



UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM
INSTITUTO DE COMPUTAÇÃO - ICOMP
PROGRAMA PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

Um método de estimação de expressões gênicas de câncer de mama com base em correlação

Beatriz Albuquerque Rodrigues

Manaus - AM

Março de 2024

Beatriz Albuquerque Rodrigues

Um método de estimação de expressões gênicas de câncer
de mama com base em correlação

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas, Campus Universitário Senador Arthur Virgílio Filho, como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador(a)

Fabíola Guerra Nakamura, Dra.

Universidade Federal do Amazonas - UFAM

Instituto de Computação - IComp

Manaus - AM

Março de 2024

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

R696m Rodrigues, Beatriz Albuquerque
Um método de estimação de expressões gênicas de câncer de
mama com base em correlação / Beatriz Albuquerque Rodrigues. -
2024.
90 f. : il., color. ; 31 cm.

Orientador(a): Fabíola Guerra Nakamura.
Dissertação (mestrado) - Universidade Federal do Amazonas,
Programa de Pós-Graduação em Informática, Universidade Federal
do Amazonas, 2024.

1. Expressão gênica. 2. Valores ausentes. 3. Estimação. 4.
Câncer de Mama. I. Nakamura, Fabíola Guerra. II. Universidade
Federal do Amazonas. Programa de Pós-Graduação em
Informática. III. Título



Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"UM MÉTODO DE ESTIMAÇÃO DE EXPRESSÕES GÊNICAS DE CÂNCER DE MAMA COM BASE EM CORRELAÇÃO"

BEATRIZ ALBUQUERQUE RODRIGUES

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA PELA BANCA EXAMINADORA CONSTITUÍDA PELOS PROFESSORES:

Profa. Dra. Fabíola Guerra Nakamura - PRESIDENTE

Prof. Dr. Vanderson de Souza Sampaio - MEMBRO EXTERNO

Prof. Dr. Juan Gabriel Colonna - INTERNO

Prof. Dr. André Luiz da Costa Carvalho - INTERNO

MANAUS, 05 DE ABRIL DE 2024.



Documento assinado eletronicamente por **Fabíola Guerra Nakamura**, **Professor do Magistério Superior**, em 03/06/2024, às 12:02, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Juan Gabriel Colonna, Professor do Magistério Superior**, em 26/06/2024, às 17:44, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **André Luiz da Costa Carvalho, Professor do Magistério Superior**, em 03/07/2024, às 17:01, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Maria do Perpétuo Socorro Vasconcelos Palheta, Secretária em exercício**, em 03/07/2024, às 18:05, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Vanderson de Souza Sampaio, Usuário Externo**, em 04/07/2024, às 11:11, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2057388** e o código CRC **463AE103**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário
Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Referência: Processo nº 23105.014554/2024-05

SEI nº 2057388

Dedico este trabalho aos meus pais, Maria e Cláudio.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus e todas as vezes que ele me direcionou a tomar a escolha certa.

Aos meus pais, Aparecida e Cláudio, que sempre fizeram de tudo para que eu tivesse a melhor educação possível e acreditaram em cada um dos meus sonhos. Ao longo desta jornada, eles estiveram ao meu lado nos momentos mais íntimos e introspectivos nos quais pensei que não iria conseguir me dizendo para olhar sempre em frente e para mirar nos meus objetivos. Cujo amor, carinho e atenção me permitiram voltar para rota.

À minha irmã e melhor amiga, Ana, que me suportou nos meus momentos mais complicados, enxugou minhas lágrimas e me fez rir insanamente depois. Sem ela eu não seria metade do que sou.

À minha orientadora, Fabíola Nakamura, por ter aceitado me orientar sem ao menos conhecer ao certo a minha trajetória. Ela me deu a oportunidade de mostrar o que eu era capaz de fazer na área de computação e me permitiu aprender bem mais do que imaginei que iria. Escolhi trabalhar com ela um tema, na época, um pouco fora da minha curva de domínio e sinto que tomei a decisão correta.

Aos meus amigos Arthur, Carlla, Mirna e Kamilla com os quais muitas vezes falei o quanto estava sendo difícil esse processo e que me ajudaram a compreender que os dias difíceis também fazem parte da caminhada e do nosso crescimento pessoal. Além disso, me fizeram refletir que, de vez enquanto, parar e respirar nos ajuda a voltar mais fortes.

Aos meus amigos Guilherme, Gabriel, Shida e Victor com os quais conversei

muitas vezes sobre meu assunto de pesquisa e me ajudaram a ter boas ideias do que fazer/melhorar. E a todos meus familiares, amigos e colegas, em geral, que estiveram presentes nesta fase tão importante da minha vida.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD.

Devemos escrever para nós mesmos, é assim que podemos chegar aos outros.

Eugène Ionesco

Um método de estimação de expressões gênicas de câncer de mama com base em correlação

Autor: Beatriz Albuquerque Rodrigues

Orientador: Fabíola Guerra Nakamura, Dra.

Resumo

Os dados de expressão de genes extraídas por meio da tecnologia de microarrays geralmente sofrem de problemas de valor perdido devido a problemas como poeira ou arranhões no slide. Em bases de dados de câncer de mama, a análise subsequente e a classificação de subtipos podem sofrer fortemente com a presença de dados omissos, sendo assim é primordial tratar esse problema. Várias abordagens para estimação desses valores em dados de expressão gênica foram desenvolvidas, mas a tarefa é difícil devido a fatores como a existência ou não de uma estrutura de correlação nos dados, à alta dimensionalidade dos dados (número de genes x número de amostras), e aos padrões e mecanismos de ausência. Nesta pesquisa, desenvolvemos um método, para estimar valores ausentes em expressões gênicas de câncer de mama, que lida com a alta dimensionalidade dos dados por meio do uso de um conjunto de genes que melhor caracterizam o câncer de mama e que podem contribuir para uma classificação mais precisa de subtipos.

Palavras-chave: Expressões gênicas, Valores ausentes, Estimação, Câncer de Mama.

Um método de estimação de expressões gênicas de câncer de mama com base em correlação

Autor: Beatriz Albuquerque Rodrigues

Orientador: Fabíola Guerra Nakamura, Dra.

Abstract

Gene expression data extracted through microarray technology often suffer from lost value issues due to problems such as dust or scratches on the slide. In breast cancer databases, the subsequent analysis and subtype classification can suffer greatly from missing data, so it is essential to address this problem. Several approaches to estimate these values in gene expression data have been developed, but the task is difficult due to factors such as the existence or not of a correlation structure in the data, the high dimensionality of the data (number of genes x number of samples), and the patterns and mechanisms of absence. In this research, we developed a method to estimate missing values in gene expressions of breast cancer, that deals with the high dimensionality of the data through the use of a set of genes that best characterize breast cancer and can contribute to a more accurate classification of subtypes.

