

UNIVERSIDADE FEDERAL DO RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

**Descoberta de Possíveis Biomarcadores Utilizando
Análise Topológica de Redes de Co-expressão Gênica**

Eduardo Soares de Abreu

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Computação

Orientador: Prof. Dr. Adriano Velasque Werhli
Co-orientador: Prof. Dr. Mariana Recamonde Mendoza

Rio Grande, 2019

Ficha catalográfica

A162d Abreu, Eduardo Soares de.

Descoberta de possíveis biomarcadores utilizando análise topológica de redes de co-expressão gênica / Eduardo Soares de Abreu. – 2019.

91 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Computação, Rio Grande/RS, 2019.

Orientador: Dr. Adriano Velasque Werhli.

Coorientadora: Dra. Mariana Recamonde Mendoza.

1. Redes Biológicas 2. Expressão Gênica 3. Aprendizado de Máquina 4. Descoberta de Biomarcadores I. Werhli, Adriano Velasque II. Mendoza, Mariana Recamonde III. Título.

CDU 577.2:004



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

DISSERTAÇÃO DE MESTRADO

Descoberta de Possíveis Biomarcadores Utilizando Análise Topológica de Redes de Co-expressão Gênica

Eduardo Soares de Abreu

Banca examinadora:

Prof. Dr. Eder Maiquel Simão

Prof. Dr. Marcelo Rita Pias

Prof. Dr. Adriano Velasque Werhli
Orientador(a)

Prof. Dr. Mariana Recamonde Mendoza
Coorientador(a)

*Dedico... à minha namorada Mirian, companheira de todos os momentos.
Aos meus familiares e aos meus orientadores.*

AGRADECIMENTOS

Agradeço à minha família pelo apoio e à minha namorada Mirian, por todo carinho, paciência, incentivo. Sem vocês eu não chegaria ao fim desta jornada.

Agradeço também aos meus orientadores, por todo apoio, paciência e ensinamentos recebidos, permitindo completar este trabalho.

Aos colegas de laboratório pela atenção e auxílio.

Ao meu colega Wolmer, companheiro de trabalho nas disciplinas e pela sua amizade.

*Não é bom ficar furioso se você empacou [em uma ideia].
O que eu faço é continuar pensando sobre o problema.
Mas trabalhando em outra coisa.*

— STEPHEN HAWKING

RESUMO

ABREU, Eduardo Soares de. **Descoberta de Possíveis Biomarcadores Utilizando**

Análise Topológica de Redes de Co-expressão Gênica. 2019. 91 f. Dissertação de (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

O avanço em técnicas de coleta de dados a partir da biologia molecular, permitiu a captura de informações de milhares de genes simultaneamente, como o Microarranjo e o RNA-Seq, tornando-se um importante meio para o entendimento dos processos biológicos envolvidos no desenvolvimento das doenças complexas como câncer, doenças cardíacas e degenerativas. A busca para extrair conhecimento a partir dos dados coletados por essas técnicas não é uma tarefa trivial, por trabalhar com o mapeamento dos níveis de transcrição dos genes em larga escala. Considerando um grande desafio enfrentado devido ao pequeno número de experimentos realizados na literatura e a grande dimensionalidade (genes), se faz necessário a proposta de novas metodologias para extração de conhecimento desses dados. Este trabalho teve por objetivo geral propor uma metodologia para identificação de possíveis novos biomarcadores por meio da análise de topologias de redes biológicas utilizando algoritmos de aprendizagem de máquina supervisionada e técnicas de teoria de redes complexas. Nessa direção esse trabalho contribuiu propondo um metodologia para avaliação dos dados de expressão gênica utilizando redes complexas e analisando suas topologias e empregando aprendizagem de máquina supervisionado para descoberta de possíveis biomarcadores. Como resultados o algoritmo SVM foi o que obteve os melhores resultados sobre o *Random Forest* e o J48.

Palavras-chave: Redes Biológicas, Expressão Gênica, Aprendizado de Máquina, Descoberta de Biomarcadores.

ABSTRACT

ABREU, Eduardo Soares de. **Methods for Discovering Biomarkers from Gene Expression Data**. 2019. 91 f. Dissertação de (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

Advances in data collection techniques from molecular biology allows the capture of information from many of simultaneous genes such as Microarray and RNA-Seq, creating an important means for understanding the biological processes involved in the development of complex diseases such as cancer, heart and degenerative diseases. The search for knowledge from the data collected by these techniques is not a trivial task due to the fact that it works with the mapping of gene transcription levels on a large scale. Considering the difficult challenge faced due to the small number of experiments performed in the literature and the large dimensionality (genes), it is necessary to propose new methodologies for knowledge extraction from these data. This work aimed to propose a methodology to identify possible new biomarkers through the analysis of biological network topologies using supervised machine learning algorithms and complex network theory techniques. Thus, this work contributed by proposing a methodology for the evaluation of gene expression data using complex networks and analyzing their topologies and employing supervised machine learning to discover possible biomarkers. As a result, the SVM algorithm obtained the best results over Random Forest and J48.

Keywords: Biological Networks, Gene Expression, Machine Learning, Discovery of Biomarkers.

LISTA DE FIGURAS

Figura 1	Representação da Dupla Helice do DNA	18
Figura 2	Representação do RNA com relação ao DNA.	19
Figura 3	Dogma Central da Biologia Molecular	20
Figura 4	Fluxo do RNA-Seq.	23
Figura 5	Esquema de hibridização e análise dos dados.	24
Figura 6	Biomarcadores identificados com base em dados Omicos.	25
Figura 7	Formação de Redes Aleatórias.	29
Figura 8	Exemplo de uma Small World Network.	30
Figura 9	Construção do modelo Barabási-Albert. Os novos vértices preferem se ligar aos nós mais conectados. Devido às propriedades de crescimento e ligação preferencial podemos perceber o surgimento de <i>hubs</i>	31
Figura 10	Grafo com 5 nós e 6 ligações sendo A nó com maior <i>degree</i> e B com menor <i>degree</i>	34
Figura 11	Exemplo da técnica de balanceamento <i>SMOTE</i>	38
Figura 12	Matriz de confusão e suas metricas.	39
Figura 13	Exemplo de Curva ROC.	40
Figura 14	Validação cruzada utilizando 4 <i>folds</i>	41
Figura 15	Visão geral do <i>framework</i>	43
Figura 16	Fluxo de trabalho da abordagem proposta	44
Figura 17	Visão geral da Metodologia	49
Figura 18	Correlação de Pearson e Descoberta do Limiar	50
Figura 19	Gráfico para Escolha do Limiar	51
Figura 20	Construção e Analise da Rede	52
Figura 21	Descoberta de Possíveis Biomarcadores	54
Figura 22	Gráfico de desbalanceamento de classes no conjunto de treinamento .	69
Figura 23	Gráfico AUC médio sem balanceamento por algoritmo	70
Figura 24	Gráfico AUC da média SVM down sample vs sem balanceamento . .	71
Figura 25	Gráfico AUC da média Random Florest down sample vs sem balanceamento	72
Figura 26	Gráfico AUC da média J48 down sample vs sem balanceamento . . .	73
Figura 27	Gráfico AUC da média SVM up sample vs sem balanceamento	73
Figura 28	Gráfico AUC da média RF up sample vs sem balanceamento	74
Figura 29	Gráfico AUC da média J48 up sample vs sem balanceamento	75
Figura 30	Arvores J48 up sample biomarcador vs não biomarcador	76
Figura 31	Gráfico AUC da média SVM Smote vs sem balanceamento	77

Figura 32	Gráfico AUC da média Random Forest Smote vs sem balanceamento	77
Figura 33	Gráfico AUC da média J48 Smote vs sem balanceamento	78
Figura 34	Gráfico da media e desvio padrão por balanceamento de classe	79
Figura 35	Gráfico da media e desvio padrão por algoritmo	80
Figura 36	Gráfico da média AUC e desvio padrão sem balanceamento vs <i>down sample</i>	80
Figura 37	Gráfico da média AUC e desvio padrão sem balanceamento vs <i>SMOTE</i>	81
Figura 38	Gráfico da média AUC e desvio padrão sem balanceamento vs <i>up sample</i>	81

LISTA DE TABELAS

Tabela 1	Matriz de Adjacência Fonte: Dados do Autor	33
Tabela 2	Comparação entre os trabalhos	48
Tabela 3	Matriz de Expressão Gênica	50
Tabela 4	Extração carasterísticas do nós	52
Tabela 5	Extração carasterísticas do nós	54
Tabela 6	Conjunto de dados balanceado para treinamento e validação	56
Tabela 7	Conjunto de dados sem balanceado para treinamento e validação	57
Tabela 8	Comparação dos trabalhos relacionados com Biomark Finder	58
Tabela 9	Limiar de correlação de Pearson usado para identificação de genes comuns	60
Tabela 10	Resultado obtidos executando SVM sem balanceamento	62
Tabela 11	Resultado obtidos executando Random Forest sem balanceamento	63
Tabela 12	Resultado obtidos executando J48 sem balanceamento	64
Tabela 13	Resultado obtidos executando SVM com balanceamento	66
Tabela 14	Resultado obtidos executando Random Forest com balanceamento	67
Tabela 15	Resultado obtidos executando J48 com balanceamento	68

LISTA DE ABREVIATURAS E SIGLAS

DNA	Ácido Desoxirribonucléico
DVC	Doenças Cardiovasculares
GEO	<i>Gene Expression Omnibus</i>
mRNA	Ácido Ribonucleico Mensageiro
OMS	Organização Mundial de Saúde
RNA	Ácido Ribonucleico
ROC AUC	Área da Curva ROC
ROC	Receiver Operating Characteristic
rRNA	Ácido Ribonucleico Ribossômico
tRNA	Ácido Ribonucleico transportador
WGCNA	<i>Weighted Correlation Network Analysis</i>
WAN	<i>Wide Area Network</i>

SUMÁRIO

1	INTRODUÇÃO	15
1.1	Motivações	16
1.2	Objetivos	16
1.2.1	Objetivos específicos:	17
1.3	Organização	17
2	FUNDAMENTAÇÃO TEÓRICA	18
2.1	Ácidos Desoxirribonucléicos e Ribonucléicos	18
2.1.1	Proteínas	19
2.2	Dogma Central da Biologia Molecular	20
2.3	Ciências Ômicas	21
2.3.1	Genoma	21
2.3.2	Transcriptoma	22
2.3.3	Dados Biológicos	22
2.4	Biomarcadores	24
2.4.1	Identificação de Genes Biomarcadores	25
2.4.2	Conjunto de Genes Biomarcadores	26
2.4.3	Vias Biológicas como Biomarcador	26
2.4.4	Redes de Biomarcadores	27
2.5	Redes Complexas	28
2.5.1	Random Networks	29
2.5.2	Small World Network	30
2.5.3	Scale-free Network	31
2.5.4	Medidas de Centralidade	32
2.6	Aprendizado de Máquina	36
2.6.1	Aprendizado Supervisionado	36
3	TRABALHOS RELACIONADOS	42
3.1	Redes de Biomarcadores em diabetes	42
3.2	Rede de Biomarcadores biomarcadores para leucemia linfocítica crônica	43
3.3	Rede de Biomarcadores de melanoma metastático	45
3.4	Redes de Biomarcador para infarto agudo do miocárdio	46
4	BIOMARKER FINDER	49
4.1	<i>Biormarker Finder</i>	49
4.1.1	Escolha dos dados de Expressão Gênica	50
4.1.2	Correlação de Pearson e Descoberta do Limiar	50

4.1.3	Construção das Redes	51
4.1.4	Extração de atributos e dados de treinamento	51
4.1.5	Descoberta de Possíveis Biomarcadores	53
5	RESULTADOS	60
5.1	Resultados	61
5.1.1	Resultado sem aplicação de balanceamento de classes	61
5.2	Resultados aplicando balanceamento de classes	65
5.3	Discussão dos Resultados	69
6	CONCLUSÃO	83
	REFERÊNCIAS	85

1 INTRODUÇÃO

Doenças complexas, como Câncer, Alzheimer, Parkinson e Cardiovasculares, são geralmente causadas por mutações genéticas e / ou influências ambientais, envolvendo vários processos biológicos. Dentre essas, as doenças cardiovasculares (DCV), causam anualmente a morte de cerca de 17,9 milhões de pessoas no mundo. Segundo a Organização Mundial de Saúde (OMS), vítimas de ataques cardíacos, representam aproximadamente 31% de todas as mortes globais (WHO et al., 2017).

Diante do impacto das doenças cardiovasculares no cenário da saúde mundial como fator gerador de incapacidade aos indivíduos afetados, assim como, os grandes gastos que implica para a saúde pública, essa enfermidade vem sendo objeto de estudo de diferentes frentes de pesquisas. Dentre estas as Ômicas (genômica, proteômica, metabolômica entre outras), e o emprego de meios computacionais para avaliação deste grande volume de dados e extração de conhecimento (DE LIZ SOUZA; AUGUSTO RHODEN; ALENCAR PAMPHILE, 2014), têm se tornado cada vez mais recorrentes na detecção precoce das cardiomiopatias e tornou-se uma forma crucial de prevenir lesões muitas vezes irreversíveis ao coração, levando a morte. O diagnóstico oportuno das cardiomiopatias fornece indicativos importantes sobre o planejamento de tratamentos precisos no estágio inicial da doença (CLAESSEN et al., 2018).

Umas das formas que vem sendo empregada é a utilização de biomarcadores para auxiliar no diagnóstico, prognóstico e na pesquisas de novos fármacos, para tratamento de doenças complexas como as cardíacas. Os biomarcadores representam entidades que podem ser medidas experimentalmente e atuam como indicativos de ocorrência de uma determinada função normal ou patológica de um organismo ou uma resposta a um agente farmacológico (ANDRE et al., 2014).

Para descoberta de possíveis novos biomarcadores uma estratégia utilizada é o emprego de dados de expressão gênica, para construção de redes de co-expressão gênica. Esta abordagem baseia-se na ideia de que uma doença complexa é raramente consequência de alterações em um único gene, mas resultado de mudanças em um conjunto de genes relacionados (ANDRE et al., 2014).

Este trabalho tem como acrônimo Biomarker Finder, e contribui para descoberta de

possíveis novos biomarcadores, através da análise de redes de co-expressão gênica por meio de caracterização topológica e visa aprendizagem de máquina.

1.1 Motivações

A abordagem reducionista frequentemente adotada pela pesquisa científica considera a divisão de sistemas em partes, que são analisadas individualmente, de modo a entender o comportamento global desses sistemas. Apesar de grandes avanços obtidos por essa metodologia, ela não se mostrou adequada para análise de sistemas ditos complexos, formados por componentes individuais que se influenciam mutuamente (BARABÁSI; GULBAHCE; LOSCALZO, 2011), (NEWMAN; BARABASI; WATTS, 2011). Um exemplo dessa limitação ocorre no contexto dos perfis de expressão gênica.

Para compreender os mecanismos que ocorrem na evolução dos processos biológicos para a progressão de uma doença complexa, são necessários além da análise de um gene exclusivo, o entendimento de como o sistema biológico se relaciona com os demais genes, podendo influenciar a função de outros genes ou ser influenciado por eles (ANDRE et al., 2014) (BARABÁSI; GULBAHCE; LOSCALZO, 2011).

Sendo assim, o estudo de sistemas complexos exige uma abordagem, que tem como princípio o fato de que o comportamento desses sistemas não pode ser explicado apenas com a soma de seus componentes individuais (BARABÁSI; ALBERT, 1999).

Uma forma de representar e analisar esses sistemas é através de grafos. Visto que estes sistemas apresentam um alto grau de irregularidade na distribuição de suas conexões, além de padrões não triviais, tais como de comunidades e motifs, esse grafos são denominados de redes complexas (BARABÁSI; GULBAHCE; LOSCALZO, 2011).

A inferência de redes a partir de dados de expressão gênica é uma tarefa desafiadora, pois geralmente o número de variáveis (genes) analisados é muito grande, dando uma característica complexa aos dados (ROY; BHATTACHARYYA; KALITA, 2014).

Uma abordagem que vem sendo empregada para contornar esse desafio é a utilização da caracterização topológica para explorar e interpretar a rede com auxílio mineração de dados para extração de conhecimento (PINHO, 2017).

1.2 Objetivos

Esta dissertação de mestrado tem como objetivo propor uma metodologia para identificação de possíveis novos biomarcadores por meio da análise de topologias de redes biológicas utilizando algoritmos de aprendizagem de máquina supervisionada e técnicas de teoria de redes complexas.

1.2.1 Objetivos específicos:

- Investigar a estrutura de redes de co-expressão geradas a partir de dados de expressão gênica;
- Comparar características topológicas entre as redes de co-expressão de pacientes com cardiomiopatia dilatada e indivíduos saudáveis;
- Descobrir biomarcadores empregando análise topológica e aprendizagem de máquina;
- Investigar a existência de padrões topológicos para biomarcadores conhecidos em contraste com genes não associados à doença.

1.3 Organização

No Capítulo 2 são apresentados conceitos fundamentais da Biologia Molecular, as formas utilizadas para descoberta de biomarcadores. No Capítulo 3 são apresentadas trabalhos relacionados . No Capítulo 4 apresentado Biomarker Finder. O Capítulo 5 são apresentados resultados utilizando Biomarker Finder. No Capítulo 6 Conclusão do presente trabalho e no Capítulo 7 o cronograma.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados conceitos necessários para realização e compreensão desta pesquisa.

2.1 Ácidos Desoxirribonucléicos e Ribonucléicos

O DNA (ácido desoxirribonucléico) consiste em duas longas fitas, cada uma composta de unidades chamadas fosfatos, moléculas de açúcar e nucleotídeos, ligados em série, formando estruturas denominadas bases nucleotídicas. Existem quatro tipos de nucleotídeos possíveis no DNA: adenina (A), guanina (G), citosina (C) e timina (T) (PASSARGE, 2009).

Assim, moléculas de DNA são usualmente representadas por uma sequência de símbolos correspondentes às bases nucleotídicas da fita, A,G,C,T. As duas fitas de DNA se encontram ligadas através de pontes de hidrogênio entre suas bases nucleotídicas, segundo regras de paridade, nas quais adenina se liga apenas com a timina (A – T), e a guanina apenas com a citosina (G – C), formando uma estrutura em dupla hélice. Dessa maneira, as fitas de DNA são exatamente complementares entre si, como pode ser visualizado na Figura 1.

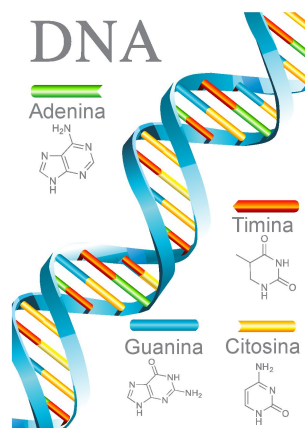


Figura 1: Representação da Dupla Hélice do DNA. Fonte: adaptada de (PASSARGE, 2009).

O RNA (ácido ribonucleico) é uma molécula de ácido ribonucleico formada por uma única cadeia com grande diversidade de conformações. A sequência de bases é similar à do DNA, exceto pela substituição da desoxirribose por ribose e da base nucleotídica timina por uracila. Essa substituição pode ser visualizada na Figura 2.

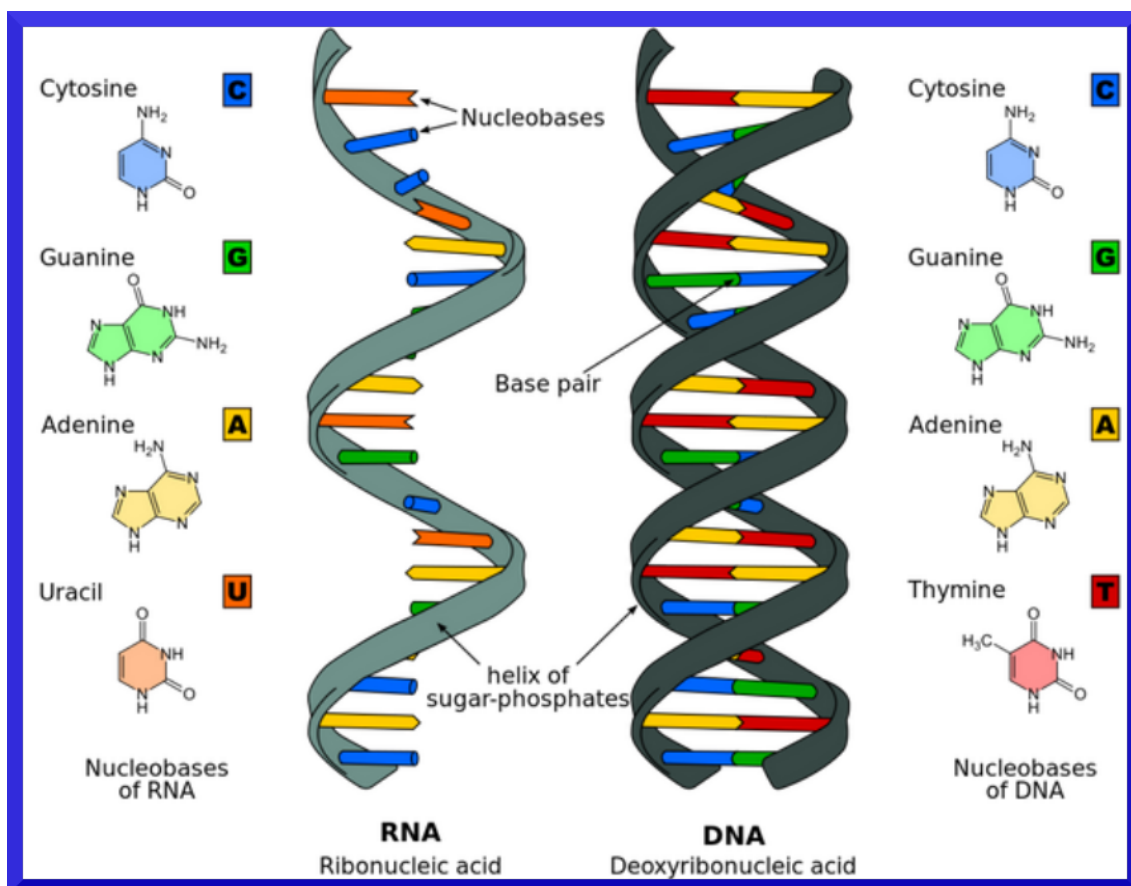


Figura 2: Representação do RNA com relação ao DNA. Fonte: Adaptado de (PASSARGE, 2009).

O RNA possui três classes principais: RNA mensageiro (mRNA), que contém a informação genética para coordenar a produção de sequência de aminoácidos; RNA transportador (tRNA), que identifica e transporta as moléculas de aminoácidos até o ribossomo e RNA ribossômico (rRNA), que acumula-se formando regiões conhecidas como nucléolos, nesses locais o rRNA combina-se com proteínas e origina os ribossomos.

2.1.1 Proteínas

As proteínas são sintetizadas a partir da molécula de DNA e atuam nos processos metabólicos e estruturais de um organismo. Cada proteína tem a sua própria forma tridimensional e tipicamente possui de 1.000 a 50.000 átomos. Embora exista uma grande variação de estrutura e funcionalidade entre as proteínas, todas elas podem ser representadas por uma sequência linear dos aminoácidos que as compõem (ZAHA; FERREIRA;

PASSAGLIA, 2014).

Esta sequência é chamada de estrutura primária da molécula de proteína. Entretanto, a estrutura primária de uma proteína, em geral, não é suficiente para determinar sua forma tridimensional, a qual está intimamente relacionada com as suas propriedades e funções num organismo. É uma tarefa extremamente difícil inferir com precisão a estrutura tridimensional de uma proteína baseado na sua sequência primária (PASSARGE, 2009).

2.2 Dogma Central da Biologia Molecular

O Dogma Central da Biologia Molecular, proposto por Francis Crick em 1958, explica como ocorre o fluxo de informações do código genético em organismos (PASSARGE, 2009). Esse modelo mostra principalmente que uma sequência de DNA pode formar uma proteína, com auxílio de diferentes atividades de RNAs. Segundo esse dogma, o fluxo da informação genética segue no sentido DNA para proteína, conforme mostrado na Figura 3.

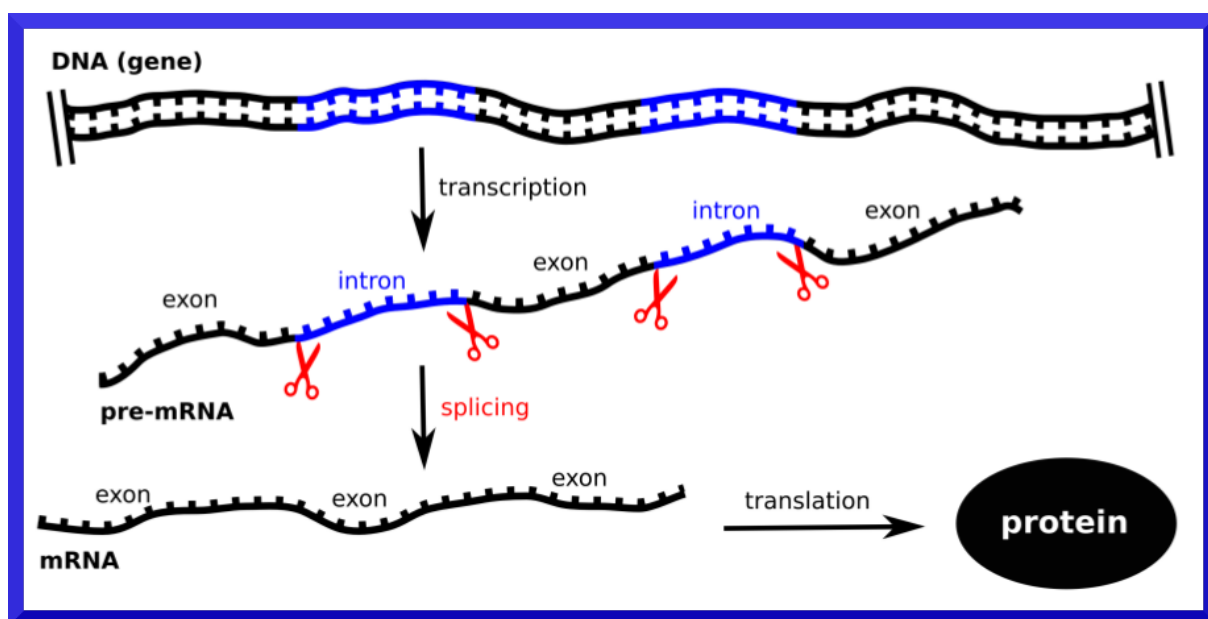


Figura 3: Dogma Central da Biologia Molecular. Fonte: Adaptado de (MACKERETH; SATTLER, 2012).

