Algoritmo para classificação de dados biológicos baseado em mineração de dados

Eduardo Moniz Garcia, Simone A. S. Fonseca, Jorge R. Beingolea
Technological Research Department (NPT - UMC)
Universidade Mogi das Cruzes – UMC - Brazil
moniz@umc.br, simoneap@umc.br, jorgegaray@umc.br

Abstract - *O estudo das alterações genéticas é considerado de extrema importância, pois permite um melhor entendimento da expressão genética e suas consequências, tais como: a previsão com antecedência de alguns tipos de doenças. A tarefa de identificar alterações na sequência de DNA*(**d**eoxyribo**n**ucleic **a**cid), *até então não descritas, após análise de sequenciamento de próxima geração, tornou-se uma das principais atividades da bioinformática, devido à capacidade de analisar e interpretar grande quantidade de dados genéticos. Vários softwares foram criados com o objetivo de alinhar sequências e, posteriormente, identificar alterações genéticas. Este estudo tem como proposta estabelecer um método que faz preparação de dados genômicos e a descoberta de correlações existentes entre alterações na sequência de DNA e demais bases nitrogenadas, com o uso de algoritmo de regra de associação utilizando mineração de dados, visando identificar correlações entre nucleotídeos de uma sequência de DNA, a correlação é feita entre os nucleotídeos que alteram significativamente a sequência de DNA e os demais nucleotídeos da sequência de DNA analisada. O objetivo desse estudo é identificar correlações entre nucleotídeos de uma sequencias de DNA ainda não conhecidas e obter um melhor entendimento da estrutura do DNA.*

Palavras chaves: DNA, Bioinformática, algoritmo.

I. INTRODUCTION

A bioinformática tem sido muito utilizada na comparação de genomas, objetivando um melhor entendimento da expressão de características gênicas presentes nos seres vivos[1][2][3][4]. O DNA é uma substância química que possui moléculas com instruções genéticas responsáveis pelo desenvolvimento e funcionamento dos seres vivos. O fosfato, o açúcar denominado desoxirribose e as quatro bases nitrogenadas :adenina, guanina, citosina e timina formam a estrutura da molécula de DNA. A estrutura da molécula de DNA é composta de duas fitas de açúcar e fosfato e a sequência de pares de bases nitrogenadas. Cada fita contém quatro tipos de subunidades de nucleotídeos que são unidas por ligações de hidrogênio entre as bases dos nucleotídeos. Os nucleotídeos são moléculas formadas por bases nitrogenadas, fosfato e pentose.

As alterações na sequência do DNA constituem uma enorme fonte de informações sobre variação genética, entre as alterações, a mudança de um único nucleotídeo, que difere no genoma entre indivíduos de uma espécie[5][6], pode contribuir como uma variabilidade genética sem, necessariamente, causar um tipo de mutação. Eles são de extrema importância na identificação de marcadores genéticos de doenças ou em estudos populacionais[7].

O presente trabalho tem o objetivo de demonstrar um método de regra de associação entre uma determinada alteração existente na sequência de DNA e suas demais bases nitrogenadas visando descobrir correlações ainda não conhecidas, e trazer uma melhora no entendimento da estrutura genômica.

A descoberta de regra de associação é o processo de analisar os relacionamentos existentes entre atributos de uma base de dados transacional, com objetivo de encontrar associações ou correlações[8]. Para tanto, desenvolveu-se um algoritmo que faz a preparação dos dados genômicos e a identificação de possíveis associações entre as alterações na sequência do DNA e demais bases nitrogenadas (arquitetura genômica) de um genoma. Um algoritmo é composto por um uma sequência de instruções que devem ser executadas em um intervalo de tempo finito para realizar uma determinada tarefa. Com o intuito de analisar a aplicabilidade do algoritmo, os experimentos foram realizados utilizando a sequência de DNA do gene humano BRCA1(breast cancer 1). Os genes são uma unidade funcional da hereditariedade contendo ácidos nucleicos ,portadores de dados genéticos.

