nucleotide, which differs in the genome among distinct individuals of the same species [5] [6], may contribute as a genetic variability without, necessarily causing a type of mutation. They are extremely important in identifying genetic markers of disease or in population studies [7].

The goal of this study is to validate an association rule method between a certain change existing in the DNA sequence and its other nitrogenous bases in order to discover correlations still unknown, and to improve the understanding of the genomic structure.

The association rule discovery is the process of analyzing the existing relationships among attributes of a transactional database, in order to find associations or correlations [8]. As a result, an algorithm was developed that prepares the genomic data and identifies possible associations between DNA sequence changes and other nitrogenous bases (genomic architecture) of a genome. An algorithm consists of a sequence of instructions that must be completed in a finite time interval to perform a certain task. In the perspective of analyzing the algorithm applicability, the experiments were performed using the DNA sequence of the human BRCA1 gene (breast cancer 1). Genes are a functional unit of heredity comprising nucleic acids, carriers of genetic data.

There are several software applications that extract knowledge about information contained in DNA sequences [9], such as: CLC genomics Workbench, Panati, Varscan, Bowtie2, SAMtools. These tools are based on the technique known as DNA sequence alignment [10], with which it is possible to compare and identify changes in certain nucleotides. Identification of changes in nucleotides is done by computational tools from a reference genome and from a file containing re-sequencing of another variety of the same species. Unlike sequence alignment, the method proposed in this study extracts knowledge of the DNA sequence using a data mining approach to correlate DNA sequence changes with other nitrogenous bases. The sequence alignment is based on a previously known reference sequence, where two sequences are compared, and their differences checked. The method proposed in this study uses the algorithm of discovery of association rules based on data mining to extract knowledge of the DNA sequences, identifying correlations between nucleotides of the same sequence.

The chief advantage of the sequence alignment technique in relation to the proposed method is the rapid identification of known changes in the analyzed DNA sequences. Nevertheless, the algorithm of discovery of association rules based on data mining has a greater capacity to infer the identification of changes in the sequence of DNA still unknown.

II. DESCRIPTION OF DATA MODEL

Data sources - The proposed method has two main activities: to prepare data from the DNA sequence (obtained in silico) and to identify the existing associations between changes of the DNA sequence and other nucleotides, of a given gene, using algorithm of discovery of association rules based on data mining.

The identification of associations between changes in DNA sequence and other nucleotides of a given gene is made from a file containing the genetic variations. The sequence

978-1-7281-3967-8/19/$31.00 ©2019 IEEE

contained in rs80357829 and sequence rs397509082, belonging to the BRCA1 gene obtained from the Ensembl biological database (European Molecular Biology LabsInstitute) on February 9, 2019, were used. The experiments to demonstrate the association rules were performed with those using the proposed method.

Preparation of nucleotide sequence data - The association rules identification algorithm uses as input files in the FASTA format used to represent nucleotide sequences, containing the nucleotide sequence with a specific DNA change, organized in the flanking sequence format [11]. The nucleotide sequences in the flanking format have a specific size and the change contained in the DNA sequence is located at the center of the sequence. The change, which is present in the BRCA1 gene transcript 202-ENST00000354071.7, inserts TT nucleotides into the DNA sequence rs80357829, is located in the center of the sequence, bounded by brackets ([-/TT]), as shown in Fig. 1.

Fig. 1: Flanking sequence rs80357829.

```
TGAAATCAGATATGGAGAGAAATCTGTATTAACAGTCTGAACTACTTCTTCATATTCTTG
CTTTTTTATTTCAGGATGCTTACAATTACTTCCAGGAAGACTTTGTTTATAGACCTCAGG
TTGCAAAACCCCTAATCTAAGCATAGCATTCAATTTTGGCCCTCTGTTTCTACCTAGTTC
TGCTTGAATGTTTTCATCACTGGAACCTATTTCATTAATACTGGAGCCCACTTCATTAGT
ACTGGAACCTACTTCATTAATATTGCTTGAGCTGGCTTCTTTAAAAACATTTTCTCTAAT
GTTATTACGGCTAATTGTGCTCACTGTACTTGGAATGTTCTCATTTCCCATTTCTCTTTC
AGGTGACATTGAATGTTCCTCAAAGTTTTCCTCTAGCAGA[-/TT]TTTTTCTTACATTT
AGTTTTAACAAATGACTTGATGGGAAAAAGTGGTGGTATACGATATGGGTTTTGTAAAAG
TCCATGTTTATTTGGAGTAATGAGTCCAGTTTCGTTGCCTCTGAACTGAGATGATAGACA
AAACCTAGAGCCTCCTTTGATACTACATTTGGCATTATCAACTGGCTTATCTTTCTGACC
AACCACAGGAAAGCCTGCAGTGATATTAACTGTCTGTACAGGCTTGATATTAGACTCATT
CTTTCCTTGATTTTCTTCCTTTTGTTCACATTCAAAAGTGACTTTTGGACTTTGTTTCTT
TAAGGACCCAGAGTGGGCAGAGAATGTTGCACATTCCTCTTCTGCATTTCCTGGATTTGA
AAACGGAGCAAATGACTGGCGCTTTG
```