O processo de síntese de uma proteína ocorre no citoplasma da célula, enquanto as informações necessárias para construí-la se encontram no DNA. Para que o processo de síntese ocorra é necessário que haja uma transferência da informação presente no DNA para os ribossomos, estruturas responsáveis pela montagem da proteína através da concatenação de aminoácidos (ZAHA; FERREIRA; PASSAGLIA, 2014).

De forma simplificada, o transporte da informação codificada ocorre da seguinte maneira. Quando a célula necessita de uma determinada proteína, a informação presente no gene que codifica esta proteína deve ser copiada. As duas fitas do DNA são então sepa-

radas com a ajuda de enzimas especiais na região correspondente ao gene que está sendo solicitado. O conteúdo do gene é então copiado de forma complementar em uma fita de mRNA, este processo é conhecido como transcrição (ZAHA; FERREIRA; PASSAGLIA, 2014).

O mRNA, por sua vez, se associa a um ribossomo presente no citoplasma. Os ribossomos são então responsáveis por interpretar a informação codificada em forma de RNA, associando as sequências de três nucleotídeos (códon) aos seus aminoácidos correspondentes, ligando-os um a um e sintetizando a proteína. Esta etapa é chamada tradução (ZAHA; FERREIRA; PASSAGLIA, 2014).

Este processo descrito pelo dogma central através do qual a informação codificada, em um gene é transformado em um produto funcional como uma proteína, é conhecido como expressão gênica.

Embora não seja o foco deste trabalho, importante mencionar que algumas mudanças já foram propostas no modelo original. Isso ocorreu em razão de hoje se saber, por exemplo, que nem todos os RNAs produzem proteínas (ZAHA; FERREIRA; PASSAGLIA, 2014).

2.3 Ciências Ômicas

Os dados ômicos são objetos dos métodos de estudo da biologia celular e molecular e visam realizar uma análise global dos sistemas biológicos.

O genoma, que corresponde a sequência completa do material genético de um organismo, e o transcriptoma, que determina quais são os genes que estão sendo expressos e como esses níveis de expressão podem mudar durante a vida de um organismo, estão relacionados com as proteínas e suas características demonstrando grandes variações entre as células desse mesmo organismo (DE LIZ SOUZA; AUGUSTO RHODEN; ALENCAR PAMPHILE, 2014).

O proteoma é o produto da proteômica, que tem como objetivo a análise das proteínas e complementa a genômica e a transcriptômica, possibilitando a identificação das proteínas de um organismo e a compreensão de suas funções (DE LIZ SOUZA; AUGUSTO RHODEN; ALENCAR PAMPHILE, 2014).

2.3.1 Genoma

O genoma é o conjunto completo das informações do DNA e contém mais de três bilhões de pares de bases em um ser humano. Em um dado organismo quase todas as células contém o mesmo genoma em seus respectivos núcleos (SILVA, 2016). Aproximadamente 3,2 bilhões de pares do genoma humano resultante, 99,5% é comum a todos os seres humanos e apenas 0,5% representa a variação genética do indivíduo (AYDAY et al., 2015).

Sendo a representação biológica de um indivíduo, o genoma contém informações sobre sua herança genética e por meio da combinação desses dados com o ambiente em que o indivíduo está inserido, é possível saber se há predisposição a determinadas condições físicas e mentais, que incluem doenças como o câncer (ZAHA; FERREIRA; PASSAGLIA, 2014).

2.3.2 Transcriptoma

Transcriptoma é o conjunto de todos os transcritos que são produzidos nas células de um tecido podendo ser interpretado como um reflexo da expressão gênica. Estes transcritos correspondem a moléculas de mRNA, tRNA, rRNA, bem como outros RNAs não codificantes, isto é, que não codificam proteínas (PASSARGE, 2009).

Possuindo maior complexidade que o genoma, responsável por sua codificação, o estudo do transcriptoma é uma tecnologia de alto rendimento que determina como esse muda de acordo com diversos fatores, em um dado período de tempo, em um determinado estado biológico (ZAHA; FERREIRA; PASSAGLIA, 2014).

De acordo com SZABO (2014), dentre os principais métodos utilizados para a avaliação dos transcriptomas, os mais utilizados são, (i) Microarranjo e o (ii) Sequenciamento do RNA (RNA-seq).

2.3.3 Dados Biológicos

RNA-Seq

O termo RNA-Seq representa o transcriptoma revelado pelo sequenciamento de DNA complementar (cDNA) e é considerado pelos pesquisadores um método revolucionário, pois possui alta sensibilidade e pode ser utilizado para caracterizar o transcriptoma de um organismo (OLIVEIRA et al., 2015). Uma das suas grandes vantagens é a ausência quase total de ruídos e a capacidade de detectar um número elevado de cópias de mRNA por célula. A Figura 4 ilustra o fluxograma do RNA-Seq.

O RNA isolado é convertido em cDNA, uma molécula estável, que pode então ser amplificada e sequenciada. As leituras de sequência são alinhadas ao conjunto do genoma (somente sequência) para identificar suas localizações com base em combinações de nucleotídeos. O mapeamento das leituras para uma lista de anotação de genes irá gerar o número de leituras que se alinham com um gene específico e são chamadas contagens. Essas contagens em qualquer gene específico são representativas da quantidade de expressão gênica na amostra e podem ser comparadas para identificar genes diferencialmente expressos (MCQUEEN et al., 2015).

Microarranjo

O chip de DNA ou tecnologia de Microarranjo permite investigar a expressão de centenas ou milhares de genes em uma dada amostra usando uma reação de hibridização.

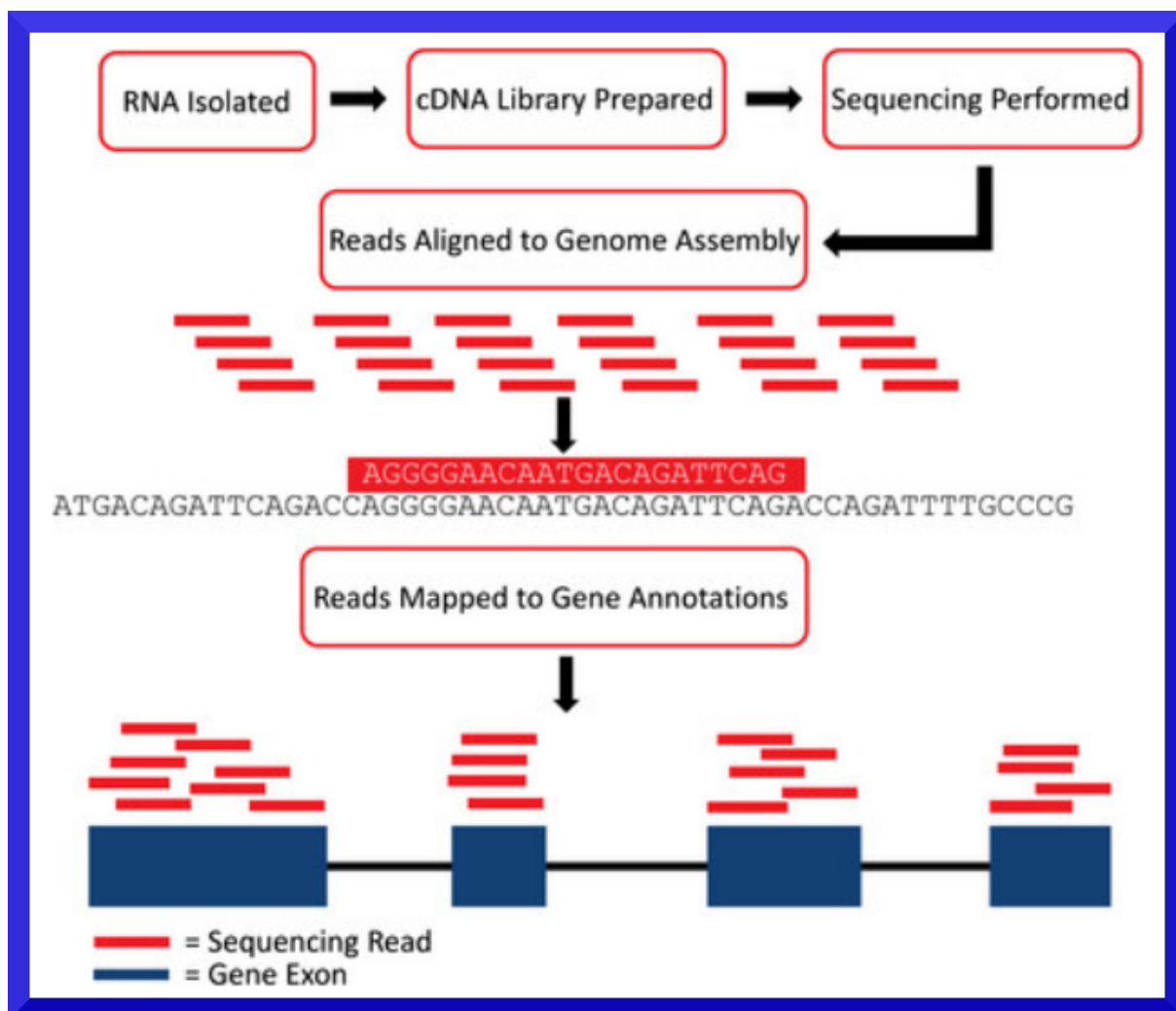


Figura 4: Fluxo do RNA-Seq. Fonte: Adaptado de (MCQUEEN et al., 2015).

Dois termos são muito utilizados no desenvolvimento da técnica: sondas e alvos. As sondas são os fragmentos gênicos conhecidos que estão imobilizados em uma matriz sólida, ou seja, as sequências correspondentes aos genes de interesse; os alvos consistem nos cDNAs marcados que estão em solução, os quais são provenientes das amostras biológicas (RUSSO; ZEGAR; GIORDANO, 2003).

Dessa forma, um arranjo pode ser definido como uma coleção ordenada de sondas, cada sonda representando uma única espécie de ácido nucleico e correspondendo ao gene de interesse. A tecnologia é baseada na hibridização entre sequências de cDNA alvos, derivadas das amostras de interesse que são marcadas com fluoróforos ou radioisótopos, com um arranjo de várias sondas ou alvos de DNA, que estão imobilizadas em uma matriz sólida. Essa matriz pode ser uma lâmina de vidro ou sílica ou uma membrana de nylon. Os alvos são produzidos a partir de RNA de amostras biológicas por transcrição reversa, marcados com fluoróforos específicos (Cy3 e Cy5) e co-hibridizados com as sondas de DNA (COLOMBO; RAHAL, 2010).

O sinal de hibridização produzido em cada sonda corresponde ao nível de expressão

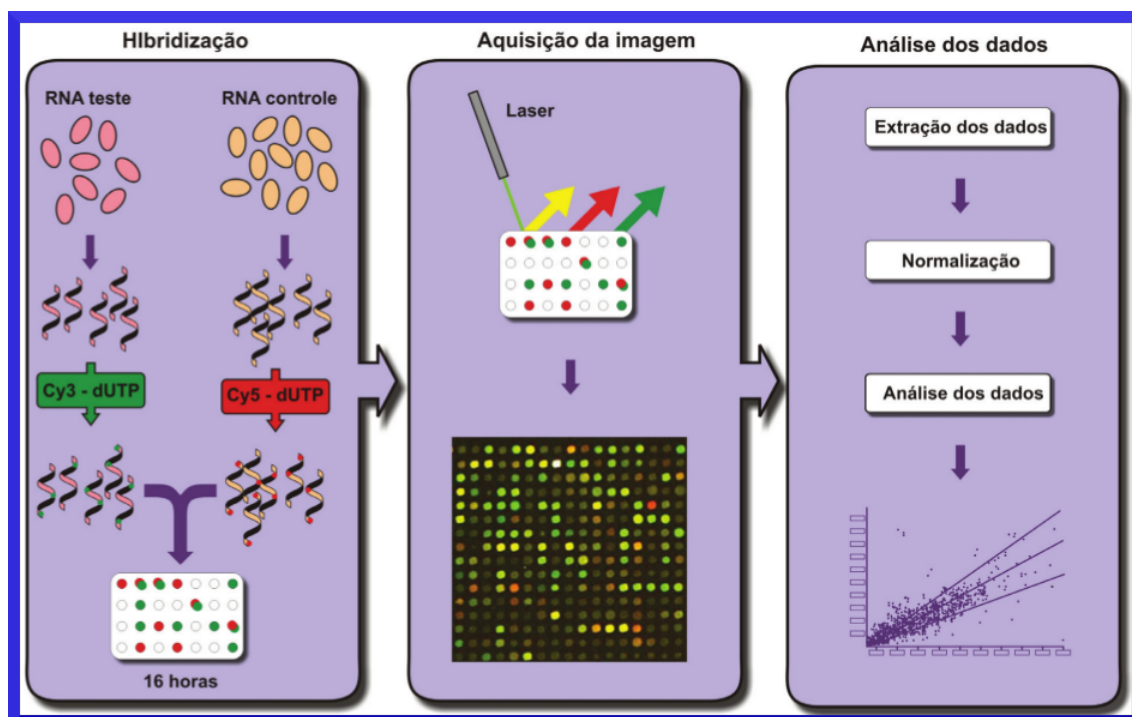


Figura 5: Esquema de hibridização e análise dos dados. Fonte: Adaptado de (COLOMBO; RAHAL, 2010).

do respectivo gene em determinada amostra no momento do estudo (RIJN; GILKS, 2004). Dessa forma, após a hibridização competitiva dos alvos marcados, os sinais são detectados, quantificados, integrados e normalizados com softwares.

2.4 Biomarcadores

As doenças geralmente são causadas por mutações genéticas e / ou influências ambientais, envolvendo vários processos biológicos. A descoberta de biomarcadores pode ajudar no diagnóstico precoce, na prevenção do desenvolvimento de doenças e no prognóstico preciso dos estados da doença, assim possibilitando administração de tratamentos necessários para bons resultados (KHERA; KATHIRESAN, 2017). Biomarcadores são marcadores de um processo biológico subjacente, seja a presença de determinada doença (*trait biomarker*), a progressão da doença (*state biomarker*) ou a resposta a um tratamento (*pharmacological biomarker*) (ANDRE et al., 2014).

Com avanços nas tecnologias e a redução de custo, uma grande quantidade de dados ômicos sobre doenças são disponíveis publicamente. Por exemplo, os dados de expressão de genes para diferentes condições fisiológicas e patológicas podem ser recuperados de *Gene Expression Omnibus* (BARRETT et al., 2008) e *ArrayExpress* (PARKINSON et al., 2008), dados de interação proteína-proteína estão disponíveis gratuitamente no BioGrid (OUGHTRED et al., 2016), entre outros meios de aquisição dos dados para análises (SILVA, 2016).

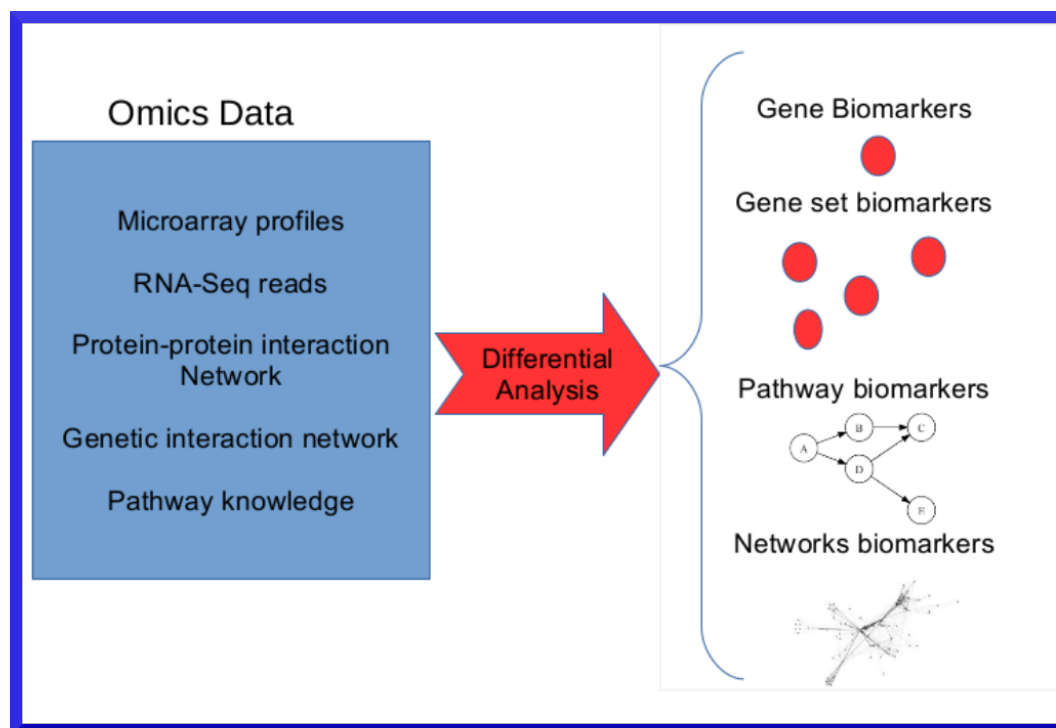


Figura 6: Biomarcadores identificados com base em dados Omicos Fonte: Adaptado de (ZHAO; QIN, 2013).

Os biomarcadores moleculares podem ser identificados com base em diferentes tipos de dados, onde os biomarcadores resultantes variam de genes individuais a conjuntos e redes de genes, como podemos visualizar na Figura 6.

2.4.1 Identificação de Genes Biomarcadores

Com a acumulação de grande quantidade de dados de expressão de genes depositados em bancos de dados públicos, está se tornando mais comum a busca por biomarcadores a partir da identificação de genes diferencialmente expressos entre amostras de casos e controles ou entre diferentes estágios de doenças (FUENTE, 2010). Geralmente os genes diferencialmente expressos têm o potencial para discriminar amostras de diferentes condições, podem ser considerados importantes para a investigação de biomarcadores (ZHAO; QIN, 2013).

As primeiras abordagens para identificar genes biomarcadores detectam genes diferencialmente expressos ao estabelecer um limiar, onde os genes cuja expressão mudam acima do limiar estabelecido são usados como biomarcadores de genes. O trabalho de DERISI; IYER; BROWN (1997), propôs para detectar genes diferencialmente expressos ao estabelecer um limiar de mudança de duas vezes seria considerado um biomarcador.

A abordagem estatística de análise de significância de microarranjo (SAM) proposta por TUSHER; TIBSHIRANI; CHU (2001) é uma das ferramentas mais utilizadas para determinar a significância das mudanças na expressão. O SAM atribui uma pontuação

a cada gene com base em sua mudança de expressão em relação ao desvio padrão das medidas avaliadas para esse gene, onde os genes com pontuação acima de um limite são considerados estatisticamente significativos.

Posteriormente, LAI et al. (2004) propôs uma estatística SAM melhorada, que utiliza o modelo de regressão linear penalizado para evitar a sobreposição considerando o grande número de genes e o número relativamente pequeno de amostras. Tanto o SAM quanto a sua versão melhorada podem ser vistos como uma redução dos testes t comuns, que geralmente são usadas para comparar duas amostras independentes quanto a um variável quantitativa.

2.4.2 Conjunto de Genes Biomarcadores

Os biomarcadores de genes geralmente se correlacionam muito bem com fenótipo de interesse e são fáceis de interpretar, no entanto o ruído nos dados podem levar a falsos positivos e falsos negativos (ZHAO; QIN, 2013). Buscando contornar os problemas com ruídos, a análise que investiga um conjunto de genes está se tornando uma tendência na interpretação de dados de expressão gênica, onde os genes no mesmo grupo são propensos a estar associados aos mesmos processos biológicos (ZAHA; FERREIRA; PASSAGLIA, 2014).

Considerada como pioneira para a análise de conjuntos de genes, a abordagem *Gene Set Enrichment Analysis* (GSEA) avalia o enriquecimento de conjuntos de genes predefinidos, e verifica se os mesmos possuem diferenças estatisticamente significativas entre duas condições interesse (SUBRAMANIAN et al., 2005).

No GSEA, o significado do score é avaliado com um teste de permutação empírica que corrige o teste de múltiplas hipóteses, em comparação com os biomarcadores de genes únicos, os conjuntos de genes identificados pela GSEA estão associados a vias ou processos biológicos são mais razoáveis para a interpretação dos dados. Além disso, em vez de se concentrar em genes com diferença significativa na expressão entre grupos ou condições de interesse, a GSEA pode detectar esses genes importantes com mudanças de expressão modestas (SUBRAMANIAN et al., 2005).

Foram propostas muitas variantes da GSEA, incluindo estatísticas de enriquecimento não paramétrico (BARRY; NOBEL; WRIGHT, 2005), (TIAN et al., 2005) e (HÄNZELMANN; CASTELO; GUINNEY, 2013), teste de bateria (DØRUM et al., 2009) e (IRIZARRY et al., 2009) e testes focados de conjuntos de genes (JIANG; GENTLEMAN, 2006) e (WU et al., 2010).

2.4.3 Vias Biológicas como Biomarcador

Uma via molecular representa as interações entre um conjunto de genes funcionalmente relacionados e são mais interessantes para os biólogos do que os conjuntos de genes (ZAHA; FERREIRA; PASSAGLIA, 2014). Considerado que, em vez das mutações

de genes individuais, a disfunção de caminhos moleculares leva ao início e ao desenvolvimento de doenças, especialmente doenças complexas, portanto, é mais razoável identificar essas vias disfuncionais subjacentes a doenças, isto é, vias biológicas como biomarcador, pode melhorar a robustez e a precisão do diagnóstico em comparação com os biomarcadores gênicos e os biomarcadores de conjuntos de genes (ANDRE et al., 2014).

Inspirado pela disponibilidades de informações, algumas abordagens computacionais foram desenvolvidas para identificar vias disfuncionais associadas a doenças. TARCA et al. (2008) propuseram uma abordagem denominada de *Signaling Pathway Impact Analysis* (SPIA) para medir o impacto de perturbações em uma determinada via sob uma determinada condição. Quando aplicado a conjuntos de dados de câncer, o SPIA supera a GSEA e identifica com sucesso caminhos conhecidos por estar envolvidos em cânceres. Mais tarde, VASKE et al. (2013) desenvolveram um modelo probabilístico baseado em gráfico conhecido como PARADIGM para identificar atividades de via específica de paciente com glioblastoma multiforme. O PARADIGM é capaz de integrar diferentes tipos de dados de informação e identificar as vias cujas atividades mudam significativamente nos pacientes, e detecta menos falsos positivos em comparação com o SPIA.

2.4.4 Redes de Biomarcadores

As redes moleculares podem dar uma visão mais global sobre os sistemas biológicos. Assim preservando as estruturas de via dentro da rede, removendo assim as limitações do conhecimento restrito a algumas vias, além disso, juntamente com os dados de alto rendimento acerca da expressão do gene, que pode descrever as atividades de moléculas individuais, as redes moleculares são capazes de caracterizar a dinâmica dos sistemas biológicos (ZHAO; QIN, 2013). Além disso, muitas doenças, especialmente doenças complexas, são causadas devido à disfunção de múltiplos genes, onde esses genes tendem a interagir uns com os outros em comparação com genes não relacionados a doença (CHEN et al., 2013).

De acordo com ZHAO; QIN (2013), muitas abordagens computacionais foram desenvolvidas para identificar sub-redes ou módulos a partir das redes moleculares, e essas sub-redes ou módulos possuem capacidade discriminativa de separar diferentes condições e, portanto, podem servir como biomarcadores. A maioria das abordagens identifica biomarcadores de rede com base na análise de redes diferenciais que integram as diferenças de genes únicos entre condições distintas e topologia de rede. Com base nas redes que eles usaram, essas abordagens podem ser categorizadas em métodos baseados em redes de associação de genes e abordagens baseadas em redes de interação proteína-proteína (TAMMINGA et al., 2017).

Nas redes de associação de genes, os nós são genes e uma aresta é colocada entre um par de genes se sua correlação de co-expressão, tipicamente o coeficiente de correlação de Pearson, estiver acima de um limiar. Ao construir diferentes redes de associação para

condições distintas com base em dados de expressão gênica, os padrões de co-expressão associados a doenças podem ser extraídos, que de outra forma são ignorados pela detecção de genes diferencialmente expressos (ZHAO; QIN, 2013).

Em CHU et al. (2011), foi descrita uma rede de associação com Modelos Gaussianos Gráficos, e detectaram as arestas que podem voltar a separar dois estados de doença comparando as probabilidades posteriores das conexões em duas condições de doença. Quando aplicados aos conjuntos de dados de câncer de mama, eles identificaram com sucesso os biomarcadores consistindo em conjuntos de genes ou caminhos, que são capazes de separar diferentes tipos histológicos de câncer de mama.

Posteriormente foi desenvolvida uma nova abordagem da Análise de Rede Diferencial (DINA) por GAMBARDILLA et al. (2013), para identificar caminhos ativos específicos da condição com a suposição de que os genes pertencentes aos mesmos caminhos tendem a ser co-regulados. A DINA tem sido utilizada com sucesso para detectar caminhos específicos do tecido e identificar via metabólica e transcricional específica do hepatocarcinoma desregulado.

Abordagem proposta por TAYLOR et al. (2009) para detectar módulos de rede que re-ligue em diferentes condições, examinando a estrutura dinâmica do interactoma humano com base em dados de expressão gênica. Quando aplicados a uma coorte de pacientes com câncer de mama, eles encontraram alguns genes que não apresentam mudanças significativas em sua expressão, mas esses genes têm diferentes parceiros de interação em pacientes sobreviventes e com desfecho ruins. Segundo os autores, esses genes podem servir como uma assinatura prognóstica para prever o desfecho e sobrevida de pacientes com câncer de mama.

2.5 Redes Complexas

As redes complexas são caracterizadas por representar um conjunto de elementos em que ligações dependem da característica do estudo que se deseja construir (BARABÁSI; GULBAHCE; LOSCALZO, 2011).

Redes complexas são modelos utilizados para representar elementos principais (também conhecidos como atores e/ou nós), e como eles se relacionam (arestas). Esses modelos a que no referimos podem ser representados matematicamente com grafos. Um grafo é um conjunto de nós conectados por arestas, e no caso das redes complexas os nós da rede representam os atores e as arestas o relacionamento entre os atores (NEWMAN; BARABASI; WATTS, 2011).