Keywords: Gene expressions, Missing values, Estimation, Breast cancer.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de conjunto de dados, figura adaptada de de Ayyad et al. (2019).	26
Figura 2 – Estimação utilizando algoritmo supervisionado, Figura adaptada de Castaneda et al., (2009). Os valores iguais a 1 representam dados não ausentes e os valores iguais a 0 representam os dados ausentes. . . .	37
Figura 3 – Método de estimação de expressões gênicas baseado em correlação (EEGBC).	52
Figura 4 – Divisão do conjunto global de genes em dois conjuntos: um contendo os genes com dados faltantes e outro contendo os genes completos. Sendo assim, valores iguais a 1 representam dados não ausentes e os valores iguais a 0 representam os dados ausentes.	53
Figura 5 – Diagrama dos métodos RFEn PAM50 e RFAn PAM50.	57
Figura 6 – Diagrama dos métodos RFECa PAM50 e RFACa PAM50.	57
Figura 7 – Processo de classificação.	58
Figura 8 – Resultados (a) RMSE e (b) R2 para estimação dos genes ausentes do suptipo Basal dado cada um dos métodos.	62
Figura 9 – Resultados (a) RMSE e (b) R2 para estimação dos genes ausentes do suptipo Her2 dado cada um dos métodos.	63
Figura 10 – Resultados (a) RMSE e (b) R2 para estimação dos genes ausentes do suptipo LumA dado cada um dos métodos.	63
Figura 11 – Resultados (a) RMSE e (b) R2 para estimação dos genes ausentes do suptipo LumB dado cada um dos métodos.	64

Figura 12 – Genes completos escolhidos em cada um dos métodos de estimação por gene incompleto no subtipo Basal. Visão de interseção na qual as cores azuis representam os genes selecionados.	65
Figura 13 – Genes completos escolhidos em cada um dos genes estimados dado o método RFECA PAM50 no subtipo Basal. Visão de interseção na qual as cores azuis representam os genes selecionados.	65
Figura 14 – Genes completos escolhidos em cada um dos métodos de estimação por gene faltante no subtipo Her2. Visão de interseção na qual as cores azuis representam os genes selecionados.	66
Figura 15 – Genes completos escolhidos em cada um dos genes estimados dado o método RFECA PAM50 no subtipo Her2. Visão de interseção na qual as cores azuis representam os genes selecionados.	67
Figura 17 – Genes completos escolhidos em cada um dos genes estimados dado o método RFECA PAM50 no subtipo LumA. Visão de interseção na qual as cores azuis representam os genes selecionados.	67
Figura 16 – Genes completos escolhidos em cada um dos métodos de estimação por gene faltante no subtipo LumA. Visão de interseção na qual as cores azuis representam os genes selecionados.	68
Figura 18 – Genes completos escolhidos em cada um dos métodos de estimação por gene faltante no subtipo LumB. Visão de interseção na qual as cores azuis representam os genes selecionados.	69
Figura 19 – Genes completos escolhidos em cada um dos genes estimados dado o método RFECA PAM50 no subtipo LumB. Visão de interseção na qual as cores azuis representam os genes selecionados.	70

LISTA DE TABELAS

Tabela 1 – Valores de força e direção referente ao coeficiente de correlação de Pearson (r).	45
Tabela 2 – Resumo dos métodos clássicos de Machine Learning.	49
Tabela 3 – Descrição do Conjunto de Dados.	55
Tabela 4 – Descrição da quantidade de genes com e sem valores ausentes na base Cptac 2C para os genes do PAM50 para cada subtipo de câncer de mama.	59
Tabela 5 – Descrição da quantidade expressões de genes com e sem valores ausentes na base Cptac 2C para os genes do PAM50 para cada subtipo de câncer de mama.	60
Tabela 6 – Resultados da estimação usando as métricas RMSE e R2 em comparação com os Baselines.	61
Tabela 7 – Resultados da classificação feita a partir das bases geradas pelos métodos de estimação, usando as métricas precisão, recall, acurácia e F_1 . Análise observando a validação cruzada k-fold com 5 e com 10 folds.	73
Tabela I – Descrição funcional dos genes do PAM50 diretamente relacionados ao câncer de mama.	80
Tabela II – Descrição dos genes do PAM50 não diretamente relacionados ao câncer	82

SUMÁRIO

1	INTRODUÇÃO	17
1.1	Contextualização e Motivação	17
1.2	Objetivos	20
1.3	Metodologia de pesquisa	21
1.4	Organização do documento	21
2	FUNDAMENTAÇÃO TEÓRICA	23
2.1	Fundamentos Biológicos	23
2.1.1	Câncer de mama e seus subtipos	24
2.1.2	Microarrays de DNA	25
2.1.3	Dados de expressões gênicas	25
2.1.4	Conjuntos de Genes	27
2.2	Dados Ausentes	27
2.2.1	Mecanismos de Ausência	28
2.2.2	Métodos de Tratamento de Valores Ausentes	29
2.2.3	Métodos de Estimação	30
2.3	Aprendizado de máquina	32
2.3.1	Aprendizado de Máquina Indutivo	32
2.3.1.1	Aprendizado não supervisionado	34
2.3.1.2	Aprendizado supervisionado	34
2.3.1.3	Aprendizado semisupervisionado	36
2.3.2	Métodos supervisionados de Machine Learning	36
2.3.2.1	Máquina de Vetores de Suporte	37
2.3.2.2	Eliminação de recursos recursivos	39
2.3.2.3	Adição de recursos recursivos	41

2.3.3	Divisão de dados	42
2.3.3.1	Validação cruzada k-fold	43
2.3.3.2	Leave-one-out cross-validation (LOOCV)	44
2.4	Correlação nos dados	44
2.5	Métricas de Avaliação	45
2.5.1	Métricas de regressão	46
2.5.2	Métricas de classificação	46
2.6	Considerações finais	47
3	TRABALHOS RELACIONADOS	48
3.1	Métodos de estimação de valore ausentes	48
3.2	Discussão	49
3.3	Considerações Finais	51
4	SOLUÇÃO PROPOSTA	52
4.1	Estimação de expressões gênicas baseada em correlação (EEGBC)	52
4.1.1	Pré-processamento e separação por subtipo	53
4.1.2	Seleção de valores ausentes	53
4.1.3	Seleção de genes correlacionados	54
4.1.4	Estimação	54
4.2	Avaliação da Solução Proposta	55
4.2.1	Conjunto de Dados e Pré-Processamento	55
4.2.2	Estimando dados faltantes	56
4.2.3	Avaliação	58
5	RESULTADOS	59
5.1	Análise de valores omissos	59
5.2	Estimação de valores omissos	60
5.3	Interseção de genes	64
5.4	Classificação de subtipos	72
5.5	Considerações Finais	73

6	CONSIDERAÇÕES FINAIS	75
6.1	Contribuições	75
6.2	Próximos passos	76
Referências		77

1

INTRODUÇÃO

Este capítulo apresenta os principais aspectos que motivam a realização desta pesquisa de mestrado, uma breve revisão da literatura, o objetivo proposto e a metodologia da pesquisa, assim como a organização deste trabalho.