Existem vários softwares que extraem conhecimento sobre informações contidas em sequências de DNA[9], tais como: CLC genomics Workbench, Panati, Varscan, Bowtie2, SAMtools. Essas ferramentas são baseadas na técnica conhecida como alinhamento de sequência de DNA[10], com ela é possível comparar e identificar alterações em determinados nucleotídeos. A identificação de alterações nos nucleotídeos é feita pelas ferramentas computacionais a partir de um genoma referência e de um arquivo contendo ressequenciamento de outra variedade de mesma espécie. Diferentemente do alinhamento de sequência, o método proposto nesse estudo extrai conhecimento da sequência de DNA utilizando uma abordagem baseada em mineração de dados para correlacionar alterações na sequência de DNA com demais bases nitrogenadas. O alinhamento de sequência se baseia em uma sequência de referência previamente conhecida, onde, são comparadas duas sequencias e verificadas suas diferenças. O método proposto nesse estudo utiliza o algoritmo de descoberta de regras de associação baseado em mineração de dados para extrair conhecimento das sequências de DNA, identificando correlações entre nucleotídeos de uma mesma sequência.

A principal vantagem da técnica de alinhamento de sequência em relação ao método proposto, está na rápida identificação de alterações conhecidas nas sequências de DNA analisadas. Porém, o algoritmo de descoberta de regras de associação baseado em mineração de dados tem maior capacidade de inferir a identificação de alterações na sequência de DNA ainda não conhecida.

II. DESCRIPTION OF DATA MODEL

a) Fontes de dados - O método proposto contém duas atividades principais: preparar dados da sequência de DNA (obtida em *silico*) e identificar as associações existentes entre alterações da sequência de DNA e demais nucleotídeos, de um determinado gene, utilizando algoritmo de descoberta de regras de associação baseado em mineração de dados.

A identificação das associações entre alterações na sequência do DNA e demais nucleotídeos de um determinado gene é feita a partir de um arquivo contendo as variações genéticas. Foi utilizada a sequência contida em rs80357829e a sequência rs397509082, pertencente ao gene BRCA1 obtido no banco de dados biológicos do Ensembl (European Molecular Biology Laboratory - European Bioinformatics Institute) em 09 de fevereiro de 2019. Os experimentos para demonstrar as regras de associação foram realizados com esses dados, utilizando o método proposto.

b) Preparação dos dados da sequência de nucleotídeos - O algoritmo de identificação de regras de associações utiliza como entrada arquivos tipo texto no formato FASTA, contendo a sequência de nucleotídeos com uma alteração de DNA específica, organizado no formato de sequência flanqueadora[11]. As sequências de nucleotídeos no formato flanqueadora tem um tamanho específico e a alteração contida na sequência de DNA está localizado no centro da sequência. A alteração, que está presente na transcrição 202-ENST00000354071.7 do gene BRCA1, insere nucleotídeos TT na sequência de DNA rs80357829, está localizado no centro da sequência, delimitado por colchetes ([-/TT]), conforme mostra a Figura 1.

Figura 1: Sequência flanqueadorars80357829.

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

Fonte: http://www.ensembl.org

Alteração presente na sequência rs397509082 confere uma inserção de nucleotídeos timina (TT), que está presente na transcrição 202-ENST00000354071.7 do gene BRCA1, conforme apresentado na Figura 2.

Figura 2: Sequência flanqueadora rs397509082.

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

Fonte: http://www.ensembl.org

III. ALGORITHM OF DISCOVERY OF ASSOCIATIONS

Esse algoritmo tem como principal objetivo descobrir regras de associações, visando identificar associações entre as alterações na sequência de DNA e demais bases nitrogenadas de um determinado gene. Dois elementos são utilizados para estabelecer as regras de associação: transação e atributo. A transação pode ser entendida como um evento a ser analisado, como compras efetuadas por cliente em uma determinada farmácia, o atributo é um dos elementos desse evento (como um determinado produto comprado pelo cliente).A associação entre atributos é estabelecida quando eles frequentemente aparecem juntos em uma mesma transação, indicando que, quando existe um atributo em um evento, outro atributo também fará parte desse mesmo evento. Ao analisarmos o evento efetuar compras (transação) de clientes em uma farmácia, podemos estabelecer uma regra de associação entre os atributos cliente e produto onde, a regra cliente -> produto, estabelece que, quando ocorre o atributo cliente o atributo produto também está presente. A regra cliente -> produto é formada pela premissa cliente e tem como consequência a conclusão produto.