As redes possuem propriedades próprias e são suportadas pela teoria de grafos apresentando-se de três formas:

- Direcionadas: que são ligações estabelecidas em uma única direção, também chamadas de cíclicas, e são representadas por um grafo orientado;

que todos os vértices de uma determinada rede têm aproximadamente a mesma quantidade de conexões e as mesmas chances de receberem novas ligações (BARABÁSI; ALBERT, 1999). Segundo os autores, quanto mais complexa for a rede, maiores serão as chances dela ser aleatória.

2.5.2 Small World Network

De acordo WATTS; STROGATZ (1998), muitas redes apresentam padrões altamente conectados, tendendo a formar pequenas quantidades de conexões em cada vértice. Assim, eles propuseram um modelo semelhante ao de Erdős e Rényi, no qual grande parte das conexões são estabelecidas entre vértices mais próximos, apresentando-se como um *small world*.

Nesse modelo, a distância média entre quaisquer dois vértices de uma rede muito grande não ultrapassa um número pequeno de vértices. Para isso, basta que algumas conexões aleatórias entre grupos sejam estabelecidas (BUCHANAN, 2003). Na Figura 8 é apresentado um exemplo de *small world network*.

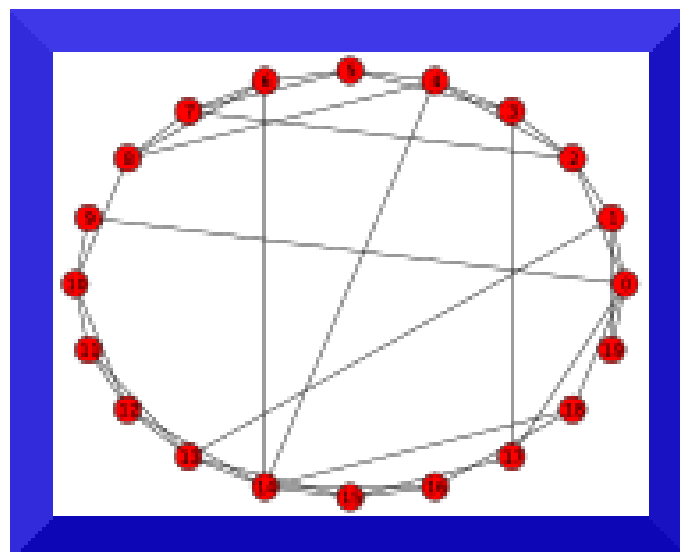


Figura 8: Exemplo de uma Small World Network. Fonte: Adaptado de (WATTS; STROGATZ, 1998).

O efeito *small world* é observado nas redes em que a maioria dos vértices se conecta a outros através de um caminho mínimo. O caminho mínimo, também chamado de caminho geodésico ou distância geodésica, é aquele formado pelo menor número de arestas que conectam um vértice origem e um vértice destino.

Para melhor ilustrar esse efeito, considere os indivíduos de uma sociedade qualquer. De acordo com o experimento conduzido por Stanley Milgram em 1960, se uma carta fosse entregue a um indivíduo, que não fosse o destinatário, e ele a repassasse a um outro e, assim, por diante, em aproximadamente seis passagens ela chegaria ao destinatário.

Esse resultado é uma demonstração direta do efeito *small world*, em que o caminho percorrido pela carta, partindo de um indivíduo qualquer até o destinatário, é mínimo.

2.5.3 Scale-free Network

Os modelos apresentados anteriormente relacionam propriedades relativas à distribuição regular de conectividade, seja com ligações determinadas ou aleatórias.

O modelo *Scale-free* as redes apresentam uma ordem na dinâmica de estruturação, com características bem específicas. Uma das principais características, denominada conexão preferencial, é a tendência de um novo vértice se conectar a um vértice da rede que tem um grau elevado de conexões.

Essa característica implica em redes com poucos vértices altamente conectados, denominados *hubs*, e muitos vértices com poucas conexões. As redes com essas características são denominadas livres de escala devido à representação matemática da rede. Ela segue uma função $f(x)$ que permanece inalterada com um fator multiplicativo sob um re-escalonamento da variável independente x .

Em outras palavras, isso significa que as *scale-free network* são aquelas em que a distribuição de graus segue a Lei de Potência, desde que exista uma solução somente para $f(ax) = bf(x)$. Conforme apresentado em (NEWMAN; BARABASI; WATTS, 2011), essas redes têm sido observadas em vários sistemas, por exemplo, na internet, na Web, em redes de metabolismos e em redes de citações de artigos científicos.

Na Figura 9 é apresentado um exemplo de rede livre de escala.

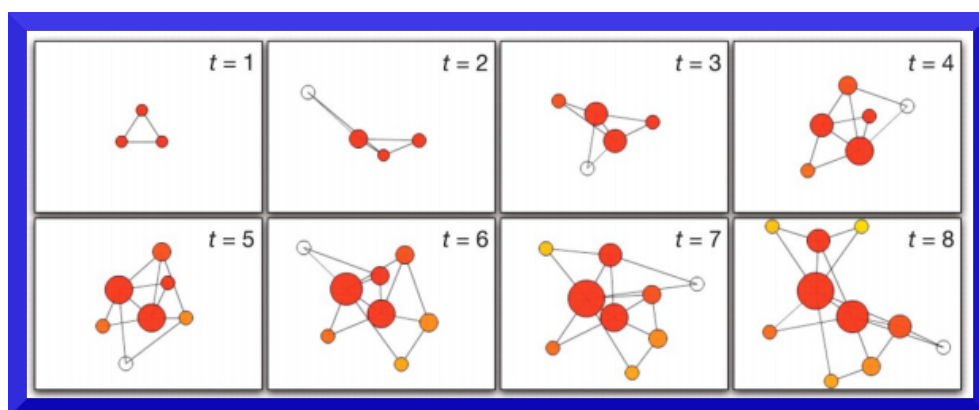


Figura 9: Construção do modelo Barabási-Albert. Os novos vértices preferem se ligar aos nós mais conectados. Devido às propriedades de crescimento e ligação preferencial podemos perceber o surgimento de *hubs* Fonte: Adaptado de (BARABÁSI, 2009)

As redes complexas têm sido aplicadas nas mais diversas áreas, para a resolução dos mais variados tipos de problemas, como nas ciências sociais, redes tecnológicas e biológicas.

Redes sociais

A sociologia é considerada entre as primeiras áreas a aplicar conceitos de redes depois da matemática. A ideia era utilizar grafos para representar as relações entre pessoas ou grupo de pessoas, Moreno (MORENO; JENNINGS et al., 1934), que posteriormente originou a área que atualmente é conhecida como análise de redes sociais. Entretanto, nos trabalhos iniciais, as redes eram chamadas de sociogramas, ao invés de redes sociais. Outro ponto interessante é que as ligações neste tipo de rede podem ter diferentes definições dependendo do que se está estudando.

As ligações, por exemplo, poderiam representar laços de amizade entre as pessoas da rede, ou relações profissionais, ou ainda, comunicação entre as pessoas ou o fluxo monetário entre elas. Um dos resultados interessantes no início dos estudos do trabalho de Moreno (MORENO; JENNINGS et al., 1934) foi quando, analisando um grupo de crianças de uma escola e suas amizades, ele descobriu que existiam poucos laços de amizade entre meninos e meninas, o que hoje é conhecido como homofilia ou mistura seletiva que é a tendência de indivíduos com alguma característica em comum terem um número maior de conexões entre si (MOODY, 2001).

Redes Tecnológicas

A Internet a que nos referimos aqui é composta pelos servidores ou computadores que são os nós da rede complexa e as ligações feitas por cabos entre eles que são as ligações. Existem várias maneiras diferentes de se obter a topologia da Internet, e várias camadas diferentes em que ela pode ser analisada. Um resultado interessante foi o de (FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999), que analisou a distribuição de graus da Internet e descobriu que ela segue aproximadamente uma lei de potência, que como é uma característica frequentemente presente nas redes reais.

Redes Biológicas

Na área das ciências biológicas as redes complexas são empregadas, devido à importância das relações entre os diversos elementos estudados, que na maioria dos casos também está associada ao funcionamento do sistema biológico. Podemos citar por exemplo, as conexões entre as células cerebrais na área da neurociência (SPORNS, 2010), ou ainda as relações entre diferentes espécies em ecossistemas (PROULX; PROMISLOW; PHILLIPS, 2005), ou ainda a interação entre proteínas (ZHANG, 2009) ou redes de expressão gênica (WANG; LIU, 2017), dentre outros.

2.5.4 Medidas de Centralidade

Existem algumas redes que usualmente compartilham características e propriedades, como os diferentes tipos de redes biológicas (PAVLOPOULOS et al., 2011). Torna-se necessário métricas para interpretar informações relevantes que essas redes venham a

fornecer.

Alguns nós de uma rede podem apresentar propriedades topológicas que os destacam em relação aos demais. Por exemplo, é possível que um nó seja tão importante para a rede a ponto de desconectá-la, caso seja excluído. Os nós capazes de afetar a topologia da rede dessa forma (ou de maneira semelhante) são chamados de nós centrais. Em redes biológicas é importante identificar nós (centrais ou intermediários) que possam afetar a topologia da rede (PAVLOPOULOS et al., 2011). Medidas de centralidade quantificam a importância de cada gene no funcionamento do sistema.

A seguir, são apresentadas as definições das medidas empregadas neste trabalho.

Degree provavelmente seja uma das medidas mais simples de centralidade, porque ela avalia a importância de um nó analisando a quantidade de nós a que ele se conecta, ou seja, quanto maior o número dos nós conectados a este, maior a importância do nó para a rede e, portanto, maior o valor atribuído para este vértice pela centralidade (PAVLOPOULOS et al., 2011). O *degree* pode ser calculado utilizando a seguinte expressão:

$$K_i = \sum_{j=1}^n A_{ij} \quad (2)$$

Em que k_i é o grau do nó i , A_{ij} são os elementos da matriz de adjacência da rede complexa e n é o número de vértice na rede complexa.

A matriz de adjacência é uma das maneiras de se representar um grafo. Nessa matriz, as linhas e colunas estão associadas aos nós da rede e quando existe uma ligação entre dois nós da rede a coluna e a linha correspondente é acrescentado o valor 1 para representar essa relação e 0 quando não há relação. Um exemplo da matriz pode ser visualizando na Tabela 1.

Tabela 1: Matriz de Adjacência Fonte: Dados do Autor

	A	B	C	D	E
A	0	1	1	1	1
B	1	0	0	0	0
C	1	0	0	1	0
D	1	0	1	0	1
E	1	0	0	1	0

Na Figura10 pode ser visualizado o grafo que representa a matriz de adjacência da tabela.

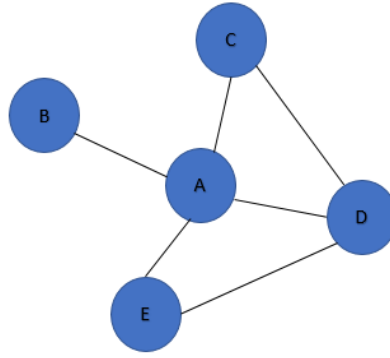


Figura 10: Grafo com 5 nós e 6 ligações sendo A nó com maior *degree* e B com menor *degree* Fonte: Dados do Autor.

O **Betweenness** mede a extensão em que um vértice se encontra em trajetos entre outros vértices. Vértices com alta interdependência podem ter considerável influência dentro de uma rede em virtude de seu controle sobre a informação que passa entre os outros. Eles também são aqueles cuja remoção da rede irá interromper a comunicação entre outros vértices, porque eles se encontram no maior número de caminhos tomados pelas mensagens (PAVLOPOULOS et al., 2011).

Matematicamente, seja $n_{s,t}^i$ o número de caminhos geodésicos de s a t que passam por i e seja $n_{s,t}$ o número total de caminhos geodésicos de s a t , ainda assim é possível descobrir o poder de intermediação de nó i .

$$b_i = \sum_{s,t} w_{s,t}^i = \sum_{s,t} \frac{n_{s,t}^i}{n_{s,t}} \quad (3)$$

Sendo assim, por convenção, a razão $w_{s,t}^i = 0$ se $n_{s,t} = 0$. Observe que cada par de vértice s, t contribui para a soma de i com um peso $w_{s,t}^i$ entre 0 e 1 expressando a distância entre i em relação ao par s, t .

A definição dada conta separadamente os caminhos geodésicos em qualquer direção entre cada par de vértices. Como esses caminhos são os mesmos em um gráfico não direcionado, isso conta efetivamente cada caminho duas vezes, e inclui caminhos que começam ou terminam com i (s pode ser igual a i ou t podem ser iguais a i), bem como caminhos de um vértice para si mesmo (s pode ser igual a t).

Closeness mede a distância média de um vértice para outros vértices e um caminho geodésico é o mais curto através de uma rede entre dois vértices (PAVLOPOULOS et al., 2011). Suponha que $d_{i,j}$ seja o comprimento de um caminho geodésico de i a j , significando o número de arestas ao longo do caminho, então a distância geodésica média para o vértice i é:

$$l_i = \frac{1 - n}{\sum_j d_{i,j}} \quad (4)$$

Essa quantidade recebe valores baixos para vértices que são separados dos outros por apenas uma pequena distância geodésica em média. Tais vértices podem ter melhor acesso à informação em outros vértices ou mais influência direta em outros vértices. Em uma rede social, por exemplo, uma pessoa com menor distância média para os outros pode achar que suas opiniões alcançam os outros na comunidade mais rapidamente do que a opinião de alguém com maior distância média. Alguns autores excluem da soma o termo $j = i$ para o qual $d_{i,i} = 0$ e, portanto, dividem a soma por $n - 1$ em vez de n , t .

A distância média l_i não é uma medida de centralidade no sentido das anteriores, pois dá valores baixos a nós mais centrais e valores altos a valores menos centrais.

O **PageRank** é um algoritmo de análise de links usado pelo mecanismo de busca do Google, que atribui um peso numérico a cada elemento de um conjunto de links com o objetivo de medir a sua importância relativa dentro do conjunto.

A definição do valor PageRank para o elemento u pode ser expressa como:

$$PR(u) = \frac{1 - d}{N} + d \sum_{v \in B_u} \frac{PR(v)}{L(v)} \quad (5)$$

O valor *PageRank* para um elemento u é dependente dos valores *PageRank* para cada elemento v dentro do conjunto B_u (este conjunto contém todos os elementos que se ligam ao elemento u) dividido pelo número $L(v)$ de links do elemento v . A variável d é um fator de amortecimento que geralmente é definida com um valor em torno de 0,85 é dividida por N que representa o número de páginas na rede.

Alpha Centrality (Bonacich Power) calcula a centralidade de cada entidade baseado na centralidade de seus vizinhos. Ele parte da premissa de que o poder surge mais da ligação com os nós fracos do que com os fortes: um nó conectado a outros igualmente bem conectados é poderoso, mas menos do que outro nó com muitos elos com outros que tenham poucas conexões, pois estes se tornam dependentes do primeiro (BONACICH, 1987).

Alpha Centrality é definido por:

$$x = \text{alphat}(A)x + e \quad (6)$$

onde A é a matriz de adjacência (não necessariamente simétrica) do gráfico, e é o vetor de fontes exógenas de status dos vértices e alpha é a importância relativa dos fatores endógenos versus exógenos, T é o autovetor e x é a quantidade de influência do nó.

Strength mede a força dos vértices em termos do peso total de suas conexões. Se imaginarmos no caso da uma rede *wide area network* (WAN), a força do vértice simplesmente representa o total tratado por um provedor internet. Portanto, é intuitivo considerar a definição alternativa de centralidade construída observando a força dos vértices como uma definição mais apropriada da importância de um vértice em redes ponderadas (BAR-RAT; BARTHELEMY; VESPIGNANI, 2007).

2.6 Aprendizado de Máquina

A clássica definição de Tom M. Mitchell (1997), diz que o programa de computador aprende com a experiência E com respeito para alguma classe de tarefas T e medida de desempenho P , se o seu desempenho em tarefas em T , medido por P , melhora com a experiência E .

A ideia por trás da utilização do conceito de aprendizagem não é apenas para saber como atuar, mas também tem como intuito aumentar a habilidade do agente para atuar no futuro. A aprendizagem ocorre através de como o agente observa as suas interações com o mundo e os seus próprios processos de decisão (SILVA; ZHAO, 2016).

De acordo com SILVA; ZHAO (2016), os algoritmos de aprendizado de máquina atualmente são agrupados segundo três paradigmas: não supervisionados, supervisionados e semi-supervisionados:

- Os algoritmos não supervisionados dispensam qualquer tipo de conhecimento externo, sendo guiados exclusivamente pela estrutura intrínseca dos itens de dados ao longo do processo de aprendizagem. Dentre as principais funções de sua aplicação pode-se destacar: agrupamento, detecção de *outlier*, redução de dimensionalidade, e associação;
- Os algoritmos supervisionados são aqueles que utilizam exclusivamente informações externas para induzir ou treinar suas hipóteses. São frequentemente empregados em problemas de classificação, quando os dados rotulados são discretos, e em problemas de regressão, quando tratam valores contínuos.
- Os semi-supervisionados, que mesclam características dos algoritmos supervisionados e dos não supervisionados, ou seja, empregam dados rotulados e não-rotulados no processo de aprendizagem.

Dentro do objetivo dessa dissertação com a descoberta de novos possíveis biomarcadores focou-se em métodos de aprendizagem supervisionado afim de de classificar genes quanto ao potencial de representar um biomarcador através do treinamento de modelos baseados em dados da literatura.

2.6.1 Aprendizado Supervisionado

Algoritmos de aprendizagem supervisionada supõem a existência de um "Professor" que ensina que tipo de comportamento você deve exibir em cada situação. Na prática, deseja-se classificar pacientes saudáveis e não-saudáveis e para fazer isso tem-se uma amostra que associa a cada paciente saudável a uma série de variáveis. Então, um algoritmo de aprendizagem supervisionado tentaria usar explicitamente essa informação para no futuro ser hábil para separar pacientes saudáveis de não-saudáveis.

No contexto dessa dissertação de mestrado utilizou-se três algoritmos de aprendizado supervisionado *Support Vector Machine (SVM)*, *Random forest* e *J48*.

Uma *Support Vector Machine* é uma representação matemática de um modelo de classificação tipicamente binária. Os casos são representados num espaço e o classificador calcula então uma fronteira (híperplano) que separa ambas as classes de forma equitativa. Para classificar novos casos, basta verificar em que lado da fronteira este é representado. Apesar de inicialmente só ser possível uma classificação através de uma função linear, ou seja, a fronteira entre classes ser uma linha reta, desenvolvimentos mais recentes tornam possível a classificação não-linear através de funções polinomiais e de base radial.

O *Random Forest* é definido formalmente de acordo com BREIMAN,(2001) como um classificador composto por uma coleção de árvores $h_k(x)$ $k = 1, 2, \dots, L$ onde T_k são amostras aleatórias independentes e identicamente distribuídas e cada árvore vota na classe mais popular para entra x .

A árvore de decisão (J48) é uma representação matemática de um modelo com regras estruturadas numa hierarquia de classes ou valores (SANTOS and AZEVEDO, 2005) aplicáveis a tarefas de classificação ou regressão, consoante se trate de classes discretas ou contínuas, respetivamente. A representação gráfica assemelha-se a uma árvore natural sendo a representação mais comum invertida pois o topo simboliza a raiz e na base encontramos as folhas.

Dentre os problemas constantemente encontrados quando trabalha-se com algoritmos de Aprendizagem de Máquina, o desequilíbrio entre classes e a seleção dos preditores adequados para criação dos modelos são mais enfrentados comumente SILVA; ZHAO (2016). Sendo o primeiro geralmente a classe que se deseja inferir é a que possui a menor quantidade de exemplos para ensinar o algoritmo os seus padrões para ele possar reconhecer corretamente e o segundo busca selecionar os atributos que busquem melhorar a classificação dos modelos.

Para enfrentar o desequilíbrio entre as classes, destacam-se as seguintes técnicas de balanceamento de classes, como *Down Sample* e *Upper Sample* descritos no clássico artigo de (PROVOST, 2000), e também o *SMOTE*, que cria instâncias artificiais para balancear as classes (CHAWLA et al., 2002) SILVA; ZHAO (2016).

Técnica de balanceamento de classe *Down Sample*, seleciona aleatoriamente subconjunto todas as classes no conjunto de treinamento para que suas frequências de coincidam com a classe minoritária. Um exemplo de sua aplicação, suponha que 80% equivalente a 800 instâncias das amostras do conjunto de treinamento seja a primeira classe e que os 20% equivalente a 200 instâncias restantes estejam na segunda classe. O *Down Sample* ira selecionar aleatoriamente 200 instancias da classe majoritária para balancear as classes.

Já O *Up Sample* realiza uma amostragem aleatoriamente com substituição, a classe minoritária até ficar equivalente a classe majoritária. Como exemplo 800 instâncias na classe majoritária e 200 na minoritária, essa técnica realiza amostragem aleatória bus-

cando 600 amostras para tornar as duas classes balanceadas.

A técnica balanceamento de classes *SMOTES*, estima novas instâncias minoria sintética em algum lugar das linhas traçadas entre instâncias reais. Na Figura 11 pode-se visualizar como é realizada a sintetização de novas instâncias minoritárias, o desequilíbrio diminui de 4 vermelho contra 13 verde para 12 vermelho contra 13 verde dessa maneira é possível adquirir um classificador que identifique as duas classes com bons índices de acerto.

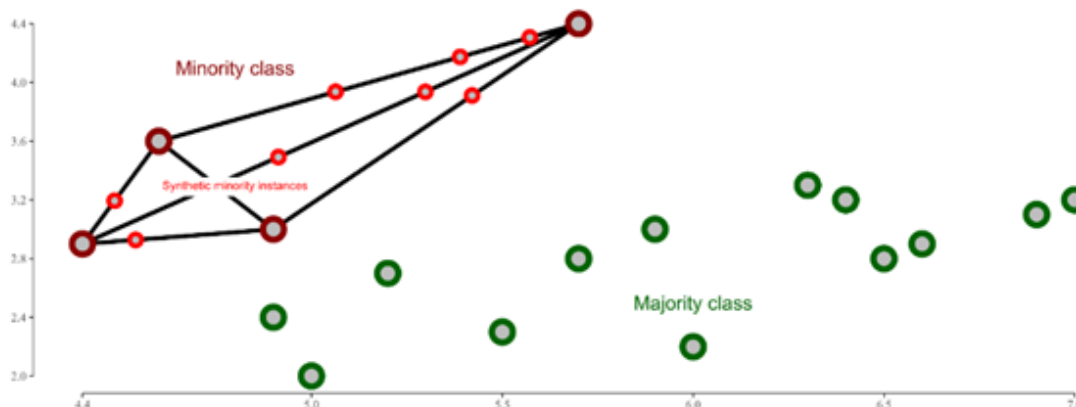


Figura 11: Exemplo da técnica de balanceamento *SMOTE*. Fonte: Adaptado de (CHAWLA et al., 2002).

De acordo com trabalho de LOPES et al. (2017), dentre as técnicas que buscam amenizar os problemas de desbalanceamento bem como sub-amostragem da classe majoritária, excesso de amostragem de classes minoritárias, penalização da classe majoritária, tem uma probabilidade maior da ocorrência de *overfitting*.

Segundo artigo original da técnica CHAWLA et al. (2002), o SMOTE, perde o seu poder balanceamento de classe quando ultrapassamos a proporção de 500% por 200% entre instância podendo gerar ruídos sendo essa hipótese confirmada por (RUFINO; VEIGA; NAKAMOTO, 2016).

O segundo problema de seleção de atributos, existem diferentes técnicas para auxiliar nesse desafio dentre essas destacam-se, algoritmos *Correlation-based Feature Selection* (CFS) e *Consistency Subset Eval* SILVA; ZHAO (2016).

O algoritmo CFS classifica os subconjuntos gerados de acordo com uma função de correlação com base em uma recompensa heurística de avaliação e também avalia a importância de um subconjunto de atributos em função da predição individual de cada atributo e o grau de correlação entre eles.

O CFS seleciona um subconjunto de recursos de forma explícita e avalia cada recurso individualmente e atribui uma pontuação com base em sua importância SILVA; ZHAO (2016).

Já o algoritmo CSE reduz o conjunto de atributos preservando a consistência original, avalia o valor de um subconjunto de atributos pelo nível de consistência dos valores verificados para cada classe quando as instâncias de treinamento são projetadas sobre o subconjunto de atributos SILVA; ZHAO (2016).

Realizado balanceamento de classes e a seleção dos atributos, para a avaliação de um modelo de classificação existem diferentes formas, dentre essas as mais utilizadas são a tabulação cruzada também conhecida como tabela de contingência, ou matriz de confusão e/ou *Receiver Operating Characteristic* (ROC), ou também conhecida como Curva ROC (PRATI; BATISTA; MONARD, 2008).

Na Figura 12 é mostrada uma matriz de confusão genérica e as medidas comuns que podem ser obtidas a partir dela. Quando um exemplo real positivo é classificado como positivo, ele é denominado verdadeiro positivo (True Positive). Quando um exemplo real negativo é classificado como positivo, ele é denominado falso positivo (False Negative). Nomenclatura similar é utilizada no caso dos exemplos classificados como negativos. Nessa tabela, **TP**, **FP**, **FN** e **TN** correspondem, respectivamente, às quantidades de verdadeiro/falso positivo/negativo. *YES* e *NO* correspondem ao número de exemplos preditos (*Classified class*) como positivos/negativos e *yes* e *no* ao número real (*Real class*) de exemplos positivos/negativos na amostra.

		Real class			
		yes	no		
Classified class	YES	TP True Positives	FP False Positives	$tp\ rate = \frac{TP}{P}$	$recall = \frac{TP}{P}$
	NO	FN False Negatives	TN True Negatives	$fp\ rate = \frac{FP}{N}$	$N = FP + TN$
				$precision = \frac{TP}{TP + FP}$	$P = TP + FN$
				$accuracy = \frac{TP + TN}{P + N}$	
				$F\text{-measure} = \frac{2}{1/precision + 1/recall}$	

Figura 12: Matriz de confusão e suas metricas. Fonte: Adaptado de (PRATI; BATISTA; MONARD, 2008).