1.1 Contextualização e Motivação

Ao longo dos anos, métodos que possibilitam obter dados, analisá-los e transformá-los em conhecimento útil foram desenvolvidos e procedimentos melhorados. No entanto, ainda hoje pesquisadores enfrentam diversos desafios quando o assunto diz respeito à mineração de informações, e ao aprendizado de máquina (AM), as quais são alternativas eficazes e têm sido bastante utilizadas para extrair conhecimento a partir de um grande número de dados [[Abdella e Marwala 2005](#)].

No campo da Biologia e Genética, o Projeto Genoma foi desenvolvido com o objetivo de identificar genes responsáveis pelas características normais e patológicas dos indivíduos, e com o tempo evoluiu de modo a, além de identificar, mapear e sequenciar o DNA humano [[Gibbs 2020](#)]. Com isso, surgiu uma grande rede de dados passíveis de análise que possibilitou a criação de base de dados públicas de modo a abrigar essa gama de informações.

A expressão genética ou expressão gênica refere-se ao processo em que a informação codificada por um determinado gene, tal como a sequência de DNA, é decodificada em RNA mensageiro e a partir disso em um produto funcional (proteína) [[Volgin 2014](#)].

Cada gene fornece instruções para um produto funcional, ou seja, uma molécula necessária para realizar um trabalho em uma dada célula. Graças a este conceito, é possível ter uma diferenciação celular, que são variedades de tipos celulares em um mesmo organismo, realizando atividades diferentes.

No contexto das bases de dados públicas, compreender como genes se relacionam e entender de modo amplo o funcionamento de sistemas biológicos torna-se algo possível, porém complexo devido, por exemplo, a presença de valores ausentes em meio aos ¹dados genômicos que ocupam as bases de dados de expressão gênica, e os quais podem reduzir a qualidade das estimativas estatísticas de um estudo de modo a produzir resultados tendenciosos, levando a conclusões inválidas [Kang 2013].

Analisar a expressão gênica permite descobrir defeitos genéticos, assim como informações sobre o funcionamento dos processos que controlam o ciclo celular. Medir de forma precisa expressões genéticas de um conjunto de células tem aplicações de extrema importância na medicina, principalmente quando o assunto é diagnóstico e caracterização do câncer [Singh et al. 2018]. De maneira genérica, o câncer abrange diferentes tipos de doenças malignas em que as células cancerígenas dividem-se de forma incontrolável e invadem tecidos ou órgãos causando a destruição dos mesmos. O câncer é uma das doenças que mais mata pessoas no mundo e a variância dos seus tipos apresenta uma discrepância de mortes por gênero (feminino ou masculino) [INCA 2021]. No Brasil, o tipo mais comum entre as mulheres é o câncer de mama, segundo dados do Instituto Nacional do câncer (INCA) em 2020 a incidência deste tipo de câncer neste grupo foi de 66.280 novos casos, o que representa uma taxa de 29,5% dos cânceres em mulheres, e 17.825 óbitos, o que representa 16,5% de mortes dos cânceres em mulheres [INCA 2021]. Além disso, existem tipos moleculares diferentes de câncer de mama. São eles: Luminal A, Luminal B, HER2 positivo e Basal. Cada tipo está associado a um comportamento biológico específico e reconhecer que subtipo um paciente possui ajuda os médicos a definirem qual a terapia mais adequada para o tratamento.

Com o avanço das tecnologias e da ciência, houve um aumento no estudo do câncer, de modo que no contexto atual há muitos repositórios públicos contendo

¹ Os dados genômicos referem-se aos dados do genoma e do DNA de um organismo. Eles são usados em bioinformática para coletar, armazenar e processar os genomas de seres vivos.

conjuntos de dados de expressões gênicas, geradas a partir da tecnologia de microarrays, que trazem amostras de tumores e os valores de expressão de genes/proteínas para aquele tumor, os quais representam o quanto determinados genes/proteínas estão ou não presentes no tecido cancerígeno [Mendonca-Neto et al. 2021].

A técnica de microarrays, tem a capacidade de gerar um grande volume de dados de expressão gênica e é bastante utilizada para análise em larga escala desses dados, sendo microarray descrito como o conjunto de pequenos fragmentos de DNA dispostos em segmentos específicos de uma superfície [Xiang et al. 2008].

No contexto das bases de dados de expressão gênica de câncer de mama, extraídas a partir de microarrays, a ausência de dados, normalmente, é ocasionada por motivos como: (i) poeira ou arranhões no slide; (ii) imagem corrompida; ou (iii) resolução insuficiente [Xiang et al. 2008].

A estimação de valores ausentes nessas bases, pode contribuir significativamente para os bases de dados públicas tendo em vista que estimar esses dados pode tornar a caracterização do câncer de mama e classificação de subtipos mais precisas, de modo que modelos mais eficientes e realistas garantam a determinação de melhores tratamentos para cada subtipo de câncer de mama, como, por exemplo, o tratamento terapêutico mais adequados conforme a evolução clínica dos pacientes, e também demonstrar padrões de possíveis ocorrências do câncer a partir da observação da diferença de expressão gênica dos organismos estudados. Além disso, a estimação permite que outros pesquisadores possam ter acesso a mais dados sobre o câncer e continuar produzindo melhorias nas bases de dados públicas, como realizar a tarefa de integração de bases de dados a qual depende de bases de dados completas.

Na literatura, alguns métodos realizam a estimação de valores faltantes em bases de dados de expressões gênicas utilizando conjunto de genes que apresentam forte relação com os genes a serem estimados. Alguns deles utilizam a abordagem de correlação entre genes, de modo que apenas um subconjunto de genes que exibe alta correlação com o gene que contém os valores faltantes é usado para computar esses valores no gene.

No entanto, não identificamos, entre os trabalhos pesquisados, métodos voltados

para estimação de expressões gênicas em bases de dados de câncer de mama que explorem a estimação por subtipo. Além disso, apesar de existirem muitos métodos e algoritmos voltados para estimação de dados, não há um que funcione de modo mais eficiente em todas as bases [Liew e Yan 2011].

Sendo assim, em nossa pesquisa buscamos estimar genes de câncer de mama com valores ausentes por meio da utilização dos genes de câncer de mama sem dados ausentes mais correlacionados a eles, por meio da separação por subtipo de câncer e usando algoritmo de aprendizado de máquina supervisionado.

Com base nisso, desenvolvemos a metodologia de Estimação de Expressões Gênicas Baseada em Correlação (EEGBC). Os experimentos, bem como a análise do método de estimação de genes faltantes proposto, são apresentados no Capítulo 4.