A change in the sequence rs397509082 confers a nucleotide insert thymine (TT), which is present in the transcript 202-ENST00000354071.7 of the BRCA1 gene, as shown in Fig. 2.

Fig. 2: Flanking sequence rs397509082.

```
CCTCATTTGTTTGGAAGAACCAATCAAGAAAGGATCCTGGGTGTTTGTATTTGCAGTCAA
GTCTTCCAATTCACTGCACTGTGAAGAAAACAAGCTAGCAGAACATTTTGTTTCCTCACT
AAGGTGATGTTCCTGAGATGCCTTTGCCAATATTACCTGGTTACTGCAGTCATTTAAGCT
ATTCTTCAATGATAATAAATTCTCCTCTGTGTTCTTAGACAGACACTCGGTAGCAACGGT
GCTATGCCTAGTAGACTGAGAAGGTATATTGTTTACTTTACCAAATAACAAGTGTTGGAA
GCAGGGAAGCTCTTCATCCTCACTAGATAAGTTCTCTTCTGAGGACTCTAATTTCTTGGC
CCCTCTTCGGTAACCCTGAGCCAAATGTGTATGGGTGAAA[-/TT]GGGCTAGGACTCCT
GCTAAGCTCTCCTTTCTGGACGCTTTTGCTAAAAAACAGCAGAACTTTCCTTAATGTCATT
TTCAGCAAAACTAGTATCTTCCTTTATTTCACCATCATCTAACAGGTCATCAGGTGTCTC
AGAACAAACCTGAGATGCATGACTACTTCCCATAGGCTGTTCTAAGTTATCTGAAATCAG
ATATGGAGAGAAATCTGTATTAACAGTCTGAACTACTTCTTCATATTCTTGCTTTTTTAT
TTCAGGATGCTTACAATTACTTCCAGGAAGACTTTGTTTATAGACCTCAGGTTGCAAAAC
CCCTAATCTAAGCATAGCATTCAATTTTGGCCCTCTGTTTCTACCTAGTTCTGCTTGAAT
GTTTTCATCACTGGAACCTATTTCAT
```

III. ALGORITHM OF DISCOVERY OF ASSOCIATIONS

This algorithm has as main objective to discover rules of associations, aiming to identify associations between the changes in DNA sequence and other nitrogen bases of a given gene. Two elements are used to establish the association rules: transaction and attribute. The transaction can be understood as an event to be analyzed, such as purchases made by a customer in a particular pharmacy, the attribute is one of the elements of that event (such as a particular product purchased by the client). The association between attributes is established when they frequently appear together in the same transaction, indicating that when an attribute exists in an event, another attribute will also be part of that event. Upon analyzing the event to make purchases (transactions) of clients in a pharmacy, we can establish a rule of association between the attributes client and product where, the rule client -> product, establishes that when the client attribute occurs the product attribute is also present. The client -> product rule is formed by the client premise and has the consequence product completion.

A. *Concepts used in the association rule*

Every transaction in the context of the association rule encompasses a series of items, which are part of a set of items belonging to the application domain. In the application domain, there exists a set of items I= {i1, ..., im}, where a transaction is composed of a subset of I, that is, T={i1,..., il} such that i1 ⊂ I and 1 <= m.

Table 1 illustrates a generic example of a transactional database in which the transaction is identified by Rid, with two forms of presentation of the data sets.