Para esta dissertação foram avaliados a acurácia, a sensibilidade e a especificidade. A sensibilidade da classificação é a razão entre os verdadeiros positivos e o total de exemplos positivos.

$$sensibilidade = \frac{TP}{TP + FN} \quad (7)$$

A especificidade da classificação é a razão entre os verdadeiros negativos e o total de exemplos negativos.

$$especificidade = \frac{TN}{FP + TN} \quad (8)$$

Como segunda media de avaliação dos modelos de *Machine Learning*, utilizou-se a

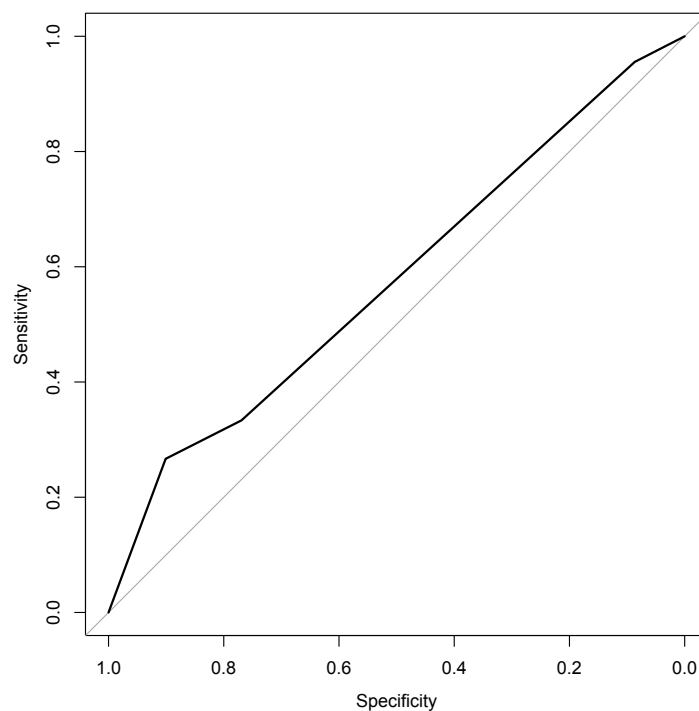


Figura 13: Exemplo de Curva ROC. Fonte: Dados do Autor.

análise de Curvas ROC, a qual representa o compromisso entre a sensibilidade (a taxa de verdadeiros positivos) no eixo y e falso alarme (a taxa de falsos positivos) no eixo x. Curvas ROC são obtidas pela variação do limiar de um modelo probabilístico de classificação (PRATI; BATISTA; MONARD, 2008). Na Figura 13 pode ser visualizada um exemplo de Curva ROC.

Outra forma de treinamento e avaliação do modelo é através da validação cruzada (*r-fold cross validation*). Este método de amostragem utilizado para análise de desempenho, consiste em dividir aleatoriamente os exemplos em r partições mutuamente exclusivas (*folds*) de tamanho aproximadamente igual a n/r exemplos, sendo n o número total de exemplos. Os exemplos nos $r-1$ *folds* são então usados para treinamento e a hipótese induzida é testada no *fold* restante. Tal processo é repetido r vezes, de modo que todos os *folds* sejam usados uma vez como conjunto de teste, pode ser visualizada na Figura 14.

Fold 1	Fold 2	Fold 3	Fold 4
TESTE	TREINAMENTO	TREINAMENTO	TREINAMENTO
TREINAMENTO	TESTE	TREINAMENTO	TREINAMENTO
TREINAMENTO	TREINAMENTO	TESTE	TREINAMENTO
TREINAMENTO	TREINAMENTO	TREINAMENTO	TESTE

Figura 14: Validação cruzada utilizando 4 *folds*. Fonte: Dados do Autor

3 TRABALHOS RELACIONADOS

3.1 Redes de Biomarcadores em diabetes

Em sua pesquisa WANG; LIU (2017) estabeleceram um *framework* utilizando estrutura de rede de co-expressão de genes para identificar biomarcadores de transcritos com diferenças nos padrões de co-expressão de genes em amostras de doença e controle.

Foi utilizada a correlação de Pearson para caracterizar as relações entre os genes e por meio de diferentes coeficientes de correlação de Pearson, os autores categorizaram os pares de genes em quatro categorias: categoria 1 refere-se à co-expressão na doença e não-expressão nos controles; categoria 2, refere-se à não-expressão na doença e co-expressão nos controles; categoria 3 refere-se à co-expressão na doença e nos controles; e categoria 4 sem co-expressão entre doença e controle.

Avaliou-se a capacidade de classificação do SVM nas diferentes categorias e também realizaram a análise de enriquecimento funcional desses genes diferenciais co-expressos. Os grupos genéticos com alto poder de diferenciação com óbvias disfunções diabéticas são selecionados como os biomarcadores identificados de diabetes.

Os autores utilizaram a base de dados disponível no NCBI GEO com o ID (GSE12368). Esses experimentos são implementados em ratos de laboratório da espécie *Rattus norvegicus*, divididos em dois grupos com *Diabetes mellitus* (GK) e ratos controle (WKY).

As expressões dos genes são medidas em cinco intervalos de tempo diferentes durante os cursos de desenvolvimento dos ratos com cinco repetições individualmente, isto é, 4, 8, 12, 16 e 20 semanas em GK e WKY e respectivamente obtiveram 25,345 genes em 50 amostras de perfis de expressão.

A Figura 15 demonstra a visão geral do *framework* para identificação de biomarcadores para diabetes a partir dados expressão de genes. Neste cenário, os autores realizaram um teste para triar os genes candidatos para calcular a correlação de Pearson a fim de construir uma rede de co-expressão empregando o SVM para avaliar a classificação desses genes selecionados pelo seu poder distintivo de classificar estados diferentes. A análise de enriquecimento GO e a curadoria da literatura fornecem mais evidências de

suas implicações disfuncionais durante a progressão diabética.

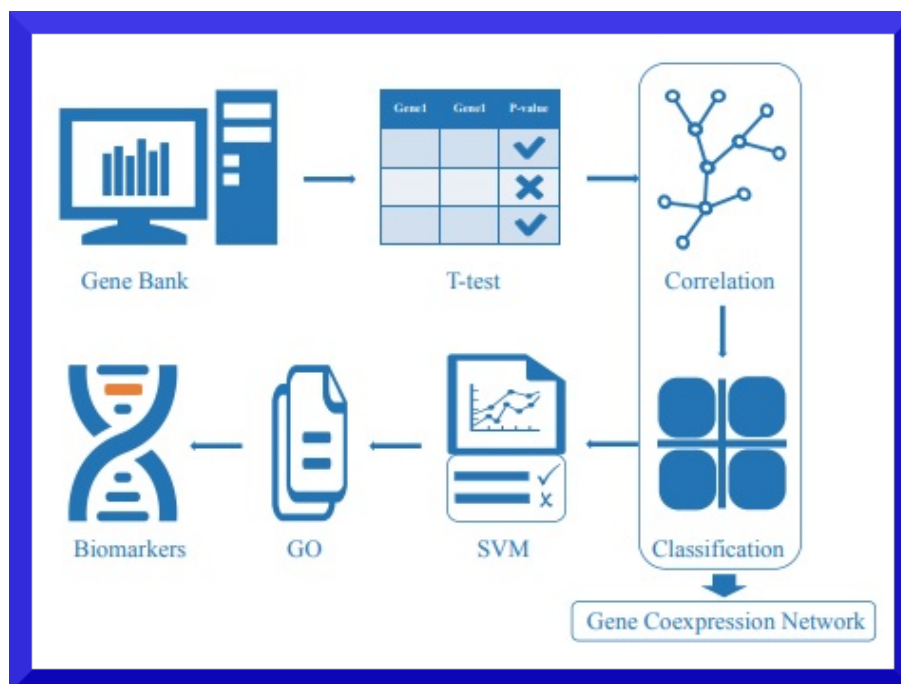


Figura 15: Visão geral do *framework*. Fonte: Adaptado de (WANG; LIU, 2017).

Como resultado os autores encontram genes como PIK3R1 e EDNRB por serem altamente relacionados à patogênese diabética. Os genes identificados podem servir como biomarcadores de diabetes e podem ser usados para o diagnóstico precoce de diabéticos após ensaios clínicos densos.

3.2 Rede de Biomarcadores biomarcadores para leucemia linfocítica crônica

Neste trabalho ZHANG et al. (2010), adotaram uma abordagem estudando genes co-expressos com ZAP70, um biomarcador bem caracterizado para LLC e múltiplos conjuntos de dados que contivessem esse gene. De forma mais específica os autores, selecionaram 23 conjuntos de dados de microarranjo correspondentes a vários tipos de cânceres do GEO.

Os autores empregaram o algoritmo CODENSE para identificar a rede de co-expressão gênica altamente conectada que abrange todo o genoma. Em seguida, reduziram a lista de genes nas redes de co-expressão nas quais a ZAP70, testando suas capacidades para prever o status de IgV H dos indivíduos usando vários métodos de seleção de recursos de aprendizado de máquina.

Está resumido o fluxo de trabalho da abordagem proposta pelos autores na Figura 16. Os biomarcadores identificados a partir do fluxo de trabalho foram então submetidos à validação testando seu poder de prognóstico em outro conjunto de dados de microarranjo

de LLC, com informações disponíveis sobre resultados clínicos de pacientes.

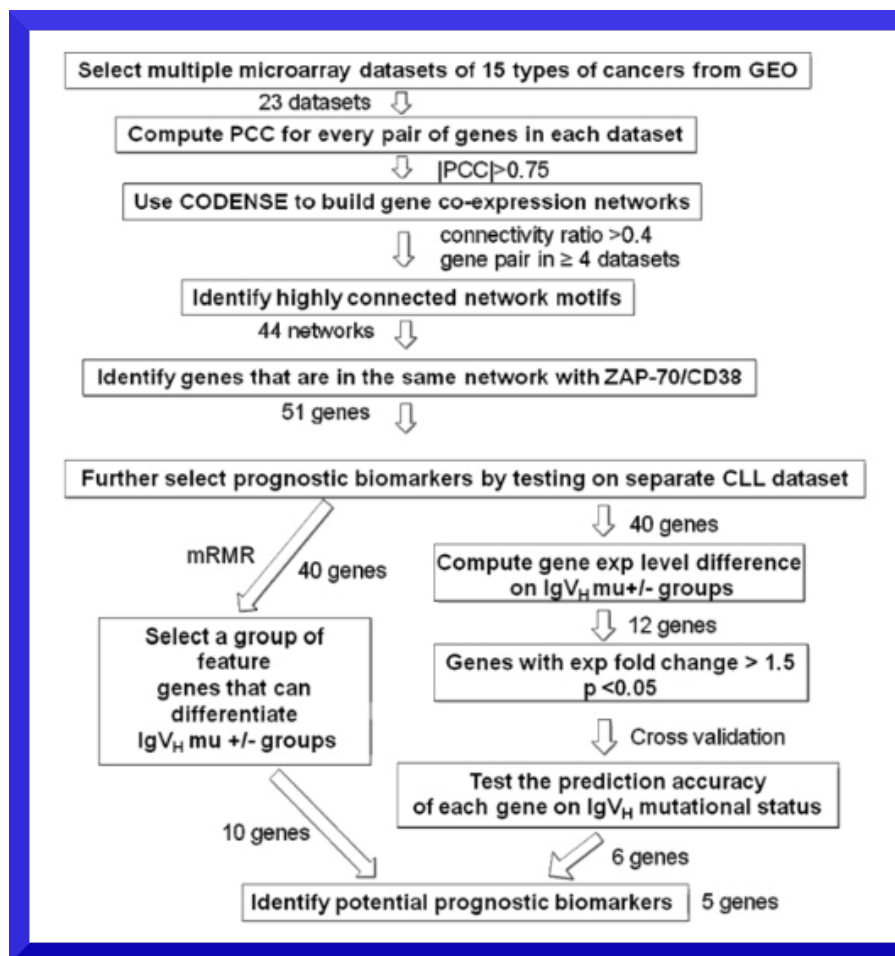


Figura 16: Fluxo de trabalho da abordagem proposta. Fonte: Adaptado de (ZHANG et al., 2010).

Os autores selecionaram 23 *dataset* de diversos tipos d canceres no GEO realizaram a correlação de Pearson para filtrar desses conjuntos de *datasets* somente os genes com limiar de correlação maior 0,75 e utilizaram o algoritmo CODENSE e identificaram 44 redes de co-expressão altamente conectadas (taxa de conectividade $r > 0,4$).

Dessas 44 redes os autores identificaram uma rede com 51 genes, dentre esse apareceu o conhecido biomarcador CLL ZAP70. Curiosamente, outro biomarcador CLL CD38 também estava incluído nesta rede e quatro genes CD8A, CD3G, CD247 e CD3D, cujos produtos são conhecidos por interagirem com o ZAP70, também estão nessa rede.

Os autores realizaram o enriquecimento na *Gene Ontology*, a qual revelou funções biológicas altamente enriquecidas relacionadas à leucemia, tais como crescimento e proliferação celular, desenvolvimento e função do sistema hematológico, resposta inflamatória e doença imunológica, filtrando-se para 40 genes para próxima etapa. Nesta pesquisa, além dos autores terem o foco em um gene já conhecido como biomarcador, nas próximas etapas eles verificaram os genes diferencialmente expressos com relação ZAP70 e genes cujos níveis de expressão podem prever o status de mutação do IgV_H.

Ao final da análise de co-expressão gênica, foi identificado o conjunto de genes IL2RB, CD8A, CD247, LAG3 e KLRK1, potenciais biomarcadores prognósticos da LLC. Os mesmos foram avaliados utilizando uma validação cruzada com 20% de retenção das amostras, repetida 100 vezes, através da aplicação destes biomarcadores para classificar os grupos de sobrevivência de pacientes utilizando conjuntos de dados de microarranjos de LLC com informações clínicas de resultados dos pacientes.

3.3 Rede de Biomarcadores de melanoma metastático

Nesta pesquisa (WANG; LI; CHEN, 2018), para explorar potenciais alvos regulatórios do melanoma metastático, analisaram os perfis de expressão de microarranjo de genes de amostras de melanoma maligno utilizando *weighted correlation network analysis* (WGCNA). Os autores utilizaram a biblioteca EdgeR desenvolvida por (ROBINSON; OSHLACK, 2010) para calcular os genes diferencialmente expressos (DEGS) e a biblioteca WGCNA para executar a análise de rede de correlação ponderada, como construção de redes, detecção de módulos, seleção de genes, simulação de dados, visualização redes de co-expressão com software externo como *cytoscape* e o *vizant* (LANGFELDER; HORVATH, 2008).

Os dados utilizados foram de perfis de microarranjo dos Institutos Nacionais de Saúde (NIH), banco de dados *Genomic Data Commons* (GDC) com um total de 471 amostras, incluindo 103 tumores sólidos primários e 368 amostras de melanoma metastático. Para construção das redes de co-expressão os autores utilizaram os DEGS identificados. Inicialmente, calcularam a correlação de Pearson entre todos os genes presentes no conjunto de dados e obtiveram uma matriz de correlação; no segundo momento, foram removidos os genes levemente correlacionados, mantendo apenas as correlações mais forte.

Com a seleção dos genes que possuíram a correlação mais forte na matriz de correlação, foi convertida em uma matriz adjacência e posteriormente para uma matriz de sobreposição topológica que combina a adjacência de dois genes e as forças de conexão que esses dois genes compartilham com outros genes. A matriz de sobreposição topológica (TOM), é uma medida altamente robusta de interconectividade de rede (proximidade). Essa proximidade é usada como entrada de *clustering* hierárquico na qual o genes são agrupados em módulos de co-expressão gênica e são representados por cores.

Para determinar os mecanismos biológicos de alteração do melanoma foi usada a regulação positiva e negativa de DEGs e a Gene Ontology (GO) para a análise de enriquecimento da via do *Encyclopedia Gene and Genomes* (KEGG). Todos os DEGs anotados no banco de dados do GO foram usados para classificar as funções do GO através do pacote *clusterProfiler* e para entender a distribuição da funcionalidade do gene no nível global. DEGs para análise de enriquecimento KEGG foram mapeados para o banco de dados KEGG. Após a análise dos dados GO e KEGG, os critérios de seleção significativos

foram definidos como taxa de descoberta falsa (FDR) e valor de *Pvalue* inferior a 0,05.

Com a análise de enriquecimento dos genes, os autores encontraram vinte e quatro genes hub importantes (RASGRP2 , IKZF1 , CXCR5 , LTB , BLK , LINGO3 , CCR6 , P2RY10 , RHOH , JUP , KRT14 , PLA2G3 , SPRR1A , KRT78 , SFN , CLDN4 , IL1RN , PKP3 , CBLC , KRT16 , TMEM79 , KLK8 , LYPD3 e LYPD5) os quais foram tratados como fatores importantes envolvidos na resposta imune e no desenvolvimento de células tumorais.

3.4 Redes de Biomarcador para infarto agudo do miocárdio

O objetivo do estudo de ZHANG et al. (2017) foi buscar potenciais biomarcadores para detecção e tratamento do infarto agudo do miocárdio (IAM). Foram utilizados um conjunto de dados denominado de GSE48060, consistindo de 52 amostras de sangue, sendo 31 amostras de pacientes após IAM e 21 amostras controle, sendo analisadas a expressão gênica daquelas amostras coletadas dentro das primeiras 48 horas após o IAM.

Empregando a biblioteca limma da linguagem R desenvolvida SMYTH et al. (2005), foram identificados genes diferencialmente expressos (DEGs) entre os 2 tipos de amostras pelos autores. Para construir e analisar a rede, utilizou-se o WGCNA e os genes com o valor limiar de $P < 0,05$ a partir dos resultados da análise de expressão diferencial entre amostras IAM e amostras normais em GSE48060 foram selecionados.

Foram realizados as seguintes etapas para a construção da rede WGCNA, através da definição da matriz de co-expressão gênica, a qual foi calculado o coeficiente de correlação de Pearson para todas as interações dos genes.

O coeficiente de ponderação β foi determinado de acordo com o princípio *scale free*. Aqui, i representa o números de nós, e $p(i)$ representa a probabilidade de um nó. Para a construção da rede utilizando o WGCNA, foi requerido que a Correlação de Pearson entre $\log(i)$ e $\log[p(i)]$ fosse pelo menos 0,8. Após o estabelecimento da rede, o *clustering de linkage* médio hierárquico foi usado para dividir diferentes módulos de co-expressão gênica. Então as correlações entre genes em um módulo e uma doença foram reconhecidas de acordo com os seguintes métodos: O valor P de um DEG entre 2 tipos de amostras foi calculado; o $\log P$ foi definido como significância gênica (GS), e a média GS em um módulo foi considerada como significância de módulo (MS). Comumente, um módulo com maior MS indica uma possibilidade maior de estar associado à doença e genes com maior *degree* foram considerados como *hubs*.

Como resultado os autores encontram potenciais biomarcadores para detecção de IAM como GZMB, GBP5, FPR1, TLR2, STAT1 e SPZ1 e esses genes podem exercer suas funções através da regulação das respostas imunológicas e inflamatórias.

Considerações Finais

Dentre os trabalhos existentes no estado da arte, escolheu-se esses como relacionados por partilharem da utilização de dados de expressão de genica, construção de redes co-expressão e pela busca de possíveis biomarcadores.

No trabalho WANG; LIU (2017), os autores buscam identificar biomarcadores para diabetes propondo um *framework*, no qual os primeiros passos como seleção dos dados de expressão gênica e construção de redes é um ponto incomum entre todos os trabalhos relacionados. Esse trabalho se diferencia frente aos outros na questão da escolha dos genes a serem analisados, realizando um filtragem inicial buscando a maior significância de *pvalue* do teste *t Student* por empregar aprendizagem de máquina para classificar as diferenças entre a doença e controle.

Na pesquisa ZHANG et al. (2010), abordagem escolhida focou em um gene que é um biomarcador bem característico da LLC, assim avaliando outros genes que se relacionam com ZAP70 e o empregou o algoritmo CONDESE para identificar redes altamente conectadas. Diferente da outra abordagem esse trabalho utilizou dados de pacientes com e sem mutação IgV H, buscando prever o status dessa mutação e aplicou um método de seleção de atributos.

Neste trabalho (WANG; LI; CHEN, 2018), a abordagem empregada foi a análise de genes diferencialmente expressos e a de rede de co-expressão gênica ponderada, para a identificação de módulos de genes relacionados a doença com a construção da rede de interação proteína-proteína onde foram extraídos os genes centrais (*hubs*). Assim, esse trabalho se diferenciou dos demais pela construção de uma rede de interação proteica para analisar a função dos genes relacionados e a identificação de fatores de transcrição importantes.

Na pesquisa de ZHANG et al. (2017), utilizou-se a análise da rede de co-expressão gênica ponderada como os trabalhos de (ZHANG et al., 2010) e WANG; LI; CHEN (2018), os quais buscam identificar os módulos mais relacionados a doença e os genes diferencialmente expresso. Foi realizada uma normalização e sumarização das sondas de coletas dados de expressão gênica, as quais que não corresponderem aos símbolos validos foram eliminados. Adicionalmente foi realizado o enriquecimento funcional através da *gene ontology* para avaliar as vias metabólicas. Um ponto de distinção dessa pesquisa com as demais, é a utilização de uma medida de centralidade tradicional (*degree*), que considera o número de interações de um nó da rede com os demais. Dessa forma, nós da rede com um alto *degree* são considerados *hubs* na rede construída.

Nos quatro trabalhos podemos constatar diferentes metodologias para descoberta de biomarcadores, empregando dados de expressão gênica e redes de co-expressão. O emprego de aprendizagem de máquina para auxiliar no reconhecimento de padrões, devido ao grande volume de informações e em alguns casos a necessidade de pré-processamento

dos dados, é feito com o objetivo de identificar biomarcadores. A utilização de medidas de centralidade são empregadas para avaliar a importância de um nó da rede em referência aos demais, estudando o comportamento dos genes como sistema integrado e não de forma separada, assim adquirindo uma maior compreensão da doença para diagnósticos e prognósticos.

Ponto em comum nos 4 trabalhos relacionados que nenhum deles utilizou balanceamento de classes, já que esse é um problema enfrentado quando se trabalha com dados reais, a abordagem adota por esse é de reduzir a quantidade de genes empregando técnicas de como WGCNA a qual separa os genes em módulos os quais são considerados devidos a seu *p-value*, deixando de considerar outros genes, pode-se considerar a possibilidade de que um genes importante não esteja presente no módulo escolhido pelo pesquisador.

Dentre os trabalhos somente (WANG; LIU, 2017), revela qual medida utilizou para analisar o resultado do seu modelo de aprendizagem de máquina, mas todos o trabalhos não especificaram de forma clara como utilizaram seus modelos, quais foram seus preditores ou quantas classes utilizaram, quando empregaram *cross validation* não relataram o número de *folds* e ou qual foi o particionamento dos dados para validação.

Nenhum dos trabalhos empregou técnicas de seleção de atributos, apenas reduziram o tamanho do seu problema diminuindo a quantidade de genes a serem estudados.

A linguagem R é amplamente empregada em todos os trabalhos para implementação de suas metodologias, mudando apenas as bibliotecas empregadas e as análises realizadas.

Na Tabela 2 é demonstrada uma comparação entre os trabalhos relacionados.

Tabela 2: Comparação entre os trabalhos

Trabalhos	Balanceamento	Aprendizagem de Máquina	Correlação de Person	Métricas de Centralidade
WANG; LIU (2017)	Não	Utiliza	Utiliza	Utiliza
ZHANG et al. (2010)	Não	Utiliza	Não	Não
(WANG; LI; CHEN, 2018)	Não	Não	Não	Utiliza
ZHANG et al. (2017)	Não	Utiliza	Não	Utiliza

4 BIOMARKER FINDER

Neste capítulo é apresentada a metodologia proposta para descoberta de possíveis novos biomarcadores a qual possui o acrônimo de *Biomarker Finder*.

4.1 *Biormarker Finder*

A metodologia proposta consiste na utilização redes para avaliação de dados de expressão gênica, empregando análise topológica e Aprendizado de Máquina para descoberta de possíveis biomarcadores. Esta metologia foi implementada com a linguagem R, usando as bibliotecas *igraph* (CSARDI; NEPUSZ, 2006) para construção e análise da rede e a *caret* KUHN (2015) para os algoritmos. A Figura 17 mostra uma visão geral do *Biomarker Finder*.

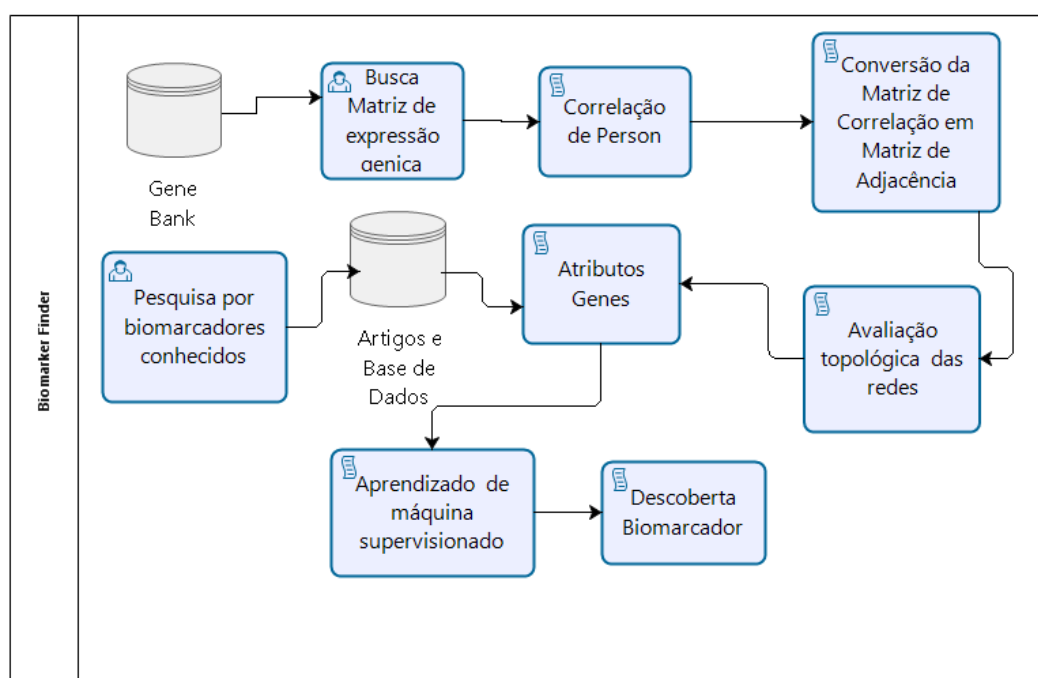


Figura 17: Visão geral da Metodologia. Fonte: Dados do Autor.