1.2 Objetivos

O objetivo principal deste trabalho é o desenvolvimento de um método de estimação de expressões gênicas de câncer de mama baseado em correlação que permita selecionar os genes completos mais correlacionados ao genes com dados faltantes, a serem estimados por subtipo de câncer de mama, através da utilização de algoritmo supervisionado de AM, para gerar bases de dados mais completas e precisas.

Os objetivos específicos são:

- Selecionar genes relevantes sem valores ausentes que ajudem a estimar os genes com dados ausentes por subtipo de câncer de mama;
- Demonstrar a eficácia do método proposto, ou seja, um estimador de maior qualidade utilizando correlação por meio da utilização de menos genes para realizar a estimação do que os métodos mais tradicionais;
- Avaliar o impacto dos valores estimados em problemas de classificação.

1.3 Metodologia de pesquisa

Para tratar este desafio e alcançar os objetivos definidos anteriormente, os seguintes passos são propostos:

- Revisão Bibliográfica: identificação do Estado-da-arte relacionado ao tema de pesquisa;
- Coleta de Dados: coleta e avaliação de Bases de Dados de câncer de mama;
- Tratamento de Dados: caracterização das bases de dados com identificação de distribuição de valores dos dados, identificação de valores faltantes, dimensão (genes x amostras);
- Dados faltantes: desenvolvimento de métodos para tratamento de dados faltantes em bases individuais;
- Métodos: avaliação dos métodos, por meio de modelo de classificação e métricas adequadas;
- Publicação de resultados em veículos de qualidade.

1.4 Organização do documento

Este primeiro capítulo apresentou o contexto, no qual este trabalho está inserido, a motivação para realizá-lo, o objetivo a ser atingido e a metodologia a ser seguida. Além deste capítulo, a organização do texto está estruturada em mais cinco capítulos, descritos a seguir:

Capítulo 2: aborda o referencial teórico, no qual esse trabalho é baseado com os seguintes conteúdos: (2.1) Fundamentos Biológicos, (2.2) Dados Ausentes, (2.3) Aprendizado de Máquina, (2.4) Correlação nos dados, (2.5) Métricas de Avaliação e (2.6) Considerações Finais.

Capítulo 3: discute os trabalhos relacionados.

Capítulo 4: descreve a solução proposta: (4.1) Estimação de expressões gênicas baseada em correlação (EEGBC) e (4.2) Avaliação da solução proposta.

Capítulo 5: descreve os resultados obtidos: (5.1) Análise de valores omissos, (5.2) Estimação de valores omissos, (5.3) Interseção de genes, (5.4) Classificação de subtipos e (5.5) Considerações Finais.

Capítulo 6: apresenta as considerações finais da pesquisa: (6.1) Contribuições e (6.2) Próximos Passos.

2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos necessários para a compreensão desta dissertação. Primeiramente, abordamos alguns conceitos biológicos como expressão gênica, usados como dados em nosso método. Em seguida, discutimos sobre dados ausentes, assunto que move o estudo proposto. A seguir, descrevemos o conceito de aprendizado de Máquina, bem com alguns de seus métodos supervisionados que utilizaremos para estimar os valores faltantes. Em frente falamos sobre os métodos de aprendizado de máquina empregados nesta proposta, incluindo métodos de aprendizado profundo, subárea de conhecimento. Depois disso, demonstramos as métricas utilizadas para avaliar o desempenho dos métodos de estimação. Após isso, falamos detecção de outliers, técnica que usamos para determinar os genes importantes a serem estimados. Em seguida, discutimos sobre correlação nos dados de expressão gênica, que nos auxilia a entender a similaridade entre os dados. Por fim, apresentamos um breve resumo do capítulo.

2.1 Fundamentos Biológicos

Nas últimas décadas, houve um aumento no estudo do câncer, visto que há muito mais bases de dados públicas contendo dados de expressão de gene do que em anos anteriores, devido tanto a disponibilização desses dados através de pacientes, quanto por meio da análise e geração desses dados por meio de tecnologias, como microarrays. Essa gama de informações trouxe a possibilidade da realização de novos estudos a respeito do cân-

cer, o desenvolvimento de estratégias diagnósticas e terapêuticas [Roukos e Ku 2012]. No entanto, ainda existem muitos impecilhos quando o assunto é estudo do câncer de forma precisa e mais determinística, como a ocorrência de expressões ausentes nas bases de dados.

2.1.1 Câncer de mama e seus subtipos

O câncer de mama é uma doença bastante heterogênea, altamente morfológica e biológica. Existem diversos subtipos moleculares do carcinoma de mama, os quais apresentam diferentes complexidades e que exigem tratamentos terapêuticos específicos. Dentro desse cenário, entender os melhores tratamentos para o câncer de mama, depende de fatores prognósticos, preditivos e patológicos obtidos a partir da análise dos diferentes tipos de câncer [Rakha et al. 2007].

Por meio da análise de mais de 8.000 genes, pesquisadores caracterizaram diferentes perfis de expressão gênica dos tumores mamários, em 2000 pacientes, sendo eles os quatro subtipos abordados abaixo [Yersal e Barutça 2014]:

1. *Luminal A*: é o subtipo mais comum. Representa 60% de todos os cânceres de mama e apresenta o melhor prognóstico.
2. *Luminal B*: compreendem 15%-20% dos cânceres de mama e têm um fenótipo mais agressivo, maior grau histológico, índice proliferativo e pior prognóstico.
3. *Superexpressão de HER2*: possui elevada expressão da oncoproteína HER2, porém apresenta negatividade para receptores hormonais. Existe dois tipos de HER2: Positivo e Negativo. O câncer de mama HER2-positivo ocorre em aproximadamente 25% dos casos de câncer de mama.
4. *Basaloide*: também chamado de "Basal". São biologicamente agressivos, contudo, têm alta taxa de resposta a agentes quimioterápicos convencionais.

Na literatura, o conjunto de genes do PAM50 são aceitos como representativos para a caracterização do câncer de mama. Nos últimos anos, a assinatura do gene

PAM50 passou a ser considerado o conjunto referencial de genes para diferenciar os subtipos se tornando uma ferramenta prognóstica valiosa na investigação do cancro da mama e orientando a tomada de decisões terapêuticas [[Mendonca-Neto et al. 2021](#)].