- **Conceitos utilizados na regra de associação** - Cada transação no contexto da regra de associação contém uma série de itens, os quais fazem parte de um conjunto de itens pertencentes ao domínio da aplicação. No domínio da aplicação, existe um conjunto de itens I= {i1,...,im}, sendo que uma transação é composta por um subconjunto de I, ou seja, T={i1,..., il} , tal que i1 ⊂ I e l <=m .

A Tabela 1 representa um exemplo genérico de uma base de dados transacional em que a transação é identificada por Rid, com duas formas de apresentação dos conjuntos de dados.

Tabela 1:Base transacional: (a) por conjuntos; (b) e matricial.

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)

Na representação matricial, os valores iguais a zero assinalam que o item não acontece na transação; se igual a 1, ele acontece. As associações entre os itens são estabelecidas por valores iguais a 1.

Um subconjunto de itens é nomeado como *itemset*; quando ele é igual a {i1, i3) e é descrito como 2-itemset, igual a {i1,i3,i4) é denominado 3-itemset, dentre outros.

O algoritmo de mineração de dados, utilizado na descoberta de regras de associação, calcula a quantidade em que cada item ocorre na base de dados, bem como as medidas de *suporte* e *confiança*, as quais são utilizadas para validar a qualidade do resultado da regra de associação.

O *suporte* sobre um *itemset* mede a frequência com que os itens aparecem juntos, expressado em percentual. Por exemplo, na Tabela 1 o 2-itemset {i1, i3} tem *suporte* igual a 50%, visto que {i1,i3} aparecem juntos em duas transações (R1 e R2), das quatros existentes. O *suporte* diz qual a frequência tem um determinado *itemset*. A medida de *confiança* expressa a confiabilidade de uma regra de associação com base na probabilidade de sua ocorrência. Representada em percentual, ela é calculada pela razão entre *suporte* da regra e o *suporte* da premissa da regra. Por exemplo: com base na Tabela 1 a regra:{i1} -> {i3}, tem a *confiança* igual a (50/75) , o *suporte* da regra {i1} -> {i3} é de 50%,pois ela ocorre nas transações R1 e R2, a premissa i1 tem *suporte* de 75%,pois aparece em três transações das quatro existentes. A *confiança* significa que, em 50% das vezes em que a premissa(i1) acontece, a conclusão(i3) também acontece, ou seja, ela expressa a probabilidade da regra {i1} -> {i3}, dada a premissa {i1}.

A regra de correlação utilizando uma sequência de DNA são regras de associações estabelecidas com o uso das medidas de *suporte* e *confiança* e pela correlação existente entre os *itemsets* da premissa e da conclusão.

Por meio da regra de correlação, é possível extrair informações relevantes sobre as sequências de nucleotídeos.

A Tabela 1 apresenta o exemplo , onde analisando a correlação entre i1 e i3 e temos a regra {i1}->{i3} com *suporte* de 50% e *confiança* de 66,66%,ela demonstra que i3 está presente em 66,66% das transações em que i1aparece, a *confiança* de 50% indica que a regra {i1}->{i3} ocorrerem 50% das transações analisadas. A medida *lift* identifica se a correlação é positiva ou negativa, utilizando-se a seguinte fórmula de cálculo: *lift* = *confiança* da regra / *suporte* da conclusão, onde *lift*> 1 indica correlação positiva; *lift*=1 indica a não existência de correlação e *lift*<1 indica correlação negativa. A Tabela 2, apresenta um exemplo fictício da medida *lift*.

Tabela 2:Exemplo com o uso da medida *lift*.