4.1.1 Escolha dos dados de Expressão Gênica

A primeira etapa da metodologia proposta é a definição de dados de expressão gênica para diferentes grupos experimentais como grupo controle (indivíduos saudáveis) e grupo doença (indivíduos acometidos de alguma enfermidade). Nesta etapa o usuário fornece uma matriz de expressão gênica já processada. A Tabela 3 representa a entrada dos dados aceita pela metodologia proposta. Na coluna genes, encontra-se a nomenclatura dos genes e as demais colunas contêm o código da amostra e os valores de expressão gênica.

Tabela 3: Matriz de Expressão Gênica

Genes	GSM123522	GSM123523	GSM123524
A1CF	5.9758162	5.948891942	5.766238159
A2M	10.85105385	10.73637596	10.76301999
A4GNT	5.872587417	5.686625374	5.798817904
...	...	...	...

4.1.2 Correlação de Pearson e Descoberta do Limiar

Na segunda etapa é calculada a correlação Pearson. Verifica-se quantos genes são comuns em cada limiar de correlação de acordo com o nível de expressão dos genes entre cada condição (1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1). Idealmente, buscamos um limiar que mantenha uma boa proporção de genes em comum entre as matrizes de correlação de diferentes grupos permitindo que exista um conjunto de elementos em comum de tamanho suficiente para coleta de atributos para nas etapas posteriores. Uma representação dessa etapa pode ser visualizada na Figura 10.

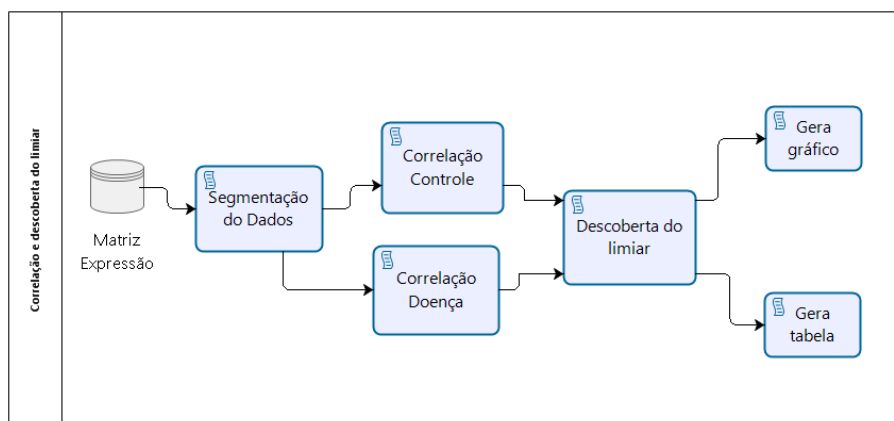


Figura 18: Correlação de Pearson e Descoberta do Limiar. Fonte: Dados do Autor.

Durante essa etapa, conforme ilustrado na Figura 18 é realizada a separação das amostras que pertencem ao controle e a doença. Após, é calculada a correlação da expressão dos genes para cada condição e logo em seguida verifica-se o percentual de genes comuns entre as duas matrizes de correlação para cada limiar. Esse resultado é exibido em um gráfico e em um arquivo de texto, como pode ser visualizado na Figura 19.

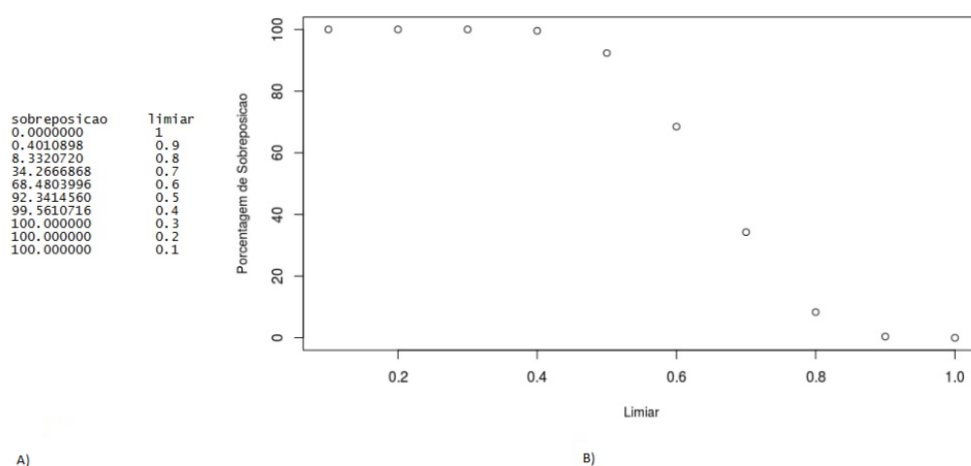


Figura 19: Gráfico para Escolha do Limiar. Fonte: Dados do Autor.

Após a seleção dos genes é realizada uma segunda verificação buscando o limiar de correlação em que a rede da doença mantenha-se conectada, para que não ocorra a existência de nós da rede sem conexão ou pequenas sub-redes as quais podem influenciar algumas métricas de rede.

4.1.3 Construção das Redes

Para construção das redes é realizada a seleção dos genes conforme o limiar de correlação da expressão escolhido. Neste trabalho optou-se por um limiar que mantenha de 80% a 90% dos genes em comum entre ambas as redes. As matrizes de correlação são convertidas em matrizes de adjacência para construção de grafos não direcionados que representam redes de co-expressão nas quais nós representam os genes e as arestas a correlação entre os níveis de expressão dos genes. Na Figura 20 são exibidos os passos realizados para construção da rede.

Com a construção do grafo não direcionado é possível fazer uma análise topológica da rede, empregando as medidas de centralidade que possibilitam verificar a importância de um nó na rede (PAVLOPOULOS et al., 2011).

4.1.4 Extração de atributos e dados de treinamento

Nesta etapa foram calculadas ao todo 9 medidas topológicas de rede, sendo destas 6 foram escolhidas utilizando o algoritmo CFS, devido a pontuação atribuída pelo algoritmo as medidas selecionadas estão descritas na seção 2.5.4 do texto. Na Tabela 4 é apresentada de forma ilustrativa 4 das 6 medidas utilizadas para treinamento.

As medidas topológicas das redes são utilizadas para o treinamento e validação dos modelos, da seguinte forma:

- Primeiro é realizada uma pesquisa no estado da arte e em base de dados buscando por genes que sejam reconhecidos como biomarcadores para criação dos

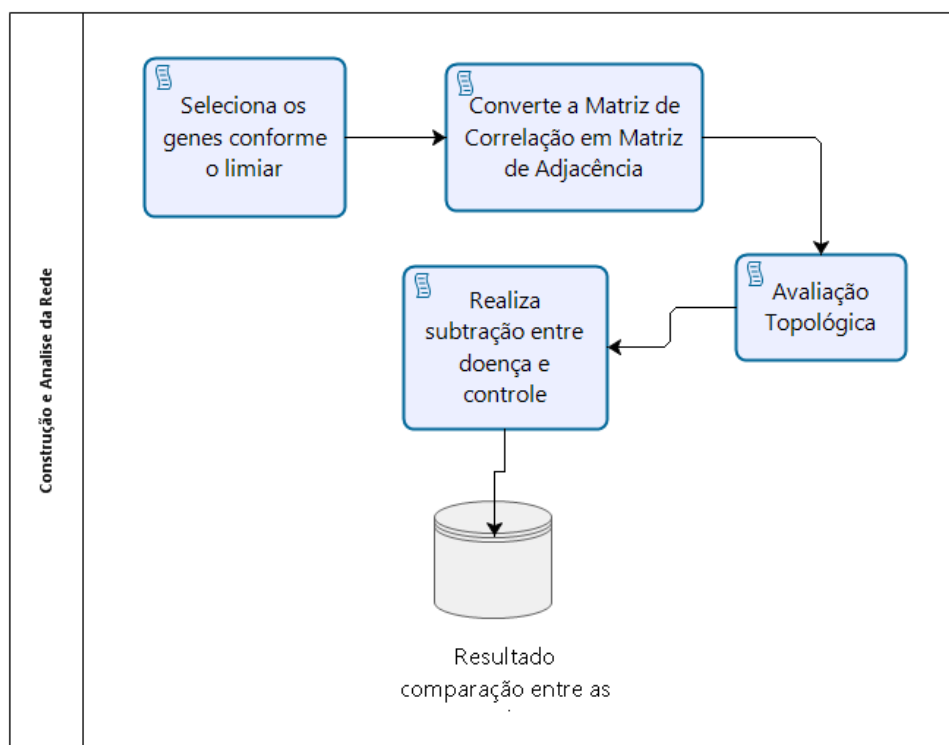


Figura 20: Construção e Análise da Rede. Fonte: Dados do Autor.

Tabela 4: Extração características do nós

Genes	Degree	Betweenness	PageRank	Strength
A1CF	1098	35352	0.000192551472601964	642.531527679267
A2M	1	0	4.27793892432442e-05	0.500581933608985
A4GNT	126	805	5.83373495606464e-05	67.5668487175874
...	...	...	...	...

preditores, assim marcando no *dataset* os genes;

- Segundo passo realiza-se um subtração entres as medidas da doença e do controle, nó a nó sendo o resultado dessa operação utilizada para treinar;
- Terceiro passo também e realizada a criação de um segundo *dataset* de treinamento e validação, considerando a divisão entre as medidas topológicas;
- Quarto passo consideramos apenas a rede da doença para treinamento e validação, por essa geralmente apresentar diferenças topológicas mais acentuada entre os genes.

Para descoberta dos genes considerados biomarcadores, empregamos na busca o termo "*list of biomarkers for dilated cardiomyopathy*" no gene cards, malacards, cardiodb, google acadêmico entre outros. A partir desse termo, obteve-se o trabalho e as fontes de dados de (HERSHBERGER; HEDGES; MORALES, 2013):

- https://www.cardiodb.org/acgv/acgv_disease.php?icc=DCM
- <https://academic.oup.com/eurheartj/article/36/18/1123/2293182>
- <https://onlinelibrary.wiley.com/doi/abs/10.1002/ejhf.255>
- https://www.genome.jp/dbget-bin/www_bget?hsa05414
- <https://www.malacards.org/>

A partir dessa pesquisa realizada criou-se um arquivo com os nomes dos genes que são considerados como biomarcadores para sua marcação no *dataset*. Dependendo das doenças que se deseja analisar, a busca por biomarcadores pode ser feita através de diferentes fontes e palavras-chaves para encontra-los e/ou ter conhecimento prévio da doença a qual deseja descobrir biomarcadores para criação dos preditores de Aprendizagem de Máquina.

Na base de dados genecards, encontrou-se um total de 926 genes relacionados a cardiomiopatia dilatada os quais são apenas associados a doença não possuindo características de um biomarcador e um total de 230 genes que são considerados biomarcadores.

Para os genes que não são considerados biomarcadores, não existem estudos para identificar quando os mesmos não o são. Para minimizar essa dificuldade, escolheu-se genes aleatoriamente para que sejam considerados não-biomarcadores. Com a grande diferença de quantidade entre as classes, empregou-se técnica de balanceamento, *Down Sample*, *Up Sample* e *Smote*, também foi utilizado o algoritmo de seleção de atributos visando alcançar melhores resultados nos classificadores.

4.1.5 Descoberta de Possíveis Biomarcadores

Para descoberta de biomarcadores empregamos o resultado da subtração das medidas topológicas das redes e da pesquisa dos genes que são considerados como biomarcadores, onde o atributo *Biomarker* será utilizado como alvo.

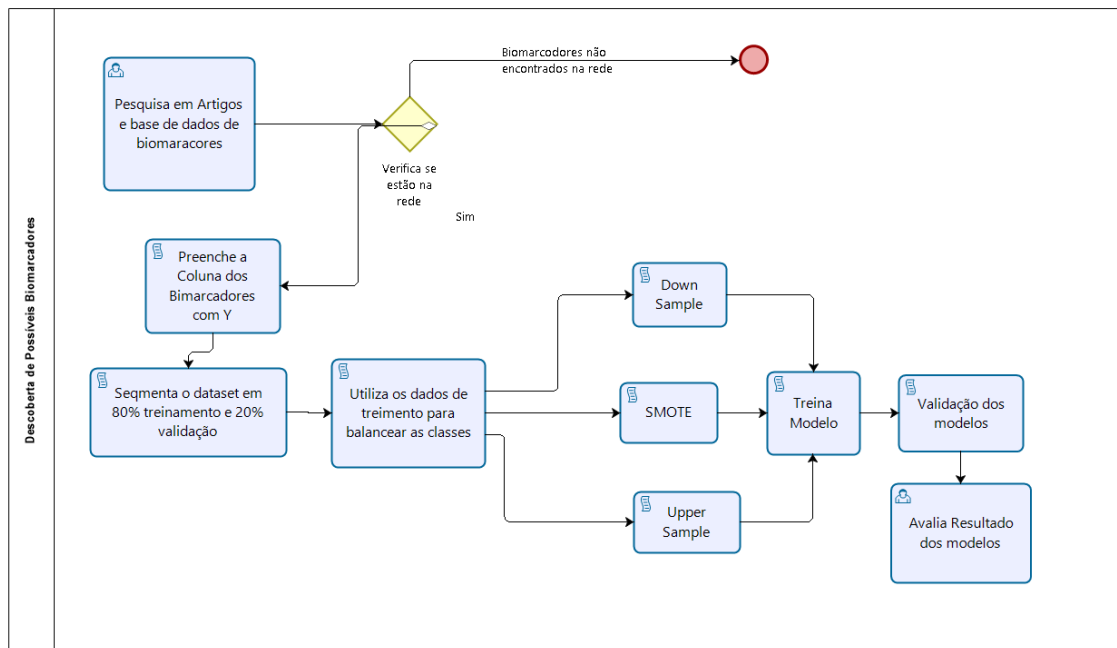


Figura 21: Descoberta de Possíveis Biomarcadores. Fonte: Dados do Autor.

Após a pesquisa realizada nos repositórios dos genes considerados biomarcadores, o *Biomarker Finder* realizou uma verificação de quais genes estão contidos nas redes; seguidamente, são adicionados aos dados um novo atributo chamado *Biomarker*, onde (Y) é para genes que são biomarcadores, (N) para genes que não são biomarcadores e (R) para genes relacionados.

Na Tabela 4 é representado um exemplo dos dados de treinamento do algoritmo.

Tabela 5: Extração características do nós

Genes	Degree	Betweenness	PageRank	Biomarker
A1CF	-1004	30436	9.39046672466202	Y
A2M	-1553	-6866	-4.51025029568884	Y
A4GNT	-518	-1159	-1.04343502864928	N
...	...	...	...	...

Para treinamento dos modelos utilizamos as seguintes combinações biomarcadores versus não biomarcadores e biomarcadores versus relacionados, buscando verificar se os biomarcadores são bons preditores com relação aos genes relacionados à doença e os não biomarcadores.

Os *datasets* foram divididos em 80% a 20%, sendo a maior porção para treinamento utilizando *cross-validation* com 10 *folds* e a menor porção, 20% é usada para validação dos modelos, pois busca-se assim manter o máximo possível de equilíbrio entre as classes.

Após a segmentação do *dataset* na parte correspondente ao treinamento, são aplicadas três técnicas de balanceamento de classes, nas quais consideramos utilizar: (i) *Down Sample* que escolhe aleatoriamente instâncias da classe majoritária, até equalizar com a

quantidade de exemplos da classe minoritária;(ii) Upper Sample que realiza sobreposições continua até que classe minoritária equiparar-se na quantidade de elementos com a maioritária; e como ultima técnica foi utiliza SMOTE esse método de sobre amostragem da classe minoritária envolve a criação de exemplos de classes minoritárias de forma sintética visando reduzir os problemas de balanceamentos de classes.

Os 20% das validação utilizou-se *Down Sample* para balancear as classes e sem balanceamento para testar a capacidade dos algoritmos de classificar.

Na tabela 6 pode-se visualizar a segmentação dos *datasets*.

Tabela 6: Conjunto de dados balanceado para treinamento e validação

Rede	Dataset	Tipo de Balanceamento	Treinamento	Validação
Doença - Controle	Biomarcadores VS Não Biomarcador	Down Sample	360	90
	Biomarcadores VS Não Biomarcador	Smote	1800	90
	Biomarcadores VS Não Biomarcador	Up Sample	17276	90
	Biomarcadores VS Relacionados	Down Sample	360	90
	Biomarcadores VS Relacionados	Smote	1800	90
	Biomarcadores VS Relacionados	Up Sample	1128	90
Doença / Controle	Biomarcadores VS Não Biomarcador	Down Sample	360	90
	Biomarcadores VS Não Biomarcador	Smote	1800	90
	Biomarcadores VS Não Biomarcador	Up Sample	17276	90
	Biomarcadores VS Relacionados	Down Sample	360	90
	Biomarcadores VS Relacionados	Smote	1800	90
	Biomarcadores VS Relacionados	Up Sample	1128	90
Doença	Biomarcadores VS Não Biomarcador	Down Sample	360	90
	Biomarcadores VS Não Biomarcador	Smote	1800	90
	Biomarcadores VS Não Biomarcador	Up Sample	17276	90
	Biomarcadores VS Relacionados	Down Sample	360	90
	Biomarcadores VS Relacionados	Smote	1800	90
	Biomarcadores VS Relacionados	Up Sample	1128	90

Tabela 7: Conjunto de dados sem balanceado para treinamento e validação

Rede	Dataset	Tipo de Balanceamento	Treinamento	Validação
Doença - Controle	Biomarcadores VS Não Biomarcador	Sem Balanceamento	8818	2204
	Biomarcadores VS Relacionados	Sem Balanceamento	744	185
Doença / Controle	Biomarcadores VS Não Biomarcador	Sem Balanceamento	8818	2204
	Biomarcadores VS Relacionados	Sem Balanceamento	744	185
Doença	Biomarcadores VS Não Biomarcador	Sem Balanceamento	8818	2204
	Biomarcadores VS Relacionados	Sem Balanceamento	744	185

Considerações

Neste capítulo foi apresentada a metodologia proposta que utiliza as medidas topológicas de rede para verificar a diferença entre doença e o controle, e busca identificar possíveis biomarcadores empregando aprendizagem na máquina para o reconhecimento dos padrões semelhantes aos dos genes que já são cientificamente comprovados como biomarcadores.

Na Tabela 8 pode ser visualizada uma comparação da metodologia proposta com trabalhos relacionados.

Tabela 8: Comparação dos trabalhos relacionados com Biomarker Finder

Trabalhos	Balancamento	Aprendizagem de Máquina	Correlação de Person	Métricas de Centralidade
WANG; LIU (2017)	Não	Utiliza	Utiliza	Utiliza
ZHANG et al. (2010)	Não	Utiliza	Não	Não
WANG; LI; CHEN (2018)	Não	Não	Não	Utiliza
ZHANG et al. (2017)	Não	Utiliza	Não	Utiliza
Biomarker Finder	Utiliza	Utiliza	Utiliza	Utiliza

Biomarker Finder diferencia-se dos demais trabalhos por utilizar mais de uma métrica de centralidade para caracterizar os nós da rede, assim buscando por genes que possam ser biomarcadores considerando seus atributos topológicos. Outro ponto dessa pesquisa é a busca por genes que já são considerados biomarcadores na literatura para a criação dos preditores para o Aprendizado de Máquina, enquanto os outros trabalhos relacionados buscam genes com limiar de correlação alto para considerar como preditores, como realizado por WANG; LIU (2017).

No estudo feito por ZHANG et al. (2010), foi considerado os genes que se relacionaram com ZAP70, um biomarcador bem conhecido dessa doença, buscando identificar redes altamente conectadas.

Já na pesquisa realizada por WANG; LI; CHEN (2018), foi aplicada a ferramenta WGCNA, para encontrar módulos que sejam mais relacionados à doença e considera que os com maior *p-value* dentro dos módulos são considerados hubs.

A pesquisa realizada por ZHANG et al. (2017), os autores consideraram que genes com alto grau dentro de um módulo é um *hub* e os genes diferencialmente expressos nos módulos são mais relacionados à doença.

Treinar um algoritmo de Aprendizagem de Máquina considerando genes que tem um limiar de correlação alto como biomarcadores, pode gerar muitos resultados falso positivos por que nem sempre os genes com o maior limiar são realmente biomarcadores.

Neste estudo, buscou-se encontrar genes considerados biomarcadores para cardiomiopatia dilatada na literatura os quais foram utilizados como preditores para o aprendizado de máquina, possibilitando dessa forma, maior precisão para a classificação dos possíveis

biomarcadores.

Uma deficiência partilhada entre todos os trabalhos é a questão de não existir outros trabalhos que identifiquem genes que não são biomarcadores.

Para solucionar esse problema, WANG; LIU (2017), considerou genes que apresentaram baixo limiar de co-expressão, para definir os preditores para não-biomarcadores, enquanto os demais estudos preocuparam-se em analisar os genes considerando suas funcionalidades, através de um enriquecimento funcional dos mesmos com a finalidade de avaliar as vias metabólicas.

Presente estudo, diferencia-se por escolher aleatoriamente os genes que serão considerados não-biomarcadores.

A metodologia proposta deste estudo, *Biomarker Finder* pode ser empregada para diferentes doenças, sendo apenas necessário uma busca pelos genes que são considerado como biomarcadores da doença que se deseja descobrir possíveis biomarcadores, enquanto as outras metodologias propostas aplicam-se a determinados casos por suas peculiaridades como o LLC por utilizar um gene específico como alvo e analisar os que se relacionam diretamente com ele.

O *Biomarker Finder* se diferencia dos demais trabalhos relacionados por ter considerado o emprego de técnicas de balanceamento de classes, visando avaliar a relação entre genes considerando aqueles contidos entre 80% a 90% do universo total da amostra de genes do *dataset*.

Outro destaque é o emprego de medidas topológicas como (*PageRank*, *Alpha Centrality*, *Strenght*, *Betweenness* e *Closeness*), que os demais estudos para avaliar a relação entre os diferentes genes.

Dessa forma, buscando reconhecer padrões existentes na rede por meio de algoritmos de aprendizado de máquina, utilizou-se três algoritmos para classificação de possíveis biomarcadores sendo eles o SVM, Random Forest e J48.

Com referência aos trabalhos relacionados pode-se considerar o *Biomarker Finder* como uma metodologia inovadora, por considerar medidas de topológicas como dados de treinamento para algoritmos de aprendizagem de máquina, já que muitos trabalhos apenas utilizam essa métrica para descobrir os *hubs* das redes.

5 RESULTADOS

Os resultados apresentados neste capítulo referem-se ao *dataset* GSE5406 da pesquisa realizada por HANNENHALLI et al. (2006), disponível no GEO, no qual esse *dataset* possui dados de pacientes com cardiomiopatia dilatada (DCM). Um total de 194 amostras de pacientes com DCM e de 16 amostras de controle de pessoas saudáveis, foram utilizadas para analisar a expressão dos genes através da técnica de microarranjo.

Como primeira etapa da metodologia proposta, realizou-se a busca de um limiar de correlação de Pearson que mantivesse um total de genes comuns entre doença e controle na faixa compreendida de 80% a 90%, considerando o universo total da amostra.

Tabela 9: Limiar de correlação de Pearson usado para identificação de genes comuns

Percentual de Genes Comuns	Limiar	Número de Genes
0.0000000	1	0
0.5448766	0.9	72
10.2164371	0.8	1350
36.2342970	0.7	4788
67.0955048	0.6	8866
88.7392160	0.5	11726
98.1761768	0.4	12973
99.9621613	0.3	13209
100.0000000	0.2	13214
100.0000000	0.1	13214

Como pode ser visualizado na Tabela 9, o limiar selecionado para construção da rede foi de 0.5, devido ao fato de estar dentro do intervalo estipulado na metodologia proposta. Foram construídas duas redes para avaliação da topologia, contendo 11726 genes. Após os cálculos das medidas de centralidade, a fim de gerar um valor de métrica por gene para comparar as medidas topológicas das redes, realizaram-se duas operações: subtração da doença pelo controle e uma divisão entre esses valores. O resultado dessa subtração e divisão entre as métricas foi utilizado para o treinamento dos modelos de aprendizagem de máquina.

O segundo passo foi a busca pelos genes que são considerados biomarcadores. Como

neste estudo de caso foi abordado DCM, buscou-se por genes que já são biomarcadores cientificamente comprovados na literatura.

Concluída a etapa de busca por biomarcadores, identificou-se um total de 230 genes considerados biomarcadores. Dentre artigos de *review* e as bases citadas na seção 4.1.4, do total de 230 biomarcadores, encontrou-se 225 que estão presentes nas redes construídas, cujo os quais receberam a letra Y na coluna *Biomarkers* dos dados para treinamento dos algoritmos de aprendizagem. Como não são realizados estudos para definir genes que não são biomarcadores, os mesmos foram escolhido de forma aleatória, para serem considerados como não biomarcadores recebendo então como atributo a letra N e para genes relacionados a doença encontrados no GENECARDS foi atribuída a letra R para um total de 920 genes.

Os resultados utilizando esse *dataset*, são apresentados na próxima secção em tabelas considerando, área da curva ROC, acurácia, sensibilidade e especificidade, os algoritmos empregados foram, SVM como kernel radial, random forest e J48, em todos utilizando configurações padrões, sendo que o SVM e random forest a função padrão busca melhor ajuste da função custo para SVM e a melhor quantidade de árvores para random forest.

5.1 Resultados

5.1.1 Resultado sem aplicação de balanceamento de classes

Nesta secção são apresentados o resultados, sem o emprego do balanceamento de classe, com objetivo demonstrar a importância de ter classes equilibradas para modelos de classificação.

Nas Tabelas(10,11 e 12) a seguir pode ser visualizado o resultado obtidos oriundos do treinamento de com as classes desbalanceadas.