2.1.2 Microarrays de DNA

Existem diferentes experimentos de microarray e cada um deles possui suas especificidades, mas ambos envolvem três procedimentos básicos comuns [[Tefferi et al. 2002](#)]:

- **Fabricação de chips:** Um microarray é um pequeno chip, onde milhares de moléculas de DNA são fixadas em grades fixas. Cada célula da grade está relacionada a uma sequência de DNA.
- **Preparação, marcação e hibridização do alvo:** Normalmente, duas amostras de mRNA (RNA mensageiro) (amostras de teste e controle) são transcritas reversamente em cDNA (DNA complementar), marcadas usando corantes fluorescentes ou isótopos radioativos e, em seguida, hibridizadas com as moléculas de DNA na superfície do chip.
- **O processo de varredura:** os chips são escaneados para ler a intensidade do sinal que é emitido dos alvos rotulados e hibridizados.

2.1.3 Dados de expressões gênicas

Cada um dos subtipos de câncer de mama foram descobertos devido a tecnologias como o microarray de expressão gênica, que no cenário atual, é o método mais utilizado para extrair, analisar e monitorar dados de expressão gênica de câncer e outros tipos de doenças malignas [[Liew e Yan 2011](#)].

As tecnologias de microarrays possibilitaram novas formas de coletar a expressão gênica de material biológico, permitindo o desenvolvimento de abordagens genômicas para prognóstico de câncer e previsão de terapia.

Expressões gênicas são a informação do processo pelo qual um gene fornece instruções para um produto funcional, ou seja, uma molécula necessária para realizar um trabalho em uma dada célula. Graças a expressão gênica, é possível ter uma diferenciação celular, que são variedades de tipos celulares em um mesmo organismo [Kundaje et al. 2015, Lonsdale et al. 2013].

A expressão gênica medida através de microarray é armazenada em conjuntos de dados, o quais consistem em uma matriz M de valor real, com linhas correspondendo aos níveis de expressões gênicas dos genes e colunas representando os perfis de expressão das amostras. Cada célula M_{ij} é o nível de expressão medido do gene i na amostra j . Isso pode ser visualizado por meio da Figura 1.

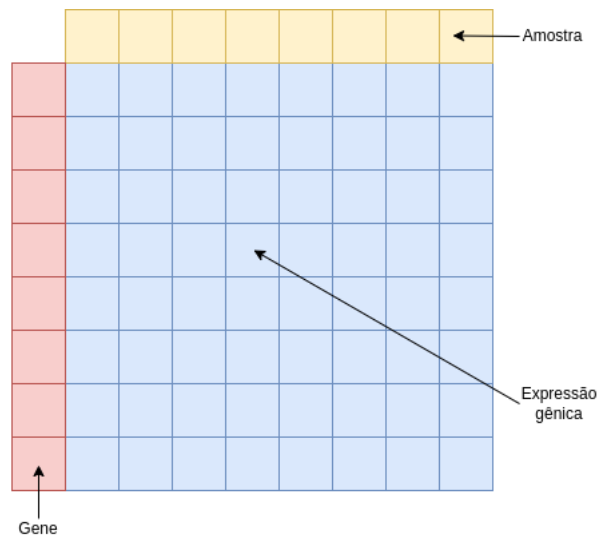


Figura 1 – Exemplo de conjunto de dados, figura adaptada de de Ayyad et al. (2019).

A matriz é composta por n linhas e m colunas, o que é conhecido como perfil de expressão gênica [Tan e Gilbert 2003]. As linhas representam os genes e as colunas representam as amostras. Comumente, o $m < n$. Este número desigual de amostras e genes é devido à dificuldade de coleta de amostra de microarray. Esses dados de alta dimensão podem causar a maldição da dimensionalidade, a qual resulta em métricas de distância imprecisas e impacto na precisão da classificação de subtipos. Para resolver o este problema em dados de expressão gênica, as abordagens de modelagem predominantes incluem filtragem de genes na fase de pré-processamento de dados e seleção de subconjunto de genes [Newman et al. 2017]. Neste contexto, quando um gene apresenta

uma ou mais expressões ausentes dizemos que ele é um gene incompleto, caso contrário nos referirmos a ele como um gene completo.

2.1.4 Conjuntos de Genes

Conjunto de genes são exatamente o que o termo diz: um conjunto de genes, uma coleção desordenada e não estruturada de genes [Liberzon et al. 2011]. Conjuntos de genes são definidos como a coleção de genes associados a um processo biológico específico (por exemplo, ciclo celular) e localização (por exemplo, no cromossomo 1). Também está associado a doenças (por exemplo, câncer de mama), ou mesmo ao conjunto de genes presentes em uma determinada via (por exemplo, o conjunto de 128 genes envolvidos na via do ciclo celular KEGG). Sendo assim, um conjunto de genes pode ser elaborado a partir de critérios arbitrários. O Molecular Signatures Database [Liberzon et al. 2011] (MSigDB) inclui mais de 10.000 desses conjuntos de genes definidos com base em muitos critérios, alguns deles aparentemente arbitrários.

2.2 Dados Ausentes

A falta de dados é um problema que afeta grande parte dos bancos de dados de domínio público, como bancos de dados de registros médicos eletrônicos e bancos de dados de cunho econômico e político. Eles podem ser ocasionados por uma série de fatores, como: a falta de respostas e falhas nos meios de extração. No contexto médico essa falta de valores poderia ser ocasionada por: (i) a variável é medida, mas por alguma razão não foi identificada, os valores não são registrados eletronicamente por desconexão de sensores, erros de comunicação com o servidor de banco de dados, omissão humana acidental e outros; ou (ii) a variável não é medida durante um determinado período de tempo devido a um motivo identificável [Data 2016]. No contexto das bases de dados de expressão gênica extraídas a partir de microarrays, a ausência pode ser ocasionada por: (i) poeira ou arranhões no slide; (ii) imagem corrompida; ou (iii) resolução insuficiente [Xiang et al. 2008].

Embora muitos algoritmos usados em mineração de dados lidem com o problema de valor ausente, espera-se que melhores resultados sejam obtidos se os dados ausentes forem removidos ou estimados, durante a fase, inicial, de pré-processamento. No pré-processamento são aplicados métodos que melhoram a qualidade dos dados, sendo uma fase semiautomática, pois depende do analista, o qual precisa ter um bom conhecimento do domínio para identificar o problema de valores omissos nos dados coletados e, então, utilizar os métodos adequados para solucioná-lo.

É primordial lidar com dados ausentes, caso contrário esses tipo de dados podem causar um efeito significativo nas conclusões tiradas pelos pesquisadores e em muitos casos invalidando a hipótese induzida.

Para fazer o tratamento de dados ausentes existe um processo conhecido como estimação que possui como objetivo substituir quaisquer valores ausentes por um determinado valor estimado com base nas informações disponíveis. Apesar de existirem outras formas de lidar com esses dados, como excluindo os mesmos das bases de dados, a estimação é dada como um dos processos mais adequados de tratamento, visto que garante bases completas e que possam vir a ser estudadas sem perda de informações.