Regra	Premissa	Conclusão	*Suporte*	*Confiança*	*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

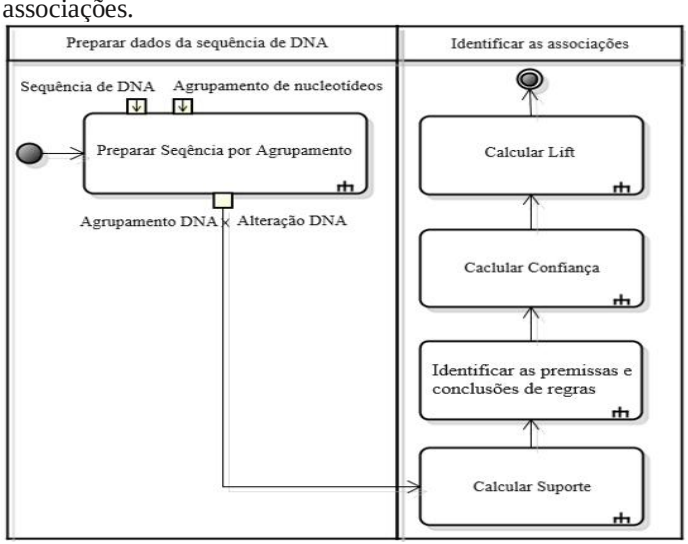
A Tabela2 apresenta r1->r2 com *suporte* de 60% e *confiança* de 70, indica que r2está presente em 70% das transações em que r1 aparece, e o *lift* de 1,17 expressa uma correlação positiva entre r1 e r2.

A regrar1->r3 contida na Tabela 2 tem *suporte* de 40% e *confiança* de 38%, indica que r3 está presente em 38% das transações em que r1 aparece, e o *lift* de 0,95 expressa uma correlação negativa entre r1 e r3.

A Figura 3 apresenta o diagrama das atividades de preparação de dados e identificação de associações.

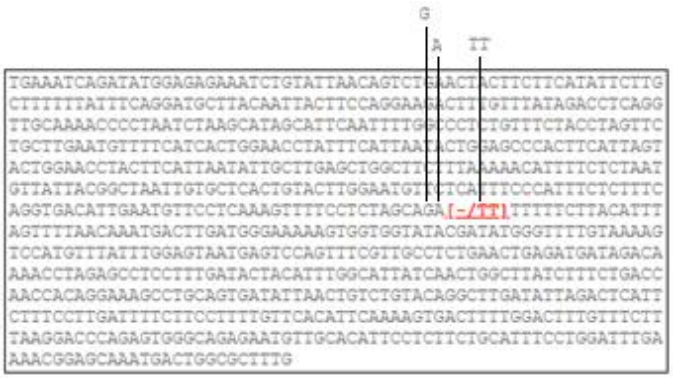
- **Preparação de Amostra -** A preparação de amostra dos dados da sequência de DNA tem como informações de entrada a sequência de DNA contendo a alteração de DNA e o agrupamento de nucleotídeos.

Figura 3: Diagrama de atividades preparar e identificar associações.



A atividade de preparar os dados da sequência de DNA gera uma lista contendo todos os agrupamentos de nucleotídeos e a alteração de DNA contidos na sequência. O agrupamento identifica quantos nucleotídeos da sequência serão utilizados para estabelecer a relação com a alteração de DNA. Na Figura 4 temos a alteração TT, que faz uso do agrupamento de tamanho igual a 1 nucleotídeo , ela estabelece regras de associação entre os demais nucleotídeos da sequência (separados de um em um) e a alteração TT, por exemplo : A->TT,G->TT, e assim por diante. A Figura 4 destaca os nucleotídeos A e T e a alteração (TT) utilizados para compor as regras A->TT e G->TT. Caso utilizássemos um agrupamento igual a 2 teríamos regras composta de dois em dois nucleotídeos e alteração TT, exemplo : AG->TT , AC->TT , e assim por diante.

Figura 4: Sequência franqueadora rs80357829 com destaques dos nucleotídeos usados nas regras A->TT e G->TT.



Fonte: http://www.ensembl.org

Na atividade de identificação de associação a partir da lista preparada verificada a quantidade de vezes que cada item da lista (os agrupamentos e a alteração) ocorrem e é calculada a medida de *suporte*, contendo as combinações entre os agrupamentos e a alteração de DNA para identificar as premissas e conclusões de regras. Em seguida são calculadas as medidas de *confiança* e lift para cada uma das regras.