Tabela 10: Resultado obtidos executando SVM sem balanceamento

Rede	Dataset	Treinamento	Validação	Algoritmo SVM			
				AUC	Acurácia	Sensibilidade	Especificidade
Doença - Controle	Biomarcadores VS Não Biomarcador	8818	2204	0,599	0,979	0	1
	Biomarcadores VS Relacionados	744	185	0,541	0,758	0	1
Doença / Controle	Biomarcadores VS Não Biomarcador	8818	2204	0,541	0,758	0	1
	Biomarcadores VS Relacionados	744	185	0,572	0,357	0	1
Doença	Biomarcadores VS Não Biomarcador	8818	2204	0,542	0,979	0	1
	Biomarcadores VS Relacionados	744	185	0,557	0,756	0	1

Tabela 11: Resultado obtidos executando Random Forest sem balanceamento

Rede	Dataset	Treinamento	Validação	Algoritmo Random Forest			
				AUC	Acurácia	Sensibilidade	Especificidade
Doença - Controle	Biomarcadores VS Não Biomarcador	8818	2204	0,575	0,979	0	1
	Biomarcadores VS Relacionados	744	185	0,557	0,736	0	1
Doença / Controle	Biomarcadores VS Não Biomarcador	8818	2204	0,557	0,736	0	1
	Biomarcadores VS Relacionados	744	185	0,545	0,534	0	1
Doença	Biomarcadores VS Não Biomarcador	8818	2204	0,665	0,979	0	1
	Biomarcadores VS Relacionados	744	185	0,549	0,724	0	1

Tabela 12: Resultado obtidos executando J48 sem balanceamento

Rede	Dataset	Treinamento	Validação	Algoritmo J48			
				AUC	Acurácia	Sensibilidade	Especificidade
Doença - Controle	Biomarcadores VS Não Biomarcador	8818	2204	0,500	0,979	0	1
	Biomarcadores VS Relacionados	744	185	0,500	0,758	0	1
Doença / Controle	Biomarcadores VS Não Biomarcador	8818	2204	0,500	0,758	0	1
	Biomarcadores VS Relacionados	744	185	0,494	0,365	0	1
Doença	Biomarcadores VS Não Biomarcador	8818	2204	0,575	0,978	0	1
	Biomarcadores VS Relacionados	744	185	0,556	0,735	0	1

5.2 Resultados aplicando balanceamento de classes

Nesta subsecção serão apresentados os resultados obtidos empregando os *datasets*, com balanceamento classes.

Nas tabelas 13, 14 e 15, é possível visualizar os resultados obtidos empregando os algoritmos SVM, Random Forest e J48, com os dados de treinamento oriundos da subtração das métricas da rede doença pela rede controle, divisão das métricas da rede doença pela rede controle, e as métricas da rede doença, utilizando as técnicas de balanceamento *down sample*, *smote* e *up sample*.

A cor laranja representa a subtração das medidas da rede doença pela rede controle, a cor amarelo a divisão para divisão entre rede doença e controle e cor azul somente as medidas da rede doença.

Tabela 13: Resultado obtidos executando SVM com balanceamento

Dataset	Balanceamento de	Treinamento	Validação	Algoritmo SVM			
				AUC	Acurácia	Sensibilidade	Especificidade
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,584	0,513	0,644	0,510
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,592	0,768	0,355	0,777
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,572	0,759	0,355	0,768
Biomarcadores VS Relacionados	Down Sample	360	90	0,516	0,494	0,433	0,514
Biomarcadores VS Relacionados	Smote	1800	90	0,554	0,618	0,344	0,705
Biomarcadores VS Relacionados	Up Sample	1128	90	0,553	0,567	0,566	0,567
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,519	0,491	0,522	0,490
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,543	0,639	0,433	0,644
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,530	0,662	0,355	0,666
Biomarcadores VS Relacionados	Down Sample	360	90	0,572	0,357	0,588	0,283
Biomarcadores VS Relacionados	Smote	1800	90	0,495	0,564	0,322	0,641
Biomarcadores VS Relacionados	Up Sample	1128	90	0,586	0,360	0,633	0,273
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,700	0,655	0,488	0,822
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,599	0,600	0,422	0,777
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,644	0,655	0,466	0,844
Biomarcadores VS Relacionados	Down Sample	360	90	0,486	0,496	0,507	0,485
Biomarcadores VS Relacionados	Smote	1800	90	0,590	0,566	0,533	0,600
Biomarcadores VS Relacionados	Up Sample	1128	90	0,534	0,532	0,635	0,428

Tabela 14: Resultado obtidos executando Random Forest com balanceamento

Dataset	Tipo de Balanceamento	Treinamento	Validação	Algoritmo Random Forest			
				AUC	Acurácia	Sensibilidade	Especificidade
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,602	0,503	0,655	0,500
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,587	0,792	0,366	0,801
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,619	0,979	0	0,999
Biomarcadores VS Relacionados	Down Sample	360	90	0,519	0,473	0,544	0,450
Biomarcadores VS Relacionados	Smote	1800	90	0,541	0,583	0,511	0,606
Biomarcadores VS Relacionados	Up Sample	1128	90	0,528	0,696	0,188	0,858
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,621	0,544	0,6	0,542
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,609	0,695	0,422	0,700
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,538	0,977	0	0,979
Biomarcadores VS Relacionados	Down Sample	360	90	0,545	0,534	0,533	0,535
Biomarcadores VS Relacionados	Smote	1800	90	0,543	0,569	0,477	0,599
Biomarcadores VS Relacionados	Up Sample	1128	90	0,524	0,534	0,533	0,535
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,655	0,644	0,577	0,711
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,586	0,588	0,333	0,844
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,633	0,655	0,466	0,844
Biomarcadores VS Relacionados	Down Sample	360	90	0,499	0,510	0,528	0,492
Biomarcadores VS Relacionados	Smote	1800	90	0,525	0,503	0,785	0,221
Biomarcadores VS Relacionados	Up Sample	1128	90	0,521	0,5	0,992	0,007

Tabela 15: Resultado obtidos executando J48 com balanceamento

Dataset		Tipo de Balanceamento		Treinamento	Validação	Algoritmo J48			
						AUC	Acurácia	Sensibilidade	Especificidade
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,516	0,255	0,788	0,244		
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,547	0,505	0,600	0,503		
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,543	0,412	0,677	0,406		
Biomarcadores VS Relacionados	Down Sample	360	90	0,469	0,252	0,888	0,049		
Biomarcadores VS Relacionados	Smote	1800	90	0,545	0,500	0,422	0,585		
Biomarcadores VS Relacionados	Up Sample	1128	90	0,475	0,448	0,588	0,404		
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,610	0,659	0,500	0,662		
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,588	0,484	0,700	0,480		
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,560	0,808	0,355	0,818		
Biomarcadores VS Relacionados	Down Sample	360	90	0,494	0,365	0,744	0,244		
Biomarcadores VS Relacionados	Smote	1800	90	0,629	0,6	0,411	0,698		
Biomarcadores VS Relacionados	Up Sample	1128	90	0,484	0,602	0,255	0,712		
Biomarcadores VS Não Biomarcador	Down Sample	360	90	0,644	0,644	0,511	0,777		
Biomarcadores VS Não Biomarcador	Smote	1800	90	0,605	0,633	0,355	0,911		
Biomarcadores VS Não Biomarcador	Up Sample	17276	90	0,623	0,633	0,422	0,844		
Biomarcadores VS Relacionados	Down Sample	360	90	0,448	0,557	0,278	0,835		
Biomarcadores VS Relacionados	Smote	1800	90	0,382	0,578	0,442	0,714		
Biomarcadores VS Relacionados	Up Sample	1128	90	0,468	0,500	0,364	0,635		

Nesta seção foram apresentados o resultado obtidos, empregando as três técnicas de balanceamento de classes e com diferentes conjuntos de treinamento.

5.3 Discussão dos Resultados

Nesta seção será apresentada a discussão e a comparação dos resultados obtidos.

Quando busca-se treinar um algoritmo de aprendizado de máquina para reconhecer um padrão (classificar), há diferentes fatores que impactam no seu desempenho como classificador. Entre eles, o desbalanceamento de classes, porque geralmente o fenômeno que é objeto de estudo acontece em menor escala, acarretando em uma menor quantidade de exemplos para o algoritmo aprender seu padrão.

Um exemplo desse problema está contido nesta pesquisa em questão, na qual realizou-se uma busca na literatura para encontrar genes que são considerados biomarcadores de cardiomiopatia dilatada, foram encontrados 230 genes e desses, 225 estavam presentes no *dataset*, que utilizou-se para testar a metodologia proposta.

O problema do balanceamento pode ser evidenciado na Figura 22, na qual é possível visualizar a diferença entre as classes para o treinamento dos modelos, na classe dos biomarcadores tem 180 instâncias para treinamento, para genes relacionados um montante de 564 e para genes que são considerados não biomarcadores um montante de 8630. O desbalanceamento impacta negativamente nos modelos de classificação por oferecer mais exemplos de uma classe do que outra, tornando o modelo muito ajustado para identificar apenas um padrão muito predominante.

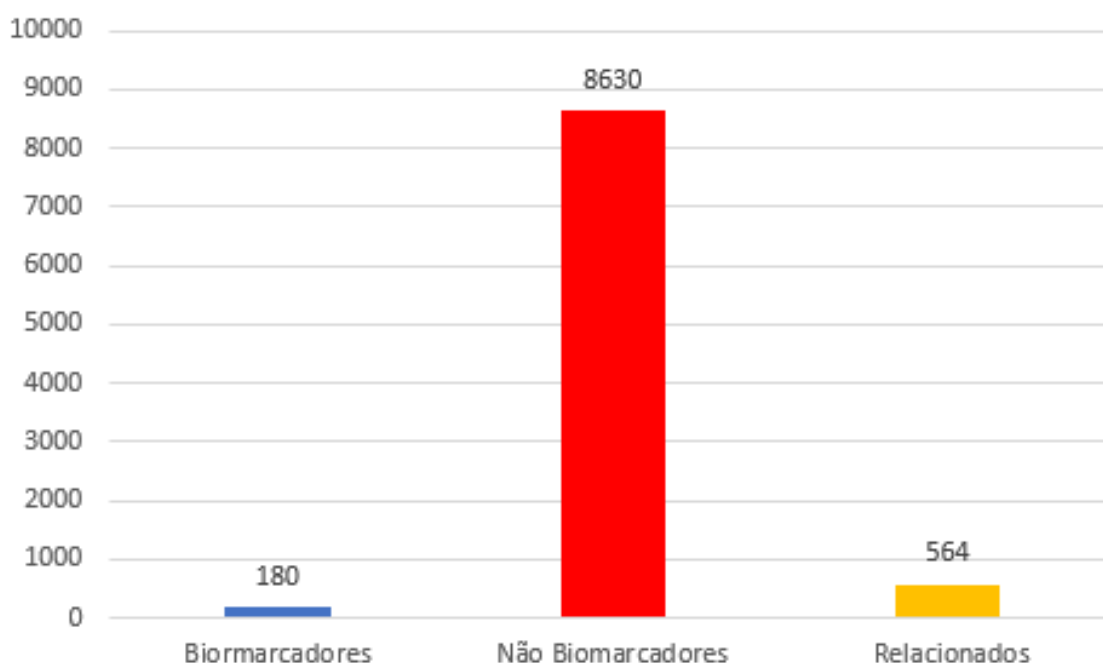


Figura 22: Gráfico de desbalanceamento de classes no conjunto de treinamento. Fonte: Dados do Autor.

Quando treinou-se os modelos com as classes desbalanceadas, não obteve-se AUC esperado para os algoritmos como pode ser observado na Figura 23, este fato ocorre devido a natureza do problema de existir um desequilíbrio na proporção entre as classes observadas.

Então buscou-se melhorar o desempenho dos classificadores considerando duas frentes, a primeira aplicando balanceamento de classes nas *feature*, e por fim a combinação das *feature* quanto ao tipo de rede, buscando verificar se a comparação entre as medidas da rede doença vs controle atrás de subtração ou uma divisão seria melhor que avaliar métricas somente da rede doença.

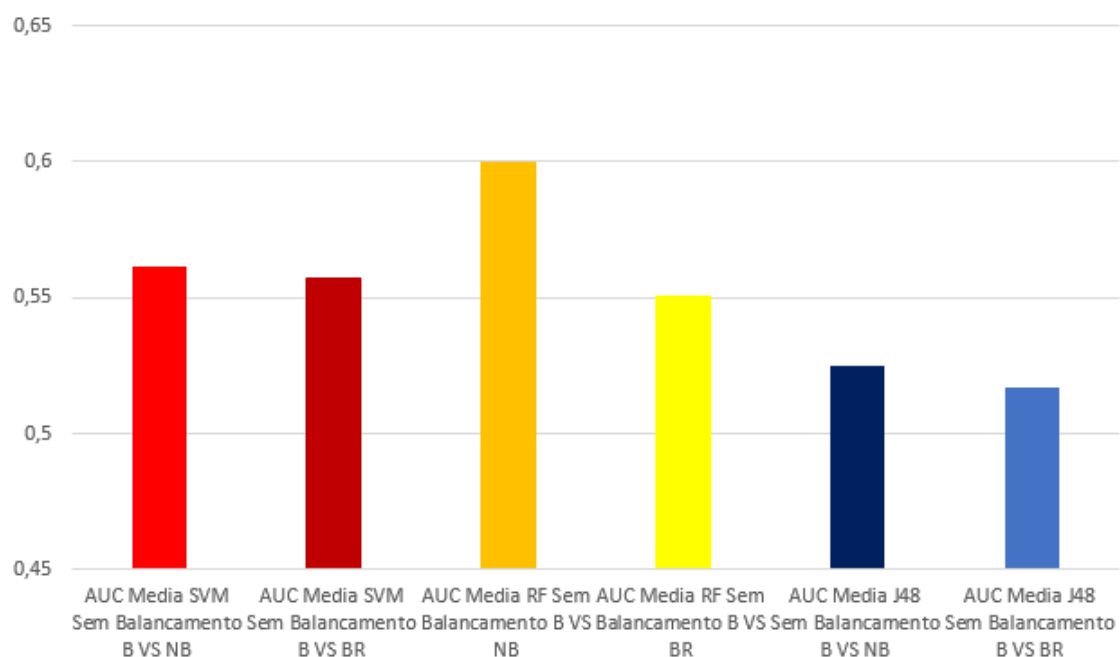


Figura 23: Gráfico AUC médio sem balanceamento por algoritmo. Fonte: Dados do Autor.

Já na execução do treinamento dos modelos verificou-se a baixo desempenho dos mesmos quando empregou-se a as *features* resultantes de subtração e divisão das redes, foi quando observamos no trabalho de WANG; LI; CHEN (2018), qual considerou apenas as amostras da doença para construção e análise de rede, sendo assim optou-se por testar essa abordagem considerando apenas a rede da doença para treinamento, foi possível observar uma melhora do AUC média em relação as outras duas *features*.

Em um segundo momento realizou o teste considerando as *features* balanceadas por rede e pode-se constatar o mesmos ganho de AUC para rede doença em relação a subtração e divisão das métricas das redes.

Uma outra análise realizada nos resultados de AUC foi comparar o ganho médio de AUC na aplicação do balanceamento versus não balanceamento, na Figura 24, quando empregamos SVM kernel radial, observou-se um ganho de AUC empregando o balanceamento *down sample* em relação as *feature* desbalanceadas utilizando biomarcadores vs

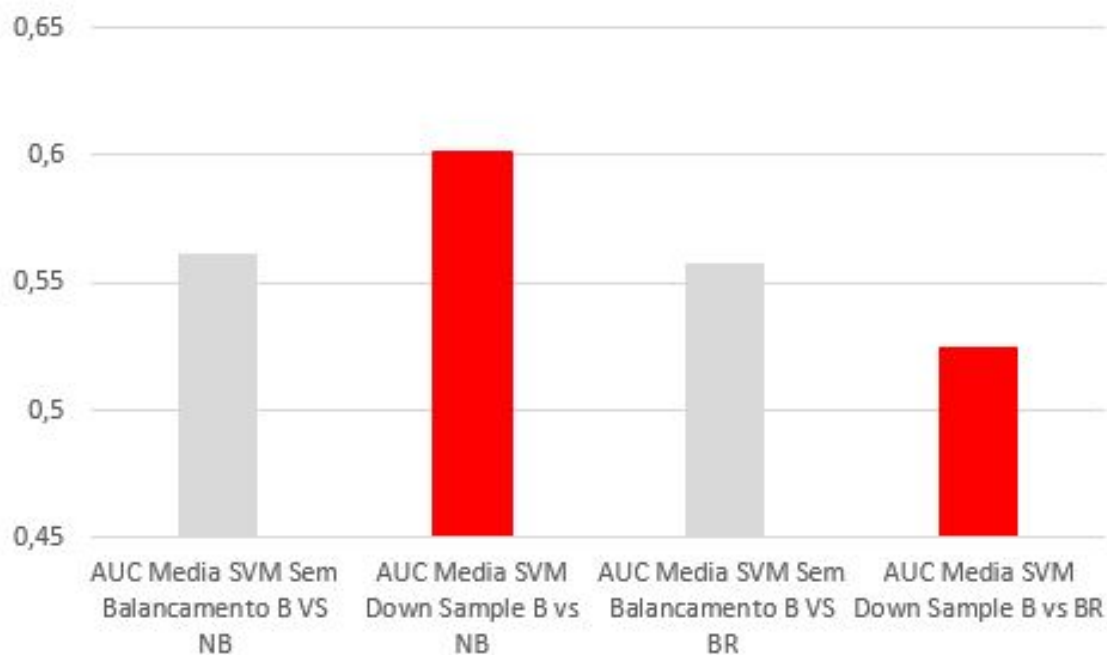


Figura 24: Gráfico AUC da média SVM down sample vs sem balanceamento. Fonte: Dados do Autor.

não biomarcadores.

Mas quando utilizou-se biomarcadores vs genes relacionados buscando classificar possíveis biomarcadores, os resultados foram inferiores aos dados não balanceados acredita-se que o motivo de perda de AUC foi pelo fato de de genes relacionados terem características topologias próxima ao dos biomarcadores e também pelo desbalanceamento induzir ao modelo a classificar a classe com maior quantidade de exemplos.

No emprego do algoritmo Random Forest analisou-se a mesma situação biomarcadores vs não biomarcadores e biomarcadores vs genes relacionados, pode ser evidenciado na Figura 25, esse modelo obteve melhores resultados quando aplicou-se *down sample* nos dados de biomarcadores vs não biomarcadores, acredita-se isso deve-se ao fato da amostra dos genes que foram considerados como não biomarcadores apresentarem diferenças topológicas entre as classes.

Quando observa-se os biomarcadores vs os genes relacionados com balanceamento down sample, a AUC foi inferior aos dados não balanceados deve-se ao fato das duas classes terem características topológicas muito próximas e pelo possível *overfitting* no modelo que atingiu uma acuracia 0,97 para não balanceamento e um sensibilidade igual 0 significando que modelo conseguiu classificar apenas os genes relacionados.

Quando utilizou-se o algoritmo J48, obteve-se resultados com mesmas tendências dos modelos anterior, que pode ser observada na Figura 27, devido essa tendência nos resultados de AUC pode considera que para esse caso os modelos não foram o fator mais impactante na melhora dos resultado AUC mas sim o balanceamento de classe quando

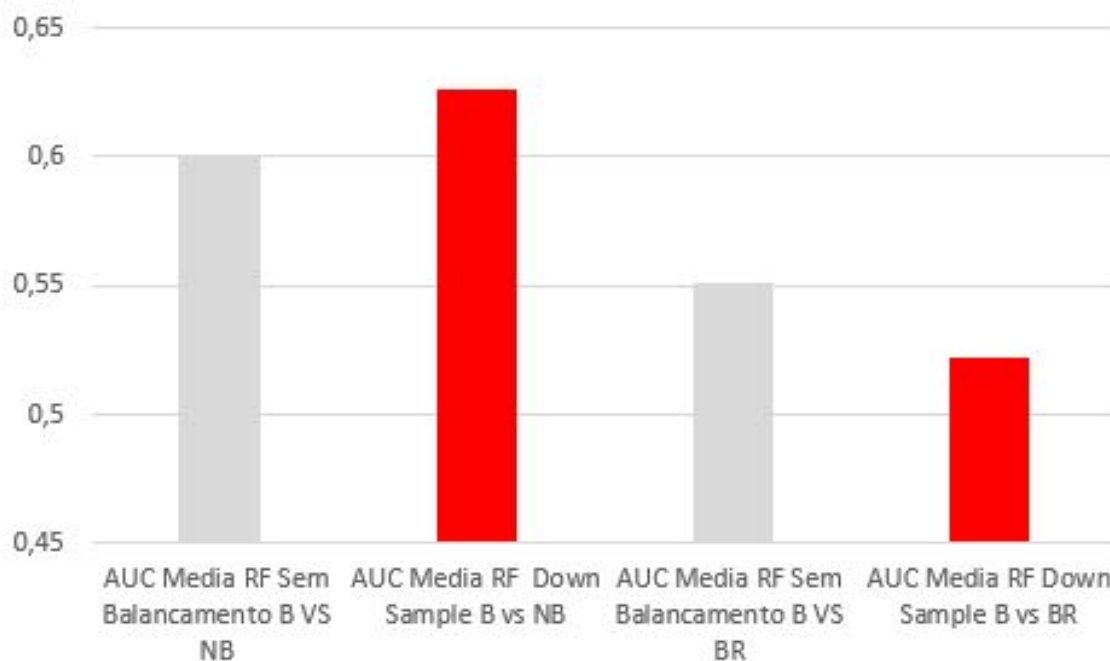


Figura 25: Gráfico AUC da média Random Florest down sample vs sem balanceamento.
Fonte: Dados do Autor.

consideramos biomarcadores vs não biomarcadores obtendo uma pequena melhora na AUC.

Quando empregou-se outro tipo de balanceamento de classe o *up sample* considerando a AUC por algoritmos, obteve outro conjunto de AUC médios.

Na utilização do algoritmo SVM com balanceamento *up sample*, obteve-se um ganho de AUC média com a combinação de biomarcadores vs não biomarcador mas com um ressalva que a sensibilidade dos modelos foi abaixo de 0,5 e obtendo uma especificidade acima de 0,70. Assim pode-se considerar que obteve um ganho com relação ao não balanceamento mais mesmo assim este tipo de balanceamento não auxilia no aumento da sensibilidade da classe positiva.

Com relação a combinação de biomarcador vs genes relacionados, obteve-se um ganho de AUC devido ao modelo ter sido a mais exemplo da classe positiva mesmos sendo exemplos sobre posto, mas também obteve uma sensibilidade abaixo de 0,50 sendo melhor que a do não balanceando por ser 0 mas mesmos assim não atingiu um valor considerável para um classificador.

Quando empregou-se o random forest com balanceamento de classe *up sample* observamos um fenômeno diferente observado até então, para esse modelo a sobre posição de amostra impactou negativamente para combinação biomarcadores vs não biomarcadores, acredita-se na ocorrência desse fato porque para validação separou-se 20% dos dados que foram balanceado com *down sample*, antes de realizar sobreposição dos dados de treinamento com intuito da não distorção dos dados de validação, buscando uma resultado

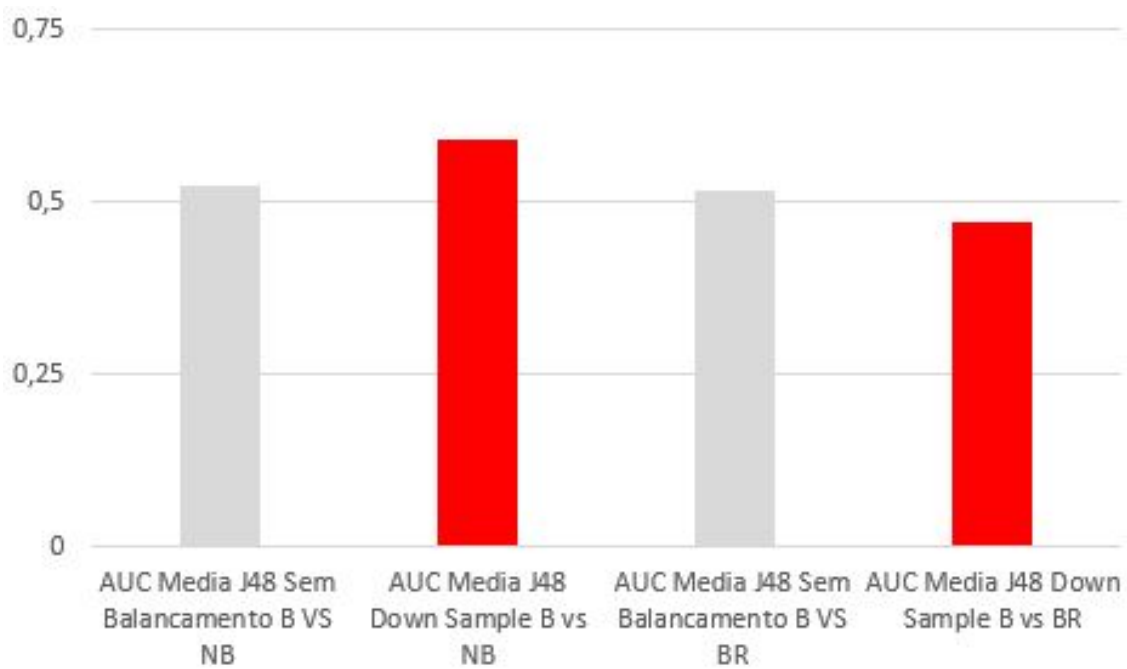


Figura 26: Gráfico AUC da média J48 down sample vs sem balanceamento. Fonte: Dados do Autor.