2.2.1 Mecanismos de Ausência

Para que seja possível lidar com os dados ausentes, de modo eficiente, é necessário compreender os mecanismos pelos quais os dados estão faltando. Esses mecanismos consistem em verificar as relações existentes entre os valores perdidos e a probabilidade de ausência desses dados, de modo a informar o que gerou a ocorrência dos mesmos [Enders 2010]. A seguir são apresentados cada um desses mecanismos:

- **Totalmente Ausente ao Acaso (MCAR):** Em um conjunto de dados, os valores podem ser ausentes totalmente ao acaso (MCAR). Isso ocorre quando o valor do atributo não está relacionado com esse atributo ou qualquer outro do conjunto de dados (dados observáveis).
- **Faltando ao Acaso (MAR):** A probabilidade de um valor estar faltando aleatoriamente (MAR) está relacionada apenas aos dados observáveis, sendo assim é

possível estimar os valores ausentes a partir dos dados observados visto que esses dados estão relacionados com as variáveis ausentes. O termo ausente ao acaso é um tanto enganoso porque pode levar a entender que os dados estão ausentes de uma forma aleatória que se assemelha ao lançamento de uma moeda. No entanto, MAR na verdade significa que existe uma relação sistemática entre uma ou mais variáveis medidas e a probabilidade de dados perdidos [Enders 2010].

- **Faltando Não ao Acaso (MNAR):** Dados ausentes não ao acaso (MNAR) ocorrem quando os dados ausentes dependem do próprio atributo e dos dados observados. Por exemplo, suponha que alunos com habilidades de leitura ruins tenham notas perdidas em um teste de literatura porque tiveram dificuldade de compreender a leitura durante o exame, nesse caso os dados são MNAR devido a probabilidade de um valor ausente ser dependente da variável que está faltando [Enders 2010]. Em contraste com o padrão MAR, a variável ausente no caso MNAR não pode ser predita apenas levando-se em consideração as variáveis do conjunto de dados.

2.2.2 Métodos de Tratamento de Valores Ausentes

Na literatura existem diferentes métodos para se tratar valores ausentes, mas muitos deles são simples e podem causar grandes distorções nos resultados das análises se forem utilizados sem os devidos cuidados. Podemos dividir esses métodos nas seguintes categoritas [Little e Rubin 2019]:

- **Descarte de dados:** Nesta abordagem os exemplos que possuem valores ausentes são descartados. Este é o método mais amplamente utilizado em muitos programas de mineração de dados, devido a sua simplicidade.
- **Estimativa de parâmetros:** Utiliza-se métodos estatísticos para estimar parâmetros de modelos e a partir desses modelos são estimados os valores ausentes.
- **Estimação:** Essa classe de métodos estima os valores ausentes através da análise de outros valores coletados. Existem diversas técnicas de estimação que vão de simples como uso da média ou moda, até as mais avançadas que usam aprendiza-

gem de máquina. Os métodos mais avançados costumam ser mais complexos e custosos, mas, geralmente, geram resultados melhores.

2.2.3 Métodos de Estimação

Como citado anteriormente, a estimação consiste em substituir os valores ausentes por valores estimados. Os métodos de substituição de valores ausentes tem sido estudados há anos. A seguir são descritos alguns dos métodos encontrados na literatura:

- **Estimação pela média, mediana ou moda:** É um dos métodos amplamente utilizados, consiste na substituição dos valores ausentes pela média (no caso de atributos numéricos) ou pela moda (atributos nominais), calculadas a partir dos valores observados dos atributos. A média normalmente é uma boa estimativa de um valor ausente, quando não se tem nenhuma outra informação a respeito do conjunto de dados. É um método conservador que possui como principal vantagem não alterar a média geral do atributo, mas altera a variância dos dados reduzindo seu valor e também pode alterar o relacionamento entre os atributos.
- **Conhecimento de domínio:** Algoritmos nesta categoria fazem a integração de conhecimento de domínio ou informação externa no processo de estimação. O uso do conhecimento de domínio tem o potencial de melhorar significativamente a precisão da estimação além do que é possível com uma abordagem puramente orientada a dados, especialmente para conjuntos de dados com pequeno número de amostras, ruidosos ou com alta taxa de falta. Em bases de expressões gênicas, algoritmos nesta categoria podem fazer uso de conhecimento sobre o processo biológico no experimento de microarray [[Liew e Yan 2011](#)].
- **Estimação Local *Hot-deck* ou *Cold-deck*:** Neste método os valores ausentes são estimados a partir da estrutura de similaridade local no conjunto de dados. Em bases de expressão de genes apenas um subconjunto de genes que exibe alta correlação com o gene que contém os valores ausentes é usado para computar os valores ausentes no gene [[Liew e Yan 2011](#)]. O nome *Hot-deck* vem da época em

que os dados eram armazenados em cartões perfurados e significa que o exemplo similar vem do mesmo conjunto de cartões (deck) que o exemplo com valores ausentes, ou seja, o conjunto de cartões que estavam sendo processados naquele momento, já no *Cold-deck* o exemplo viria de um outro conjunto do qual o exemplo com valores ausentes não fazia parte. O processo consiste em duas etapas, na primeira é usado um método de agrupamento não supervisionado para se separar os exemplos similares. Posteriormente, os exemplos com valores ausentes são associados a cada um desses agrupamentos e os seus valores são preenchidos usando a média ou moda do agrupamento escolhido ou simplesmente escolhendo aleatoriamente um exemplo similar e utilizando os valores de seus atributos para preencher os valores que estão ausentes no outro exemplo.

- **Estimação com base em k-vizinhos mais próximo (KNNImpute):** Assim como a estimação *Hot-deck*, o KNNImpute estima os valores ausentes partir da estrutura de similaridade local no conjunto de dados. Os valores faltantes de cada amostra são imputados usando o valor médio dos vizinhos mais próximos encontrados no conjunto de treinamento. Duas amostras são próximas se os recursos que não estão faltando em nenhuma delas estiverem próximos.
- **Estimação Global:** Neste método os valores ausentes são estimados com base nas informações de correlação global derivadas de toda a matriz de dados. Em bases de expressão de genes, os algoritmos desta categoria assumem a existência de uma estrutura de covariância global entre todos os genes ou amostras na matriz de expressão. Quando esta suposição não é adequada, ou seja, quando os genes exibem estruturas de similaridade local dominante, sua imputação torna-se menos precisa.
- **Modelos preditivos:** Os modelos preditivos atualmente são o estado da arte no tratamento de valores ausentes. São utilizados métodos estatísticos ou de aprendizado de máquina para a criação de um modelo preditivo para estimar os valores que serão usados para substituir os valores ausentes. O atributo com valor ausente é utilizado como atributo de classe e o restante dos atributos são utilizados como entrada para o modelo preditivo. Os modelos preditivos utilizam

correlações entre os atributos para estimarem os valores ausentes, por isso é necessário haver correlações entre os genes do conjunto de dados, caso contrário os valores ausentes não serão corretamente estimados.