IV. DESCRIPTION RESULTS

Experimentos – identificação de associações existentes nas bases nitrogenadas das sequências de DNA rs80357829 ers397509082 - O objetivo deste experimento foi identificar possíveis correlações existentes entre a alteração (TT) existente nas sequencias rs80357829e rs397509082e os demais nucleotídeos contidos na transcrição 202-ENST00000354071.7 do gene *BRCA1*. Utilizou-se, neste experimento, uma amostra de nucleotídeos contendo a alteração rs80357829 e rs397509082 e uma sequência flanqueadora com tamanho total de 800 (oitocentos)nucleotídeos, sendo 400 posicionados à esquerda e 400 localizados à direita da alteração. A sequência utilizada foi obtida no banco de dados do *Ensembl*nodia09 jan. /2019. A alteração presente nas sequências (rs80357829 e rs397509082) confere uma inserção de nucleotídeos timina (TT), alterando significativamente a sequência de DNA e, caso seja um exon, irá gerar uma sequência transcrita e consequentemente a sequência de aminoácidos na proteína muda significativamente. O experimento foi realizado aplicando-se o algoritmo de identificação de regras de associação sobre as sequências à esquerda e à direita da alteração [TT] contida em rs80357829 e rs397509082, considerando-se um agrupamento de um nucleotídeos para estabelecimento das regras de correlação com a alteração [TT] e os demais nucleotídeos.

Tabela 3:Correlação entre a alteração TT e os nucleotídeos localizados a sua esquerdadas sequências rs80357829 e rs397509082.

Sequência	Regra	Premissa	Conclusão	*Suporte*	*Confiança*	*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**

Com base na Tabela 3, observa-se que a regra T->TT extraída da sequência rs397509082 apresenta o maior *suporte* no valor de 19,13 % , indicando que o nucleotídeo T junto com a alteração TT está presente em 19,61% das sequências localizadas à esquerda da alteração TT. A *confiança* igual a 54,64% indica que, em 54,64 % das vezes que ocorre a alteração TT, o nucleotídeo T também ocorre. O *lift* de 1,09 estabelece forte correlação entre o nucleotídeo T e a alteração TT, sendo que a regra T->TT ocorreu 153 vezes na sequência rs397509082. No experimento, destacam-se também a regra T->TT da sequência rs397509082 com valor *suporte* superior a 1%.

Em seguida, as regras de correlação foram identificadas sobre os nucleotídeos das sequências rs80357829 e rs397509082 localizados à direita da alteração TT, utilizou-se uma medida de *confiança* e *suporte* mínimo de 6%.

Tabela 4:Correlação entre a alteração TT e os nucleotídeos localizados a sua direita das sequências rs80357829 e rs397509082.

Sequência	Regra	Premissa	Conclusão	*Suporte*	*Confiança*	*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**

A Tabela 4 evidencia que na sequência rs80357829 a regra TT->T apresenta o maior *suporte* no valor de 17,88%, indicando que a alteração TT junto com o nucleotídeo T está presente em 17,88 % das sequências localizadas à direita da alteração TT.

A *confiança* de 50,53% significa que, em 50,53% das vezes em que a conclusão (T) acontece, a premissa (TT) também ocorre. O *lift* de 1,01 indica forte correlação entre a alteração TT e a sequência T, sendo que a regra TT->T ocorreu 143 vezes na sequência rs80357829. Destacam-se também no experimento a sequência rs397509082 que apresentou a regra TT->T com valor de *suporte* igual a 17,50 %.

V. DISCUSSION OF RESULTS

Os maiores *suporte*s por experimento contidos nas Tabelas 3 e 4 foram utilizados para compor a Tabela 5.

Tabela 5: Maiores *suporte*s por experimento.

Sequência	Regra	Premissa	Conclusão	*Suporte*	*Confiança*	*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

Tendo a Tabela 5 como base, é possível inferir que existe uma probabilidade em torno de 17,00 % da premissa TT aparecer junto com a alteração Te a probabilidade em torno de 16 % da alteração T ocorrer com a conclusão TT (conclusão) nas sequências rs80357829 e rs397509082.