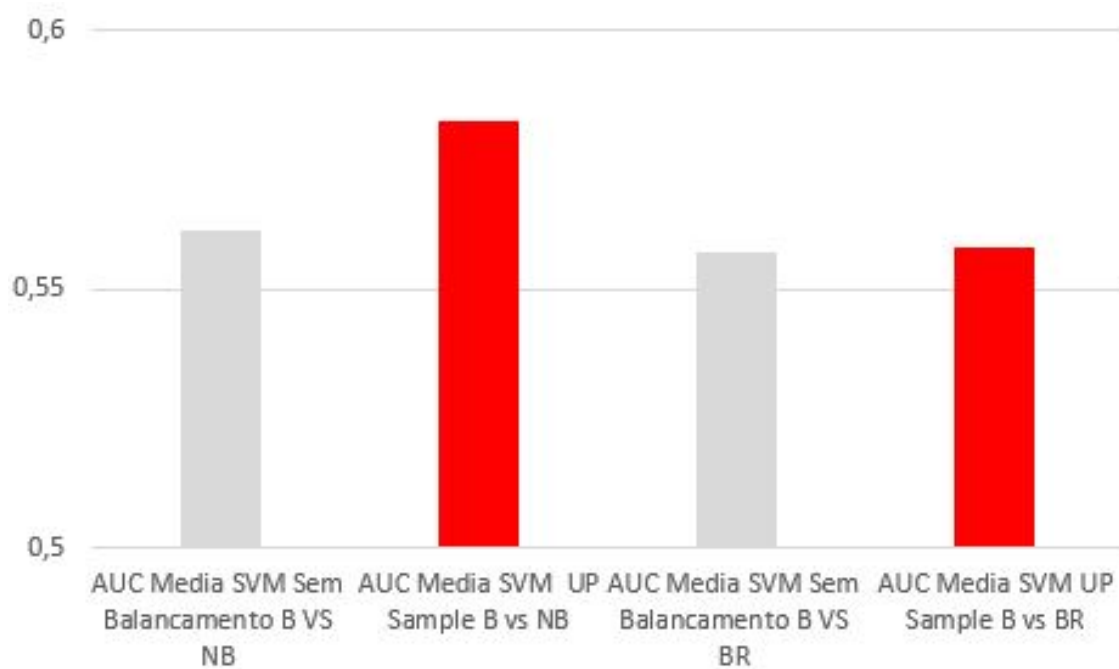


Figura 27: Gráfico AUC da média SVM up sample vs sem balanceamento. Fonte: Dados do Autor.

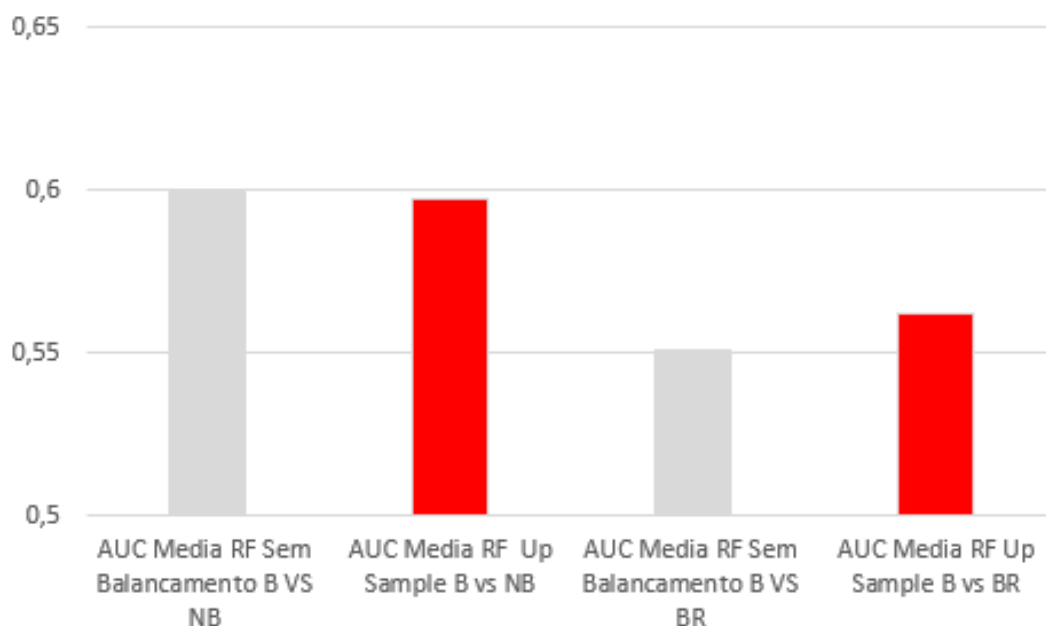


Figura 28: Gráfico AUC da média RF up sample vs sem balanceamento. Fonte: Dados do Autor.

próximo de uma amostra real.

Já o algoritmo J48 manteve os comportamento apresentado até então pelos outros modelos quando empregou-se balanceamento *down sample*, acredita-se que sobre posição de amostras para treinamento o *up sample*, não impactou tanto como no random forest quando observa-se o biomarcador vs não biomarcador, no qual J48 manteve um ganho AUC com relação ao não balanceamento, como pode ser observado na Figura ??, as árvores construídas para esse modelo não necessitaram de pode por terem sido árvores de poucos níveis como pode ser visualiza na Figura 30

Já quando observamos biomarcadores vs genes relacionas, acredita-se e perda no balanceamento de seja por aumentar um sensibilidade do modelos com relação ao não balanceamento fazendo com que a diferença entre sensibilidade e especificidade impacte na AUC do modelo.

Quando aplicou-se a técnica de balanceamento Smote para algoritmo SVM observou-se um ganho AUC com relação ao desbalanceamento considerando a combinados de dados biomarcadores vs não biomarcadores, esse ganho pode ser observados na Figura 31, mas quando observa-se o biomarcadores vs relacionado o modelo não obteve ganho de AUC, acredita-se seja pelo motivo anteriormente citado de semelhança topologia entre biomarcadores e relacionados.

No utilização do random forest com balanceamento de classe smote acredita-se, que balanceamento obteve um resultado inferior ao do conjunto de treinamento não balanceado, porque criou-se numero superior de instâncias artificias com relação as originais, foi seguida a orientação do autor da técnica de proporção máxima para criação de instâncias

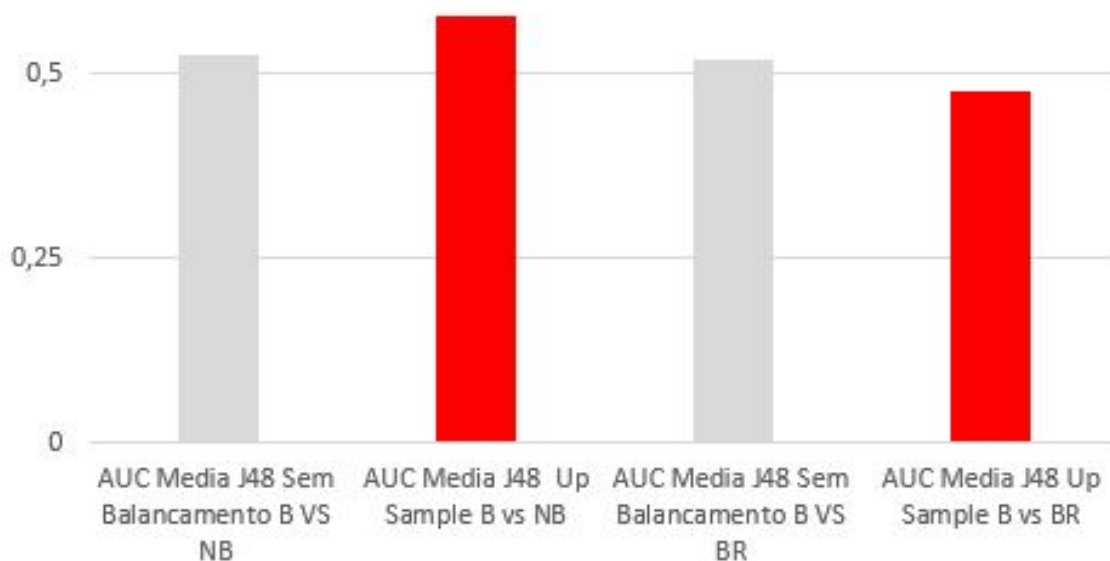


Figura 29: Gráfico AUC da média J48 up sample vs sem balanceamento. Fonte: Dados do Autor.

artificiais, possivelmente essa proporção precisa ser revista para buscar melhores resultados empregando esse balanceamento de classes.

Já utilização do balanceamento smote, contribui para ganhos AUC no algoritmo J48 não sofrendo com a possível distorção dos dados treinamento pelo numero superior de instâncias criadas artificialmente por este tipo de balanceamento, mesmo com esse ganhos de AUC não pode-se considerar esse classificador aplicavel para descoberta de biomarcadores por ele ter uma AUC muito baixa e um sensibilidade baixa para reconhecimento do padrão da classe positiva.

O melhor resultado obtido pode ser visualizado da seção 5.2 , empregando SVM com kernel radial, na qual a melhor área de curva ROC foi de 0,70 quando utilizou-se as medidas da rede doença, aplicando balanceamento de classe *down sample*. Acredita-se que esse resultado tenha influência do emprego da rede doença por não perder as características topológicas dos genes e pelo balanceamento de classe *down sample* evitar o super ajustamento do classificador.

O segundo melhor resultado foi empregando o algoritmo *Random Forest*, também foi utilizando as medidas da rede doença, mas sem balanceamento de classe, atingindo 0,66 de área da curva ROC. Esse resultado é o segundo melhor quando consideramos apenas questão da AUC, mas quando olhamos a sua sensibilidade, esse resultado é um classificador da classe negativa por ter uma sensibilidade igual 0, não sendo possível utilizá-lo para classificação de biomarcador.

Apesar desses classificadores, atingirem um área de curva ROC que os distingue de

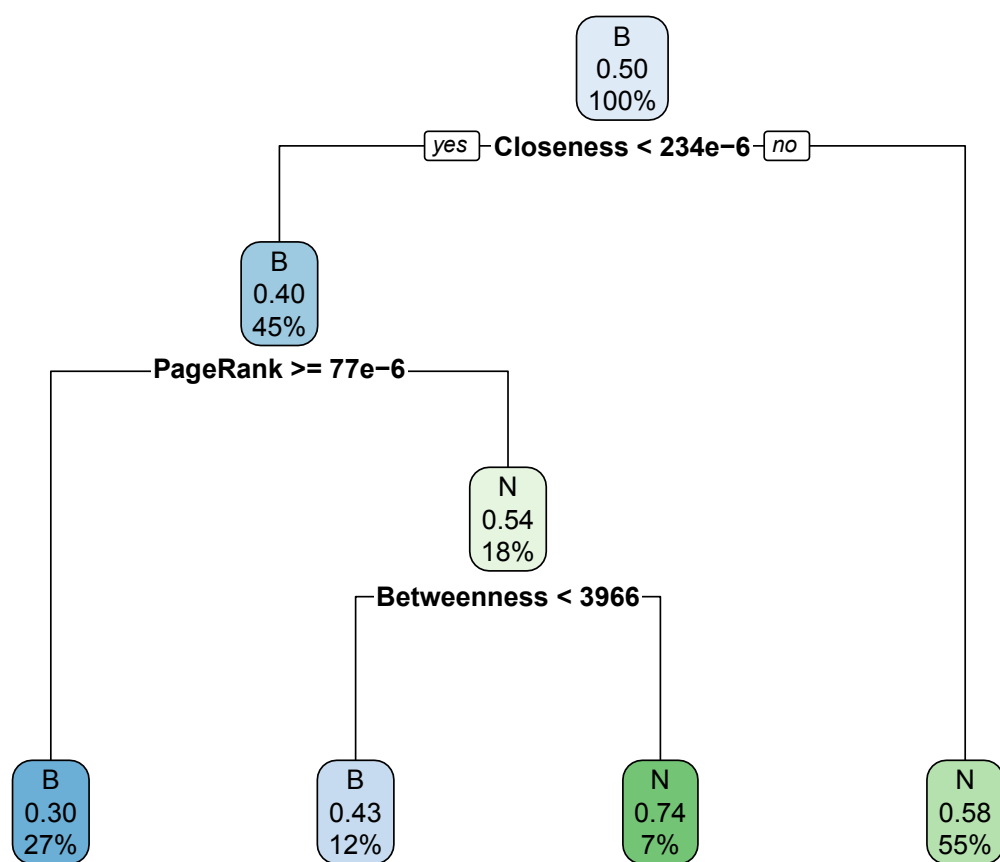


Figura 30: Árvores J48 up sample biomarcador vs não biomarcador. Fonte: Dados do Autor.

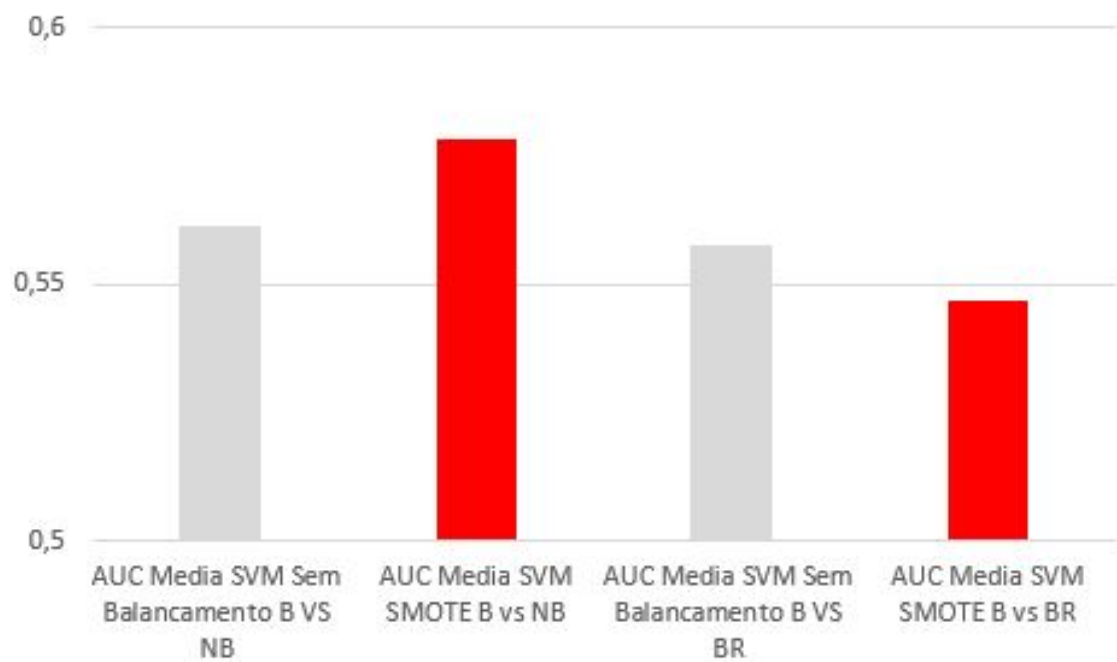


Figura 31: Gráfico AUC da média SVM Smote vs sem balanceamento. Fonte: Dados do Autor.

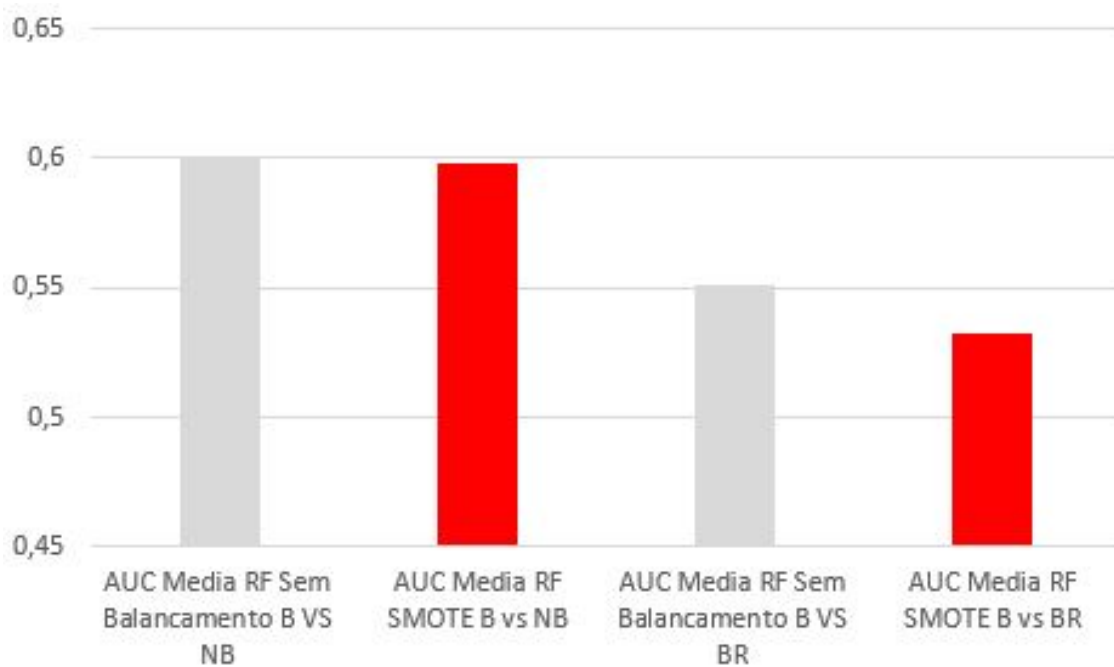


Figura 32: Gráfico AUC da média Random Forest Smote vs sem balanceamento. Fonte: Dados do Autor.

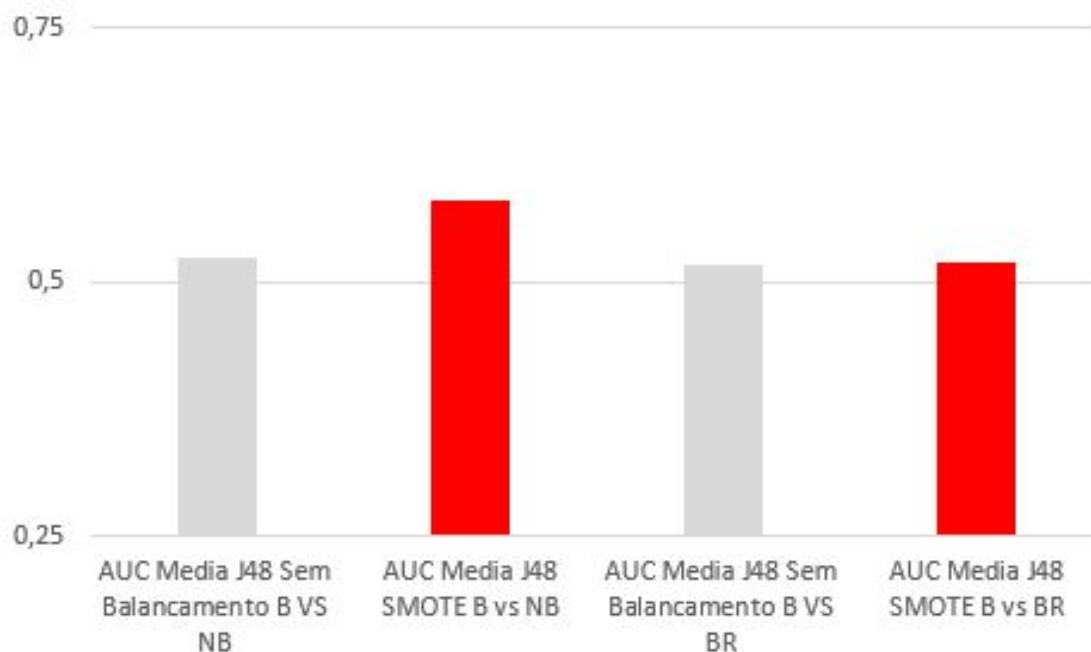


Figura 33: Gráfico AUC da média J48 Smote vs sem balanceamento. Fonte: Dados do Autor.

um classificador aleatório, a sua sensibilidade é muito baixa para classe positiva que foi considerada a dos biomarcadores. Sendo assim, pode-se dizer que obteve-se um classificador de não biomarcadores.

Com relação a Tabela 14, na qual são exibidos o resultados obtidos pelo *Random Forest*, é possível observar a mesma tendência do SVM, onde os melhores resultados foram aqueles obtidos empregando as medidas da rede doença e o balanceamento de classe *down sample*, chegando a 0,6553 de AUC.

O algoritmo J48 manteve o mesmo comportamento do SVM e do *Random Forest*, obtendo seu melhor resultado empregando o *dataset* das medidas da rede doença e com o balanceamento de classe *down sample*. O resultado de 0,6444 de área da curva ROC, e a sua segunda melhor área da curva foi com balanceamento de classe *smote*, atingindo 0,629, indicando dessa forma que o J48 não é um classificador aleatório mas ainda inferior ao SVM com 0,70 AUC.

Observando os resultados da área de curva ROC, obteve-se apenas um classificador acima de 0,70, quanto os demais ficaram muito abaixo do esperado. Idealmente buscou-se classificadores com uma AUC superior a 0,70, buscando atingir um valor mais próximo possível de 0,90. Acredita-se que um dos motivos para AUC inferiores a 0,70 seja o fato de não existir uma definição clara para genes que não são biomarcadores, impactando negativamente os resultados por não ter certeza dos falsos negativos.

Na comparação entre os balanceamento de classe utilizando média e desvio padrão, entre *down sample*, *smote* e *up sample*, pode-se visualizar na Figura 34 que não houve ganho na área de curva ROC dos classificadores. Esse fato pode ser justificado quando

observamos que os desvios padrão estão contidos entre si, demonstrando que não houve ganho de eficiência.

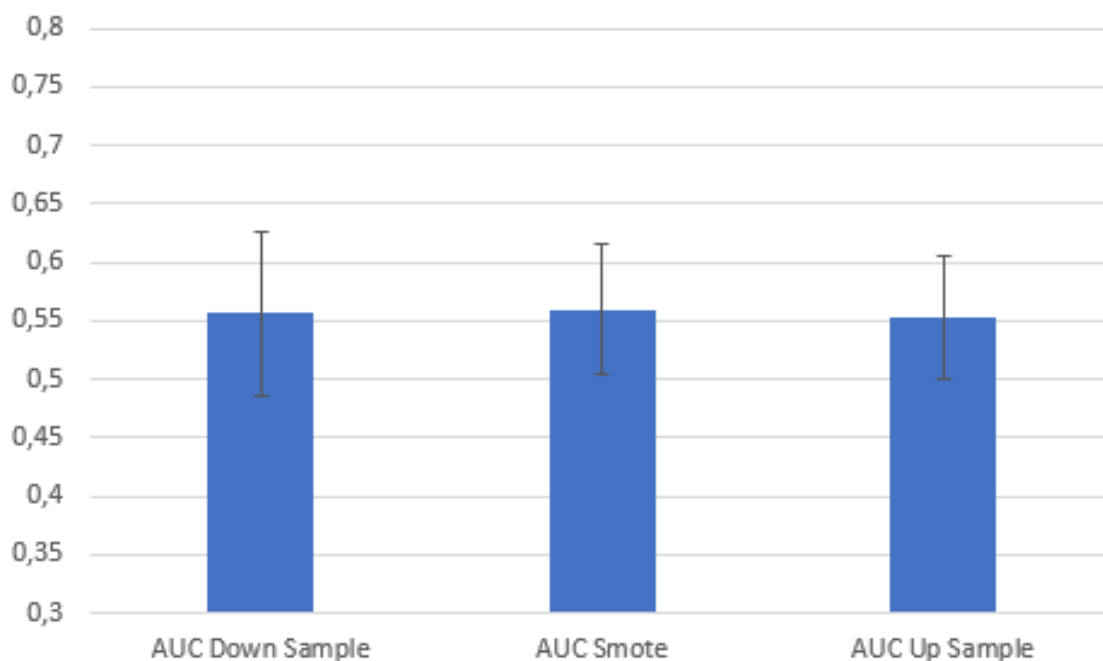


Figura 34: Gráfico da média e desvio padrão por balanceamento de classe. Fonte: Dados do Autor.

Quando comparados os algoritmos SVM, random forest e J48, empregando média e desvio padrão da área da curva ROC, na Figura 35, verifica-se que não há diferença de aplicação dos classificadores, porque os mesmos apresentam médias e desvios padrão muito próximos.

Os resultados obtidos empregando os *datasets* sem balanceamento e com balanceamento *up sample* de classe encontram-se na Figura 38.

As figuras 27 a 29 foram elaboradas para demonstrar um comparativo entre os resultados do balanceamento versus o não balanceamento de classes. Os resultados em alguns casos, mostram ganho de AUC quando não foi utilizado o não balanceamento de classes sobre o balanceamento, isso se observarmos apenas a AUC, mas é necessário avaliar ainda a sensibilidade resultante dos algoritmos. Por esse motivo esse ganho de AUC para os modelos treinados sem balanceamento torna-se irrelevante, porque esse modelo apenas classifica uma classe, a dos não biomarcadores.

Devido a esse motivo, buscou-se empregar o balanceamento de classe para contornar o problema de desequilíbrio, objetivando aumentar a sensibilidade dos modelos para classificação de biomarcadores.

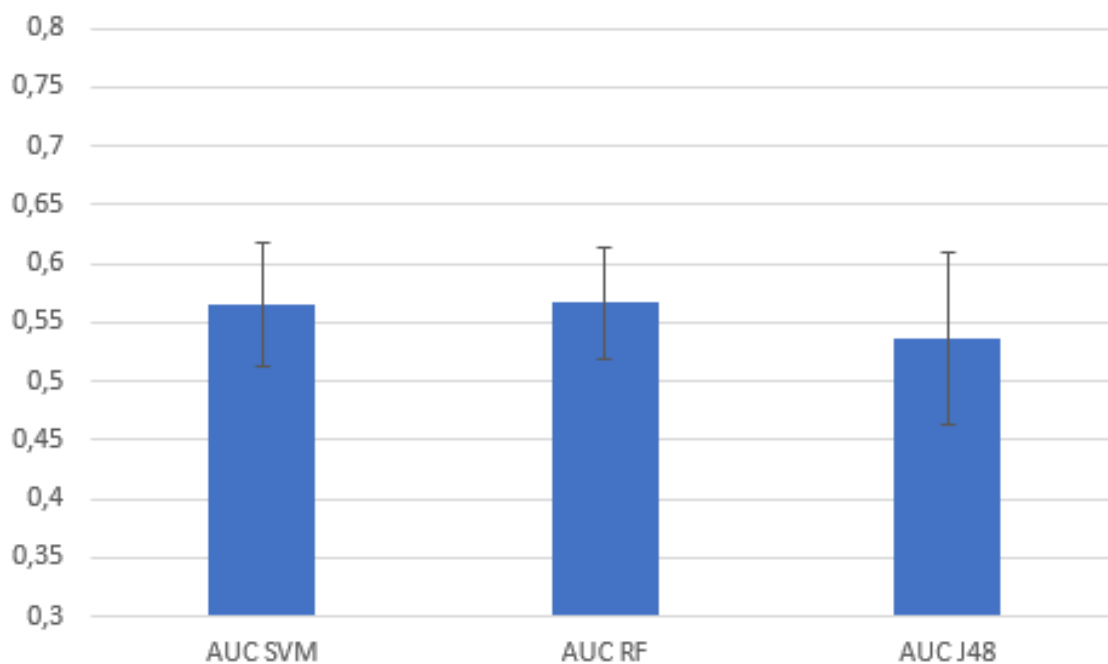


Figura 35: Gráfico da media e desvio padrão por algoritmo. Fonte: Dados do Autor.