2.3 Aprendizado de máquina

O aprendizado de máquina (AM) (*do inglês*, Machine Learning - ML) trata-se da capacidade que um artefato construído pelo homem tem de analisar dados para automatizar a criação de modelos analíticos. Portanto, é uma vertente da inteligência artificial, um conceito mais amplo, que diz respeito à capacidade que uma máquina tem de tomar decisões a partir de um raciocínio que lembra o pensamento humano. No caso da AM, espera-se que essas decisões tomadas pelos equipamentos tenham como base o aprendizado com dados e a identificação de padrões com o mínimo de (ou sem nenhuma) intervenção humana [Russel e Norvig 2015].

Desde sua criação, na década de 50, a IA tem evoluído bastante, no entanto, ainda é necessário muito estudo para conseguir criar máquinas que pensem ou ajam como seres humanos de forma geral. O grande desafio desta tarefa está na compreensão de como o ser humano aprende e como melhora o seu desempenho cada vez que realiza uma tarefa.

Existem diversas estratégias de aprendizagem, as quais em ordem crescente de complexidade são: hábito, instrução, dedução, analogia e indução [Monard et al. 1997]. Em nossa pesquisa, fazemos o uso do aprendizado indutivo, visto que utilizamos algoritmos de aprendizado supervisionado para realizar a estimação de valores ausentes. Descrevemos a seguir cada um desses conceitos.

2.3.1 Aprendizado de Máquina Indutivo

Apesar do aprendizado indutivo ser o de maior complexidade, ou seja, o que o aprendiz gasta mais tempo para aprender [Monard et al. 1997], ele é o método mais útil pois permite obter novos conhecimentos a partir de um conjunto de exemplos, ou casos

observados previamente. Este tipo de abordagem, induz descrições gerais de conceitos utilizando-se de exemplos específicos desse conceito. Por exemplo, um aluno que estuda para uma prova resolvendo uma lista de exercícios similares aos que existirão na prova, extrai conhecimento desses exemplos e quando encontra exercícios similares na prova tem mais chance de resolvê-los. Vários algoritmos de AM utilizam aprendizado indutivo por meio de exemplos, neste caso, estes algoritmos também são chamados de indutores e realizam generalizações e especializações na indução de uma hipótese h . Esses sistemas tomam decisões baseadas em experiências acumuladas contidas em outros casos resolvidos com sucesso. Basicamente, o aprendizado de máquina indutivo pode ser dividido em aprendizado supervisionado, semissupervisionado e não supervisionado.

No aprendizado supervisionado, a indução dos conceitos é realizada a partir de exemplos pré-classificados, ou seja, os exemplos já estão rotulados com uma classe ou conceito conhecido. No aprendizado não supervisionado as classes são desconhecidas e no semissupervisionado existe a mistura do aprendizado supervisionado com o não supervisionado, sendo assim consegue induzir conceitos de exemplos rotulados e não rotulados.

Dado um conjunto de exemplos, é importante distinguir os quatro tipos de conjuntos comumente utilizados em grande parte dos algoritmos de AM:

- **Conjunto de treinamento ou treino:** é a entrada dos algoritmos de AM, também conhecido como casos vistos pois refere-se aos exemplos vistos antecipadamente, pelo algoritmo de AM. A partir dele os modelos ou hipóteses são construídos.
- **Conjunto de teste:** é o conjunto utilizado para avaliar o modelo. Também conhecido como casos não vistos. Ele deve ser "visto" pelo algoritmo na fase de indução do modelo e seus exemplos devem ser disjuntos do conjunto de treino.
- **Conjunto de validação:** Em alguns casos é necessário realizar ajustes no modelo, para isso é necessário um conjunto de exemplos que não foram utilizados na construção do modelo. Como o conjunto de teste seria "visto", é necessário um novo conjunto com exemplos não vistos.

- **Conjunto de predição:** Conjunto de exemplos não rotulados que devem ter seu rótulo previsto pelo modelo induzido.

2.3.1.1 Aprendizado não supervisionado

O aprendizado não supervisionado possui dados não rotulados que o algoritmo deve tentar entender por conta própria, com base no próprio conjunto de dados. Sendo assim, a sua entrada é um conjunto de dados $X = \{x_1, x_2, \dots, x_n\}$, onde cada exemplo consiste somente do vetor $\vec{x}$, sem o atributo de classe Y . A tarefa mais frequente no aprendizado supervisionado tem como objetivo encontrar alguma regularidade no conjunto de dados, formando agrupamentos de exemplos que tem alguma similaridade, ou seja, os exemplos são descritos em agrupamentos (*do inglês*, cluster) de acordo com alguma medida de similaridade. As tarefas mais comuns desse tipo de aprendizado incluem:

- **Agrupamento:** o objetivo é agrupar os exemplos usando uma medida de similaridade;
- **Deteção de novidade:** identifica os poucos exemplos que diferem da maioria;
- **Redução de dimensionalidade:** cujo objetivo é reduzir a dimensão do vetor de características enquanto mantém características essenciais da amostra de treino;
- **Regras de associação:** o objetivo é encontrar associações interessantes, como relacionamentos e dependências em grandes conjuntos de dados.

A maior parte dos algoritmos de estimação de valor ausentes utilizam a abordagem supervisionada de estimação, com base em similaridade entre os dados, ou seja, usam o agrupamento.

2.3.1.2 Aprendizado supervisionado

No aprendizado supervisionado, existe um rótulo $y_i \in Y$ que é a predição desejada para um exemplo $x_i \in X$ na amostra de teste. Os rótulos podem vir de um conjunto

de valores infinitos. Esses valores distintos entre si, são definidos como classes. Por exemplo, ao classificar um paciente com subtipo de câncer de mama Basal ou Não Basal, tem-se a variável tipo de câncer como sendo uma classe (definida por rótulos Basal ou Não Basal). Na realidade, normalmente, esses rótulos são definidos como 0 ou 1, respectivamente, verdadeiro ou falso. Sendo assim, estaria se olhando para a seguinte pergunta: *O paciente tem o subtipo de Câncer Basal?* Cujas respostas seriam 0, se sim, e 1, se não.

Com esses conceitos pode-se definir o aprendizado supervisionado como sendo o domínio dos exemplos X e o domínio dos rótulos Y , tendo $P(x, y)$ como uma junção de distribuições desconhecidas dos exemplos e rótulos $X \times Y$. Dada uma amostra de treinamento, o aprendizado supervisionado induz a função $h : X \rightarrow Y$ que se aproxime da verdadeira função desconhecida $f : X \rightarrow Y$, com o objetivo de $h(x)$ prever o verdadeiro rótulo y_i em um exemplo futuro x_i . Dependendo, do domínio do rótulo y_i , os problemas de aprendizado supervisionado podem ser divididos em classificação ou regressão:

- **Classificação:** a classificação é caracterizada pelo tipo do valor do atributo Y que é discreto. Após isso, o algoritmo irá induzir uma função $h : X \rightarrow Y$, onde, por exemplo, $y \in \{\text{Basal}, \text{Não Basal}\}$, com o objetivo de prever o verdadeiro rótulo.
- **Regressão:** constituem problemas de aprendizado supervisionado onde o conjunto de valores do atributo de classe Y é contínuo, ou seja, $Y \in \mathbb{R}$.