O aumento da acuracidade da probabilidade dos indicadores de *suporte* contidos na Tabela 5 está diretamente relacionada ao aumento da quantidade de experimentos, quando maior a quantidade de experimentos realizados com as sequências de DNA, mais preciso serão os indicadores de *suporte*, *confiança* e *lift*. A indicação de uma correlação forte entre alterações na sequência de DNA e demais nucleotídeos pode ser uma informação de extrema importância para entendimento da estrutura de DNA. Considerado um agrupamento de um nucleotídeos para estabelecimento das regras de correlação com a alteração TT contida nas sequênciasrs80357829 e rs397509082. Porém, esse agrupamento poderia ser de2,3,4 ou quantos nucleotídeos for necessário comparar com a alteração na sequência de DNA.O método proposto pode ser aplicado em qualquer sequência de DNA basta, selecionar a sequência de DNA desejada, identificar a sequência de alteração no DNA e eleger o agrupamento de nucleotídeos.

A análise dos resultados obtidos com os experimentos feitos demonstra que, com o uso do algoritmo, correlações entre as alterações nas sequencias de DNA e demais nucleotídeos podem ser identificadas e, obter-se um melhor entendimento da estrutura do DNA.

VI. CONCLUSION AND PERSPECTIVES

Este trabalho teve como objetivo criar um método de análise de sequencias de DNA, composto pelas atividades de preparação de dados genômicos e aplicação de algoritmo baseado na regra de associação da mineração de dados para identificar possíveis correlações entre alterações na sequência do DNA e demais nucleotídeos na sequência franqueadora.

Os experimentos foram realizados para avaliar a probabilidade de determinadas regras de correlação que ocorrerem entre as alterações de sequência do DNA e os demais nucleotídeos pertencentes ao um determinado gene.

Com base nas análises e comparações feitas sobre os experimentos realizados, constatou-se que o algoritmo de descoberta de regra de associação pode ser utilizado para um melhor entendimento da estrutura de DNA. Porém, os experimentos ficaram restritos a uma única sequência de DNA do gene *BRCA1* e o algoritmo deve ser testado com uma sequência de maior tamanho e assim poderá validar sua eficiência.

As principais contribuições deste estudo são o método proposto, que contém as atividades de preparação de dados genômicos e o uso do algoritmo de descoberta de regras de associações entre alterações das sequencias do DNA e demais nucleotídeos, bem como a análise dos experimentos realizados com o algoritmo.

Como trabalhos futuros pretende-se aplicar a preparação de dados e o algoritmo de regras de associação sobre uma quantidade maior de alterações de sequência de DNA, permitir o armazenamento em banco de dados dos resultados gerados pelo algoritmo de regra de associação e, dessa forma, facilitar as comparações entre as regras geradas, e por fim criar uma ferramenta com interface visual amigável com objetivo de tornar fácil o uso do método proposto de preparação de dados genômicos e de descoberta de regras de correlação entre dados de sequência de DNA. Inclui-se a possibilidade de utilizar uma infraestrutura de nuvem para fornecer o acesso como serviço à análise dos dados [12].

REFERÊNCIAS

[1]. VERLI, H. **Bioinformática da Biologia à Flexibilidade Molecular**. 1a. ed. São Paulo: Sociedade Brasileira de Bioquímica e Bilogia Molecular - SBBq, 2014. v. 53

[2]. FERNANDES, L. A. Montagem e anotação funcional de sequências gênicas de Handroanthus impetiginosus (Mart. ex DC.) Mattos. p. 1–63, 2015.

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

[4]. HU, W. et al. Association mining of mutated cancer genes in different clinical stages across 11 cancer types. **Oncotarget**, v. 7, n. 42, 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, n. 6, 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, n. 1–2, p. 53–61, 2003.

[8]. (SILVA; PERES; BOSCARIOLI, 2016)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.

[10]. KOOGAN, G. (ED.). Introdução à Genética. 11a edição ed. Rio de Janeiro: Guanabara, 2016.

[11]. NCBI. The NCBI Handbook - NCBI Bookshelf. n. Md, p. 1–10, 2012.

[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, p. 69-76, 2015.