TABLE 1: TRANSACTIONAL BASE: (a) BY SETS; (b) AND MATRIX.

R1 = { i1,i3}
R2 = {i1,i2,i3}
R3 = {i2}
R4 = {i1,i4}
(a)

Rid	*i1*	*i2*	*i3*	*i4*
R1	1	0	1	0
R2	1	1	1	1
R3	0	1	0	1
R4	1	0	0	1

(b)

In the matrix representation, values equal to zero show that the item does not happen in the transaction; if it equals 1, it occurs. The associations among items are set by values equal to 1.

A subset of items is named as itemset; when it equals {i1, i3} and is described as 2-itemset, equal to {i1, i3, i4} is called 3-itemset, among others.

The data mining algorithm, used in the discovery of association rules, calculates the amount in which each item occurs in the database, as well as the measures of *support* and *confidence*, which are used to validate the quality of the association rule result.

The *support* for an itemset measures the frequency with which the items appear together, expressed as a percentage. For example, in Table 1 the 2-itemset {i1, i3} has a *support* equal to 50%, since {i1, i3} appear together in two transactions (R1 and R2) of the existing four ones. The *support* indicates the frequency a given itemset possesses. The measure of *confidence* expresses the reliability of an association rule based on the probability of its occurrence. Represented as a percentage, it is calculated by the ratio between the rule *support* and the rule premise *support*. For example, on the basis of Table 1 the rule: {i1} -> {i3} has *confidence* equal to (50/75), the *support* of the rule {i1} -> {i3} is 50%, since it occurs in transactions R1 and R2, the premise i1 has *support* of 75%, as it appears in three transactions of the four existing ones. *Confidence* means that, at 50% of the times the premise (i1) occurs, the *conclusion* (i3) also occurs, it expresses the probability of the rule {i1} -> {i3}, given the premise {i1}.

Correlation rules using a DNA sequence are rules of associations established with the use of measures of *support* and *confidence* and by the existing correlation between the itemsets of the premise and of the *conclusion*. Using the correlation rule, it is possible to extract relevant information about the nucleotide sequences.

Table 1 presents the example, where analyzing the correlation between i1 and i3 and we have the rule {i1} ->

{i3} with *support* of 50% and *confidence* of 66,66%, it shows that i3 is present in 66,66 % of the transactions in which i1 appears, the 50% *confidence* indicates that the rule {i1} -> {i3} occurs in 50% of the transactions analyzed. The *lift* measure identifies whether the correlation is positive or negative, using the following calculation formula: *lift* = rule *confidence*/ *conclusion support*, where *lift*> 1 indicates positive correlation; *lift* = 1 indicates no correlation exists and *lift* <1 indicates negative correlation. Table 2 shows a fictitious example of the *lift* measure.

TABLE 2: EXAMPLE USING THE *LIFT* MEASURE.

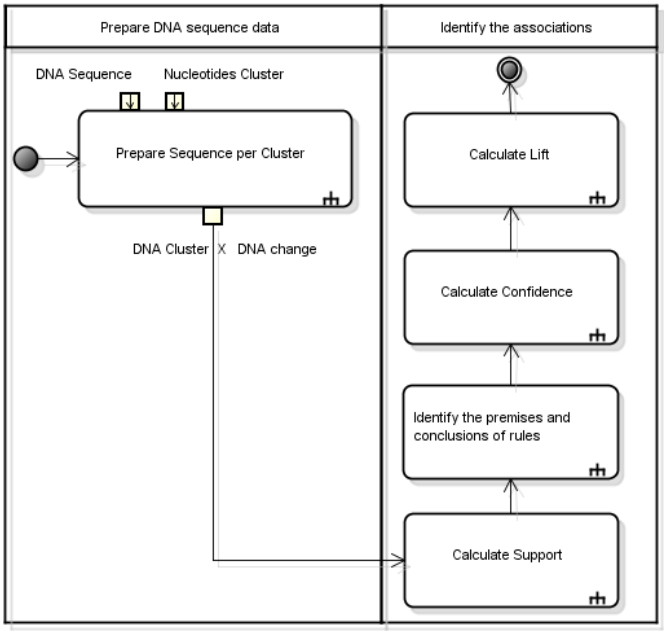
Rule	*Premise*	*Conclusion*	*Support*	*Confidence*	*Lift*
r1->r2	r1	r2	60%	70%	1.17
r2->r1	r2	r1	50%	40%	0.80
r1->r3	r1	r3	40%	38%	0.95
r3-->r1	r3	r1	70%	70%	1.00

Table 2 presents r1->r2 with *support* of 60% and *confidence* of 70% and indicates that r2 is present in 70% of transactions where r1 appears, and the *lift* of 1.17 expresses a positive correlation between r1 and r2.