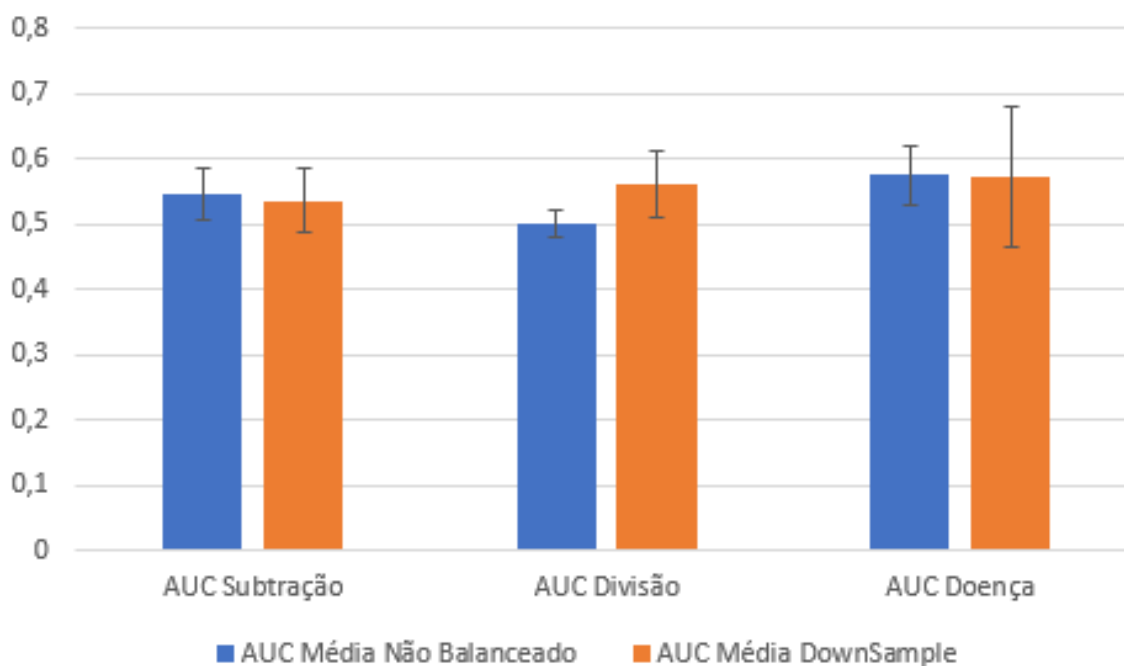


Figura 36: Gráfico da média AUC e desvio padrão sem balanceamento vs *down sample*. Fonte: Dados do Autor.

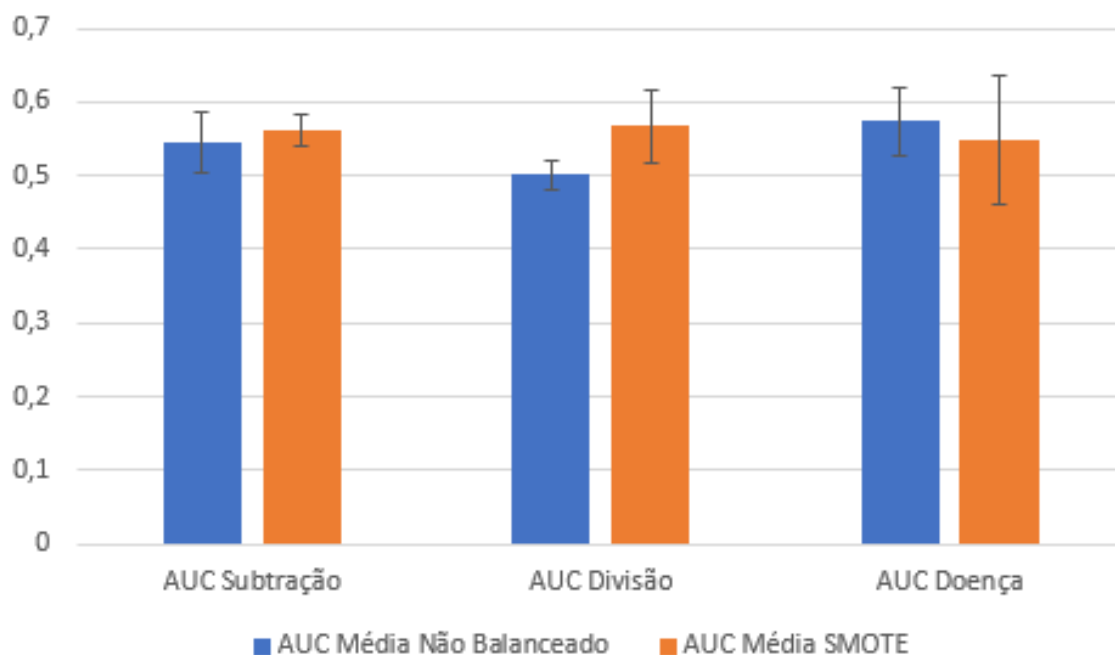


Figura 37: Gráfico da média AUC e desvio padrão sem balanceamento vs *SMOTE*. Fonte: Dados do Autor.

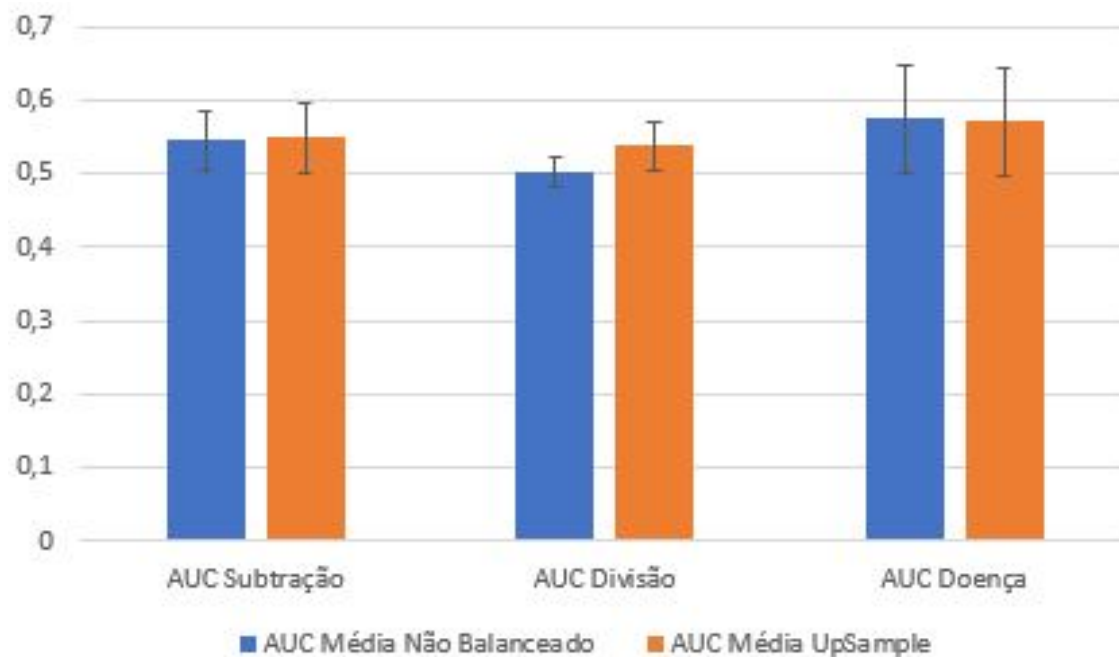


Figura 38: Gráfico da média AUC e desvio padrão sem balanceamento vs *up sample*. Fonte: Dados do Autor.

Considerações

A utilização de métricas de centralidade de redes, pode-se ser considerado como um proposta inovadora, não foi encontrado até o momento trabalhos que empregassem a metodologia proposta para descoberta de possíveis biomarcadores.

O melhor resultado obtido empregando a metodologia proposta foi o SVM com *down sample* atingindo um área da curva ROC 70%, já os outros modelos não passaram de 60% de área.

Quando considerou apenas a rede da doença e não a comparação entre as duas redes os classificadores apresentaram um melhora na sua classificação, isso provavelmente deve-se ao fato que mesmo com mudanças topológicas entre doença e controle, não torna possível a identificação de biomarcadores.

No teste considerando biomarcadores vs relacionados, foi possível visualizar uma sensibilidade a classe positiva que neste caso era os biomarcadores, isso reforça porque os genes relacionados ainda não são considerados biomarcadores por não caracterizar tão bem a condição da doença.

6 CONCLUSÃO

Apesar do aumento significativo em nossa capacidade de produzir evidências biológicas e dados experimentais, devido aos avanços tecnológicos e de inúmeros projetos sobre o genoma realizados na última década, ainda não conseguimos compreender plenamente o significado destas informações geradas, seja pelo ruído ou por dados faltantes dos conjuntos de dados gerados, juntamente com nossos conhecimentos parciais.

O entendimento dos processos biológicos envolvidos em doenças complexas pode levar a criação de novos tratamentos farmacológicos mais eficientes e terapias.

Estudar uma doença de forma estática e considerar apenas um fator, como uma mera expressão de genes ou um perfil, não deve ser o caminho a ser seguido, pelo simples fato que praticamente nenhuma interação biológica acontece de forma isolada em um organismo, sendo ela dependente de um grande número de interações com outras moléculas e sistemas envolvidos.

Dessa forma tanto a integração de dados como a análise de elementos topológicos numa rede, parecem ser fundamentais para que se obtenha êxito neste tipo de análise. Outro aspecto que justifica a integração é a própria natureza da aquisição dos dados biológicos, ainda passíveis de erros e limitações metodológicas (sobretudo na captação de ruído).

Trabalhar com redes no âmbito de doenças complexas é um grande desafio científico, uma vez que ela pode ser vital para descoberta de biomarcadores e pela questão da alta dimensionalidade dos dados.

Os resultados para esse estudo de caso não alcançaram os resultados esperados, uma possibilidade é questão das classe estarem desbalanceadas, e a sobre posição de não biomarcadores sobre os biomarcadores.

A avaliação dos dados e técnicas utilizada, permitiram concluir que o problema é de difícil discriminação e mesmos técnicas avanças de *machine learning* podendo não ser eficientes para separar as classes.

Como trabalhos futuros, serão realizados estudo empregando outros modelos de aprendizagem de máquina como Redes Neurais Profundas, modelos de semi-supervisionado, classificadores de uma classe buscando mitigar o problema dos falsos

negativos e como também outros *datasets*.

Este trabalho foi apresentado no MPU 2018 e no *Worshop* do Projeto de Medicina Personalizada, convenio de Universidade Federal do Rio Grande do Sul e a Universidade Federal do Rio Grande

REFERÊNCIAS

ANDRE, R.; SCAHILL, R. I.; HAIDER, S.; TABRIZI, S. J. Biomarker development for Huntington's disease. **Drug discovery today**, [S.l.], v.19, n.7, p.972–979, 2014.

AYDAY, E.; DE CRISTOFARO, E.; HUBAUX, J.-P.; TSUDIK, G. Whole genome sequencing: Revolutionary medicine or privacy nightmare? **Computer**, [S.l.], v.48, n.2, p.58–66, 2015.

BARABÁSI, A.-L. Scale-free networks: a decade and beyond. **science**, [S.l.], v.325, n.5939, p.412–413, 2009.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, [S.l.], v.286, n.5439, p.509–512, 1999.

BARABÁSI, A.-L.; GULBAHCE, N.; LOSCALZO, J. Network medicine: a network-based approach to human disease. **Nature Reviews Genetics**, [S.l.], v.12, n.1, p.56–68, 2011.

BARRAT, A.; BARTHELEMY, M.; VESPIGNANI, A. The Architecture of Complex Weighted Networks: Measurements and Models. In: **Large Scale Structure And Dynamics Of Complex Networks: From Information Technology to Finance and Natural Science**. [S.l.]: World Scientific, 2007. p.67–92.

BARRETT, T.; TROUP, D. B.; WILHITE, S. E.; LEDOUX, P.; RUDNEV, D.; EVANGELISTA, C.; KIM, I. F.; SOBOLEVA, A.; TOMASHEVSKY, M.; MARSHALL, K. A. et al. NCBI GEO: archive for high-throughput functional genomic data. **Nucleic acids research**, [S.l.], v.37, n.suppl_1, p.D885–D890, 2008.

BARRY, W. T.; NOBEL, A. B.; WRIGHT, F. A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. **Bioinformatics**, [S.l.], v.21, n.9, p.1943–1949, 2005.

BUCHANAN, M. **Nexus: small worlds and the groundbreaking theory of networks**. [S.l.]: WW Norton & Company, 2003.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, [S.l.], v.16, p.321–357, 2002.

CHEN, W.-H.; ZHAO, X.-M.; NOORT, V. van; BORK, P. Human monogenic disease genes have frequently functionally redundant paralogs. **PLoS computational biology**, [S.l.], v.9, n.5, p.e1003073, 2013.

CHU, J.-h.; LAZARUS, R.; CAREY, V. J.; RABY, B. A. Quantifying differential gene connectivity between disease states for objective identification of disease-relevant genes. **BMC systems biology**, [S.l.], v.5, n.1, p.89, 2011.

CLAESSEN, G.; SCHNELL, F.; BOGAERT, J.; CLAEYS, M.; PATTYN, N.; DE BUCK, F.; DYMARKOWSKI, S.; CLAUS, P.; CARRÉ, F.; VAN CLEEMPUT, J. et al. Exercise cardiac magnetic resonance to differentiate athlete's heart from structural heart disease. **European Heart Journal-Cardiovascular Imaging**, [S.l.], 2018.

COLOMBO, J.; RAHAL, P. A tecnologia de microarray no estudo do câncer de cabeça e pescoço. **Revista Brasileira de Biociências**, [S.l.], p.64–72, 2010.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. **InterJournal, Complex Systems**, [S.l.], v.1695, n.5, p.1–9, 2006.

DE LIZ SOUZA, L.; AUGUSTO RHODEN, S.; ALENCAR PAMPHILE, J. A IMPORTÂNCIA DAS ÔMICAS COMO FERRAMENTAS PARA O ESTUDO DA PROSPECÇÃO DE MICRORGANISMOS: PERSPECTIVAS E DESAFIOS. **Uningá Review**, [S.l.], v.18, n.2, 2014.

DERISI, J. L.; IYER, V. R.; BROWN, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. **Science**, [S.l.], v.278, n.5338, p.680–686, 1997.

DØRUM, G.; SNIPEN, L.; SOLHEIM, M.; SÆBØ, S. Rotation testing in gene set enrichment analysis for small direct comparison experiments. **Statistical applications in genetics and molecular biology**, [S.l.], v.8, n.1, p.1–24, 2009.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. In: ACM SIGCOMM COMPUTER COMMUNICATION REVIEW, 1999. **Anais...** [S.l.: s.n.], 1999. v.29, n.4, p.251–262.

FUENTE, A. De la. From ‘differential expression’ to ‘differential networking’ – identification of dysfunctional regulatory networks in diseases. **Trends in genetics**, [S.l.], v.26, n.7, p.326–333, 2010.

GAMBARDELLA, G.; MORETTI, M. N.; DE CEGLI, R.; CARDONE, L.; PERON, A.; DI BERNARDO, D. Differential network analysis for the identification of condition-specific pathway activity and regulation. **Bioinformatics**, [S.l.], v.29, n.14, p.1776–1785, 2013.

HANNENHALLI, S.; PUTT, M. E.; GILMORE, J. M.; WANG, J.; PARMACEK, M. S.; EPSTEIN, J. A.; MORRISEY, E. E.; MARGULIES, K. B.; CAPPOLA, T. P. Transcriptional genomics associates FOX transcription factors with human heart failure. **Circulation**, [S.l.], v.114, n.12, p.1269–1276, 2006.

HÄNZELMANN, S.; CASTELO, R.; GUINNEY, J. GSVA: gene set variation analysis for microarray and RNA-seq data. **BMC bioinformatics**, [S.l.], v.14, n.1, p.7, 2013.

HERSHBERGER, R. E.; HEDGES, D. J.; MORALES, A. Dilated cardiomyopathy: the complexity of a diverse genetic architecture. **Nature reviews cardiology**, [S.l.], v.10, n.9, p.531, 2013.

IRIZARRY, R. A.; WANG, C.; ZHOU, Y.; SPEED, T. P. Gene set enrichment analysis made simple. **Statistical methods in medical research**, [S.l.], v.18, n.6, p.565–575, 2009.

JIANG, Z.; GENTLEMAN, R. Extensions to gene set enrichment. **Bioinformatics**, [S.l.], v.23, n.3, p.306–313, 2006.

KHERA, A. V.; KATHIRESAN, S. Genetics of coronary artery disease: discovery, biology and clinical translation. **Nature Reviews Genetics**, [S.l.], 2017.

KUHN, M. Caret: classification and regression training. **Astrophysics Source Code Library**, [S.l.], 2015.

LAI, Y.; WU, B.; CHEN, L.; ZHAO, H. A statistical method for identifying differential gene–gene co-expression patterns. **Bioinformatics**, [S.l.], v.20, n.17, p.3146–3155, 2004.

LANGFELDER, P.; HORVATH, S. WGCNA: an R package for weighted correlation network analysis. **BMC bioinformatics**, [S.l.], v.9, n.1, p.559, 2008.

LOPES, M. V. d. S. et al. Aplicação de classificadores para determinação de conformidade de biodiesel. , [S.l.], 2017.

MACKERETH, C. D.; SATTLER, M. Dynamics in multi-domain protein recognition of RNA. **Current opinion in structural biology**, [S.l.], v.22, n.3, p.287–296, 2012.

MCQUEEN, C.; DINDOT, S.; FOSTER, M.; COHEN, N. Genetic susceptibility to *Rhodococcus equi*. **Journal of veterinary internal medicine**, [S.l.], v.29, n.6, p.1648–1659, 2015.

MILO, R.; SHEN-ORR, S.; ITZKOVITZ, S.; KASHTAN, N.; CHKLOVSKII, D.; ALON, U. Network motifs: simple building blocks of complex networks. **Science**, [S.l.], v.298, n.5594, p.824–827, 2002.

MOODY, J. Race, school integration, and friendship segregation in America. **American journal of Sociology**, [S.l.], v.107, n.3, p.679–716, 2001.

MORENO, J. L.; JENNINGS, H. H. et al. Who shall survive? , [S.l.], 1934.

NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. **The structure and dynamics of networks**. [S.l.]: Princeton University Press, 2011. v.19.

OLIVEIRA, L. S. d. et al. **Método de identificação de genes taxonomicamente restritos em dados de RNA-seq em organismo não modelo**. 2015. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Pará.

OUGHTRED, R.; CHATR-ARYAMONTRI, A.; BREITKREUTZ, B.-J.; CHANG, C. S.; RUST, J. M.; THEESFELD, C. L.; HEINICKE, S.; BREITKREUTZ, A.; CHEN, D.; HIRSCHMAN, J. et al. BioGRID: a resource for studying biological interactions in yeast. **Cold Spring Harbor Protocols**, [S.l.], v.2016, n.1, p.pdb–top080754, 2016.

PARKINSON, H.; KAPUSHESKY, M.; KOLESNIKOV, N.; RUSTICI, G.; SHOJA-TALAB, M.; ABEYGUNAWARDENA, N.; BERUBE, H.; DYLAG, M.; EMAM, I.; FARNE, A. et al. ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. **Nucleic acids research**, [S.l.], v.37, n.suppl_1, p.D868–D872, 2008.

PASSARGE, E. **Genética texto y atlas**. [S.l.]: Ed. Médica Panamericana, 2009.

PAVLOPOULOS, G. A.; SECRIER, M.; MOSCHOPOULOS, C. N.; SOLDATOS, T. G.; KOSSIDA, S.; AERTS, J.; SCHNEIDER, R.; BAGOS, P. G. Using graph theory to analyze biological networks. **BioData mining**, [S.l.], v.4, n.1, p.10, 2011.

PINHO, J. R. R. Precision Medicine. **Einstein (São Paulo)**, [S.l.], v.15, n.1, p.VII–X, 2017.

PRATI, R.; BATISTA, G.; MONARD, M. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, [S.l.], v.6, n.2, p.215–222, 2008.

PROULX, S. R.; PROMISLOW, D. E.; PHILLIPS, P. C. Network thinking in ecology and evolution. **Trends in ecology & evolution**, [S.l.], v.20, n.6, p.345–353, 2005.

PROVOST, F. Machine learning from imbalanced data sets 101. In: AAAI’2000 WORKSHOP ON IMBALANCED DATA SETS, 2000. **Proceedings...** [S.l.: s.n.], 2000. v.68, p.1–3.

RIJN, M. Van de; GILKS, C. Applications of microarrays to histopathology. **Histopathology**, [S.l.], v.44, n.2, p.97–108, 2004.

ROBINSON, M. D.; OSHLACK, A. A scaling normalization method for differential expression analysis of RNA-seq data. **Genome biology**, [S.l.], v.11, n.3, p.R25, 2010.

ROY, S.; BHATTACHARYYA, D. K.; KALITA, J. K. Reconstruction of gene co-expression network from microarray data using local expression patterns. **BMC bioinformatics**, [S.l.], v.15, n.Suppl 7, p.S10, 2014.

RUFINO, H. L. P.; VEIGA, A. C. P.; NAKAMOTO, P. T. SOMOTE_EASY: AN ALGORITHM TO TREAT THE CLASSIFICATION ISSUE IN REAL DATABASES. **JISTEM-Journal of Information Systems and Technology Management**, [S.l.], v.13, n.1, p.61–80, 2016.

RUSSO, G.; ZEGAR, C.; GIORDANO, A. Advantages and limitations of microarray technology in human cancer. **Oncogene**, [S.l.], v.22, n.42, p.6497–6507, 2003.

SILVA, F. A. B. da. Big Data e Nuvens Computacionais: Aplicações em Saúde Pública e Genômica. **Journal of health Informatics**, [S.l.], v.8, n.2, 2016.

SILVA, T. C.; ZHAO, L. **Machine learning in complex networks**. [S.l.]: Springer, 2016. v.1.

SMYTH, G. et al. **Limma**: linear models for microarray data. Bioinformatics and computational biology solutions using R and bioconductor. Edited by: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. 2005. [S.l.]: New York: Springer, 2005.

SPORNS, O. **Networks of the Brain**. [S.l.]: MIT press, 2010.

SUBRAMANIAN, A.; TAMAYO, P.; MOOTHA, V. K.; MUKHERJEE, S.; EBERT, B. L.; GILLETTE, M. A.; PAULOVICH, A.; POMEROY, S. L.; GOLUB, T. R.; LANDER, E. S. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences**, [S.l.], v.102, n.43, p.15545–15550, 2005.

TAMMINGA, C. A.; PEARLSON, G. D.; STAN, A. D.; GIBBONS, R. D.; PADMANABHAN, J.; KESHAVAN, M.; CLEMENTZ, B. A. Strategies for advancing disease definition using biomarkers and genetics: The Bipolar and Schizophrenia Network for Intermediate Phenotypes. **Biological Psychiatry: Cognitive Neuroscience and Neuroimaging**, [S.l.], v.2, n.1, p.20–27, 2017.

TARCA, A. L.; DRAGHICI, S.; KHATRI, P.; HASSAN, S. S.; MITTAL, P.; KIM, J.-s.; KIM, C. J.; KUSANOVIC, J. P.; ROMERO, R. A novel signaling pathway impact analysis. **Bioinformatics**, [S.l.], v.25, n.1, p.75–82, 2008.

TAYLOR, I. W.; LINDING, R.; WARDE-FARLEY, D.; LIU, Y.; PESQUITA, C.; FARIA, D.; BULL, S.; PAWSON, T.; MORRIS, Q.; WRANA, J. L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. **Nature biotechnology**, [S.l.], v.27, n.2, p.199–204, 2009.

TIAN, L.; GREENBERG, S. A.; KONG, S. W.; ALTSCHULER, J.; KOHANE, I. S.; PARK, P. J. Discovering statistically significant pathways in expression profiling studies. **Proceedings of the National Academy of Sciences of the United States of America**, [S.l.], v.102, n.38, p.13544–13549, 2005.

TUSHER, V. G.; TIBSHIRANI, R.; CHU, G. Significance analysis of microarrays applied to the ionizing radiation response. **Proceedings of the National Academy of Sciences**, [S.l.], v.98, n.9, p.5116–5121, 2001.

VASKE, C. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 2010. , [S.l.], 2013.

WANG, L.-x.; LI, Y.; CHEN, G.-z. Network-based co-expression analysis for exploring the potential diagnostic biomarkers of metastatic melanoma. **PloS one**, [S.l.], v.13, n.1, p.e0190447, 2018.

WANG, Y.; LIU, Z.-P. Identifying biomarkers of diabetes with gene coexpression networks. In: CHINESE AUTOMATION CONGRESS (CAC), 2017, 2017. **Anais...** [S.l.: s.n.], 2017. p.5283–5286.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. **nature**, [S.l.], v.393, n.6684, p.440, 1998.

WHO et al. **International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)-2015-WHO Version for; 2015.**

WU, D.; LIM, E.; VAILLANT, F.; ASSELIN-LABAT, M.-L.; VISVADER, J. E.; SMYTH, G. K. ROAST: rotation gene set tests for complex microarray experiments. **Bioinformatics**, [S.l.], v.26, n.17, p.2176–2182, 2010.

ZAHA, A.; FERREIRA, H. B.; PASSAGLIA, L. M. **Biologia Molecular Básica-5**. [S.l.]: Artmed Editora, 2014.

ZHANG, A. **Protein interaction networks: computational analysis**. [S.l.]: Cambridge University Press, 2009.

ZHANG, J.; XIANG, Y.; DING, L.; BORLAWSKY, T. B.; OZER, H. G.; JIN, R.; PAYNE, P.; HUANG, K. et al. Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. In: BMC BIOINFORMATICS, 2010. **Anais...** [S.l.: s.n.], 2010. v.11, n.9, p.S5.

ZHANG, S.; LIU, W.; LIU, X.; QI, J.; DENG, C. Biomarkers identification for acute myocardial infarction detection via weighted gene co-expression network analysis. **Medicine**, [S.l.], v.96, n.47, 2017.

ZHAO, X.-M.; QIN, G. Identifying biomarkers with differential analysis. In: **Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases**. [S.l.]: Springer, 2013. p.17–31.