The r1->r3 rule contained in Table 2 has a *support* of 40% and a *confidence* of 38%, and indicates that r3 is present in 38% of transactions where r1 appears, and the *lift* of 0.95 expresses a negative correlation between r1 and r3.Fig. 3 shows the activity diagram to prepare and identify associations.

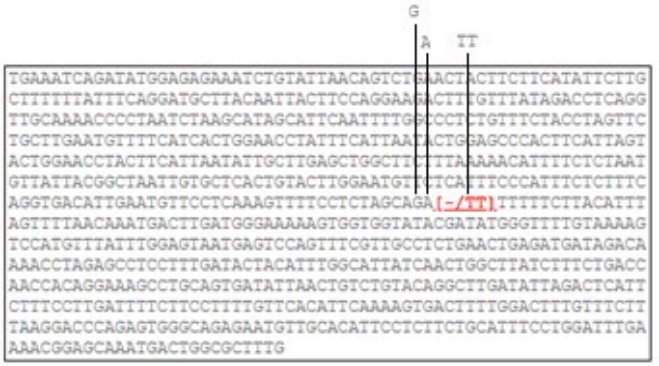
- **Sample Preparation -** The preparation of the DNA sequence data has as input information the DNA sequence containing the DNA change and the nucleotide cluster.

Fig. 3: Activity diagram to prepare and identify associations.



The activity of preparing the DNA sequence data generates a list containing all nucleotide clusters and DNA changes contained in the sequence. The cluster identifies how many nucleotides in the sequence will be used to establish the relationship with the DNA alteration. In Fig. 4 we have the change TT, which makes use of the cluster of size equal to 1 nucleotide, it establishes rules of association between the other nucleotides of the sequence (separated one by one) and the change TT, for example: A-> TT , G-> TT, and so on. Fig. 4 highlights the nucleotides A and T and the change (TT) used to compose the A-> TT and G-> TT rules. If we used a cluster equal to 2 we would have rules composed of two nucleotides and TT change, example: AG-> TT, AC-> TT, and so forth.

Fig. 4: Flanking Sequence rs80357829 with highlights of the nucleotides used in the rules A->TT and G->TT.



In the activity of association identification from the prepared list, the number of times that each item in the list occurs (the clusters and the change) is checked and the *support* measure is calculated, containing the combinations between the clusters and the DNA change to identify the premises and conclusions of rules. Next, the *confidence* and *lift* measures are calculated for each of the rules.

IV. DESCRIPTION RESULTS

Experiments - identification of associations existing in the nitrogenous bases of the DNA sequences rs80357829 and rs397509082 - The objective of the present experiment was to identify possible correlations between the change (TT) in sequences rs80357829 and rs397509082 and the other nucleotides contained in the BRCA1 gene 202-ENST00000354071.7 transcription. In this experiment, a sample of nucleotides containing the change rs80357829 and rs397509082 and a flanking sequence with a total size of 800 (eight hundred) nucleotides was used, with 400 positioned to the left and 400 located to the right of the change. The sequence used was obtained from the Ensembl database on Jan 09 2019. The present change in the sequences (rs80357829 and rs397509082) confers a nucleotide insert thymine (TT), significantly altering the DNA sequence and if it is an exon, it will generate a transcribed sequence and accordingly the amino acid sequence in the protein changes significantly. The experiment was performed by applying the association rules identification algorithm on the left and right sequences of the [TT] change contained in rs80357829 and rs397509082, considering a cluster of a nucleotide to establish the rules of correlation with the change [TT] and the other nucleotides.

TABLE 3: CORRELATION BETWEEN THE TT CHANGE AND THE NUCLEOTIDES LOCATED TO ITS LEFT OF THE SEQUENCES RS80357829 AND RS397509082.

Sequence	*Rule*	*Premise*	*Conclusion*	*Support*	*Confidence*	*Lift*	*Total*
rs80357829	A->TT	A	TT	12.88%	48.58%	0.97	103
rs80357829	C->TT	C	TT	10.50%	50.30%	1.01	84
rs80357829	G->TT	G	TT	7.50%	42.55%	0.85	60
rs80357829	T->TT	T	TT	19.13%	54.64%	1.09	153
rs397509082	A->TT	A	TT	13.63%	51.42%	1.03	109
rs397509082	C->TT	C	TT	10.38%	49.70%	0.99	83
rs397509082	G->TT	G	TT	10.13%	57.45%	1.15	81
rs397509082	T->TT	T	TT	15.88%	45.36%	0.91	127

Based on Table 3, it is observed that the T->TT rule extracted from the sequence rs397509082 shows the highest *support* in the value of 19.13%, indicating that the nucleotide T is present in 19.61% of the sequences located to the left of the TT change. The *confidence* equal to 54.64% indicates that in 54.64% of the times the T sequence occurs, the TT change is present. The *lift* of 1.09 establishes a strong correlation between the T nucleotide and the TT change, and the T->TT rule occurred 153 times in the sequence rs397509082. In the experiment, the T->TT rule of the sequence rs397509082 with a *support* value equal to 15.88% is also highlighted.

Subsequently, the correlation rules were identified on the nucleotides of sequences rs80357829 and rs397509082 located to the right of the TT change, a *confidence* measure and a minimum *support* of 6% were used.

TABLE 4: CORRELATION BETWEEN THE TT CHANGE AND THE NUCLEOTIDES LOCATED TO ITS RIGHT OF THE SEQUENCES RS80357829 AND RS397509082.

Sequence	*Rule*	*Premise*	*Conclusion*	*Support*	*Confidence*	*Lift*	*Total*
rs80357829	TT->A	TT	A	12.88%	49.05%	0.98	103
rs80357829	TT->C	TT	C	9.13%	44.51%	0.89	73
rs80357829	TT->G	TT	G	10.13%	56.64%	1.13	81
rs80357829	TT->T	TT	T	17.88%	50.53%	1.01	143
rs397509082	TT->A	TT	A	13.38%	50.95%	1.02	107
rs397509082	TT->C	TT	C	11.38%	55.49%	1.11	91
rs397509082	TT->G	TT	G	7.75%	43.36%	0.87	62
rs397509082	TT->T	TT	T	17.50%	49.47%	0.99	140

Table 4 shows that in the sequence rs80357829 the TT->T rule presents the highest *support* in the value of 17.88%, indicating that in 17.88% of the cases in which the TT change appears in the nucleotide sequences, the sequence T is also present. The *confidence* of 50.53% means that in 50.53% of the times that the change (TT) happens, the *conclusion* (T) also occurs. The *lift* of 1.01 indicates a strong correlation between the TT change and the T sequence, whereas the TT->T rule occurred 143 times in the rs80357829 sequence. The sequence rs397509082 that presented the rule TT->T with *support* value equal to 17.50% is also highlighted in the experiment.

V. DISCUSSION OF RESULTS

The largest *supports* per experiment contained in Tables 3 and 4 were used to *compose Table 5.*

TABLE 5: LARGEST SUPPORTS PER EXPERIMENT.

Sequence	*Rule*	*Premise*	*Conclusion*	*Support*	*Confidence*	*Lift*	*Total*
rs80357829	TT->T	TT	T	17.88%	50.53%	1.01	143
rs397509082	TT->T	TT	T	17.50%	49.47%	0.99	140
rs80357829	T->TT	T	TT	19.13%	54.64%	1.09	153
rs397509082	T->TT	T	TT	15.88%	45.36%	0.91	127

With Table 5 as a basis, it is possible to assume that there is a probability of around 17.00% of the TT premise to appear together with the T change and the probability around 16% of the T change to occur with the *conclusion* TT (*conclusion*) in the sequences rs80357829 and rs397509082.

The probability accuracy increase of the *support* indicators contained in Table 5 is directly interrelated with the increase of the number of experiments, the greater the number of experiments performed with the DNA sequences, the more accurate the *support*, *confidence* and *lift* indicators. The indication of a strong correlation between changes in the DNA sequence and other nucleotides may provide extremely important information for understanding the DNA structure. A nucleotide cluster for establishing correlation rules with the TT change contained in the sequences rs80357829 and rs397509082 was taken into account. However, this cluster could be of 2, 3, 4 or as many nucleotides as required for comparison with the DNA sequence change. The proposed method can be applied in any DNA sequence, just select the desired DNA sequence, identify the DNA change sequence and elect the nucleotide cluster.

The analysis of the results obtained with the experiments done shows that, with the use of the algorithm, correlations between the changes in DNA sequences and other nucleotides can be identified and obtain a better understanding of the structure of DNA.

VI. CONCLUSION AND PERSPECTIVES

The objective of this work was to create a DNA sequence analysis method. It consists of the activities of preparation of genomic data and application of algorithm based on the data mining association rule to identify possible correlations between changes in DNA sequence and other nucleotides in the flanking sequence.

The experiments were conducted to assess the probability of certain correlation rules occurring between DNA sequence changes and the other nucleotides belonging to a given gene.

On the basis of the analysis and comparisons made on the experiments performed, it was verified that the association rule discovery algorithm can be used for a better understanding of the DNA structure. Yet, the experiments were restricted to a single DNA sequence of the BRCA1 gene and the algorithm should be tested with a larger sequence number to corroborate its effectiveness.

The chief contributions of this study are the proposed method, which comprises the activities of genomic data preparation and the use of the association rule discovery algorithm among changes in DNA sequences and other nucleotides, as well as the analysis of the experiments performed using the algorithm.

With future works, it is intended to apply the data preparation and the association rule algorithm on a larger number of DNA sequence changes, to allow the database storage of the results generated by the association rule algorithm and, in this way, to facilitate comparisons between the generated rules, and finally to create a tool with a user-friendly visual interface to render easy use of the proposed method of preparation of genomic data and the discovery of correlation rules between DNA sequence data. This includes the possibility of using a cloud infrastructure to provide access to service data analysis [12].

REFERENCES

[1] VERLI, H. Bioinformática da Biologia à Flexibilidade Molecular. 1a. ed. São Paulo: Sociedade Brasileira de Bioquímica e Bilogia Molecular - SBBq, 2014.

[2] FERNANDES, L. A. Montagem e anotação funcional de sequências gênicas de Handroanthus impetiginosus (Mart. ex DC.) Mattos. 2015.

Tese (Mestrado em Genética e Biologia Molecular) – Instituto de Ciências Biológicas, Universidade Federal de Goiás. 2015.

[3] LIMA, R. S. Sistema Multiagente para Anotação Manual em Projetos de Sequenciamento de Genomas. 2007.

Tese (Mestrado em Informática) – Instituto de Ciências Exatas, Universidade de Brasília. 2007.

[4] HU, W. et al. Association mining of mutated cancer genes in different clinical stages across 11 cancer types. Oncotarget, v. 7, 2016,

[5] MOREIRA, L. M. Ciencias genomicas:Fundamentos e aplicações. 1a. ed. Ribeirão Preto - SP: Genética, Sociedade Brasileira de gentica, 2015.

[6] PANG, , H.; HAUSER, M.; MINVIELLE, S. Pathway-based identification of SNPs predictive of survival. European Journal of Human Genetics, v. 19, 2011.

[7] SALISBURY, B. A. et al. SNP and haplotype variation in the human genome. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis, v. 526, 2003.

[8] SILVA, L. A. DA; PERES, S. M.; BOSCARIOLI, C. Introdução à mineração de dados com aplicações em R. 1a. ed. São Paulo: Elsevier, 2016.

[9] ASSIS, H. DE; RIBEIRO, L. Desenvolvimento de um serviço de análise de sequências utilizando um modelo baseado em atributos de resultados de PSI-BLAST. 2013.

Tese (Doutorado em Bioinformática) – Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais. 2013.

[10] KOOGAN, G. Introdução à Genética. 11.ed. Rio de Janeiro: Guanabara, 2016.

[11] McEntyre , J.; Ostell, J. The NCBI Handbook. 3.ed. ethesda:National Center for Biotechnology Information , 2002.

[12] GARAY, Jorge R.B.; DE OLIVEIRA, A. M. ; TORRES, J. C. Z. ; ZUFFO, M. K. ; LOPES, R. D. Cloud Application Platform as a Service in Educational Environments. Recent Patents in Engineering, v. 9, 2015.