Como já mencionado, o conjunto de dados é dividido em dois conjuntos disjuntos: treino, usado para induzir a hipótese, e teste, usado para validar a hipótese, verificando se ela é genérica, mostrando que o algoritmo aprendeu com os exemplos "vistos". A generalização do conceito aprendido pode sofrer os seguintes problemas:

- **Superajuste (*do inglês, overfitting*):** Ocorre quando a hipótese induzida é altamente especializada para o conjunto de treino, atingindo uma alta taxa de acertos neste conjunto, mas uma baixa taxa de acertos no conjunto de teste, o que significa que o algoritmo decorou os exemplos e não consegue generalizar o conceito com outro conjunto de dados.

- **Subajuste (*do inglês, underfitting*):** Ocorre quando o algoritmo não consegue abstrair o conhecimento do conjunto de treino, isso pode ocorrer devido a um subdimensionamento de parâmetros do algoritmo, ou seja, o modelo é incapaz de capturar a relação entre as variáveis de entrada e saída com precisão, gerando uma alta taxa de erro no conjunto de treinamento e nos dados não vistos.

2.3.1.3 Aprendizado semisupervisionado

Como o próprio nome sugere, o aprendizado semissupervisionado é o meio termo entre o aprendizado supervisionado, o qual generaliza padrões nos dados rotulados de treinamento, e o não supervisionado, que identifica propriedades intrínsecas do conjunto de dados não rotulados. Tem surgido como uma alternativa quando existe uma grande quantidade de dados não rotulados disponíveis e poucos dados rotulados.

2.3.2 Métodos supervisionados de Machine Learning

Como descrito anteriormente, métodos de aprendizado de máquina (ML) podem ser usados para resolver problemas de classificação ou regressão e, na literatura, aparecem em uma gama de estudos. Pesquisas recentes usam algoritmos de ML com o objetivo de estimar dados ausentes em bases de diferentes tipos de dados. Nesta pesquisa estudaremos a eficiência da estimação de dados de expressão gênica de câncer de mama utilizando algoritmos de regressão supervisionados, nos quais o processo básico de estimação pode ser visualizado por meio da Figura 2.

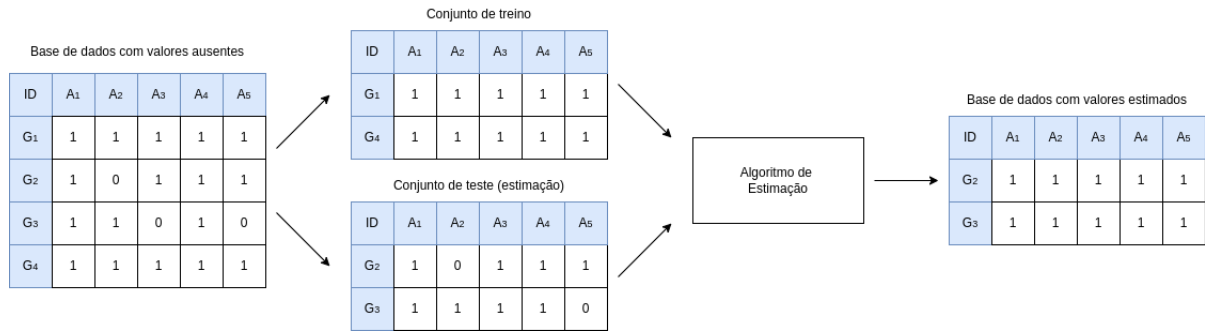


Figura 2 – Estimacao utilizando algoritmo supervisionado, Figura adaptada de Castaneda et al., (2009). Os valores iguais a 1 representam dados não ausentes e os valores iguais a 0 representam os dados ausentes.

O conjunto de dados é dividido em dois conjuntos: o de treino, que possui apenas os exemplos completos, e o de estimacao, que possuem os exemplos com valores ausentes. Após isso, os conjuntos são processados por um algoritmo de estimacao que preenche os valores ausentes e como resultado se obtém dados completos. Note que, apesar de qualquer algoritmo de estimacao poder ser usado, a divisao em treino e estimacao é semelhante à divisao treino e teste do aprendizado supervisionado. E apenas dados completos do conjunto de treino são usados para induzir a hipótese que irá estimar os valores ausentes. Sendo assim, o conjunto de estimacao não é utilizado pelo algoritmo de estimacao na inducao da hipótese.

A seguir descrevemos o algoritmo Máquina de Vetores de Suporte que é utilizado nesta pesquisa.

2.3.2.1 Máquina de Vetores de Suporte

O algoritmo de Máquina de Vetores de Suporte (SVM, *do inglês*: support-vector machine) é um algoritmo de kernel esparsos usado em problemas de classificacao e regressao, baseado na teoria do aprendizado estatístico (TAE), introduzida em 1992. Aqui discutiremos a estrutura SVM para o problema de regressao em sua abordagem matemática, no entanto os mesmos caminhos se seguem em seu problema de classificacao.

Dado um conjunto de dados de treinamento onde x_n é um conjunto multivariado de N observacoes com valores de resposta observados y_n . Para encontrarmos a funcao linear:

$$y_n = f(x) = x_n w^T + b,$$

onde w^T é vetor de pesos e b é o bias (termo de viés da equação do hiperplano), é necessário termos certeza de que $f(x)$ seja mais plana possível, logo precisamos encontrar $f(x)$ com valor de norma mínima ($w^T w$). Isto pode ser formulado como um problema de otimização convexo para minimizar:

$$J(w) = \frac{1}{2} w^T w,$$

sujeito a todos os resíduos terem um valor inferior a um valor ε ; ou, na forma de equação:

$$\forall n : |y_n - (x_n w^T + b)| \leq \varepsilon.$$

É possível que não exista tal função $f(x)$ ou, melhor dizendo, tal hiperplano que satisfaça essas restrições para todos os pontos. Para lidar com restrições que de outra forma seriam inviáveis, podemos introduzir variáveis de folga ξ_n e ξ_n^* para cada ponto. Esta abordagem é semelhante ao conceito de “margem suave” na classificação, porque as variáveis de folga permitem a existência de erros de regressão até o valor de ξ_n e ξ_n^* , mas ainda assim satisfazem as condições exigidas.

A inclusão de variáveis de folga leva à função objetivo, também conhecida como fórmula primal [Vapnik, 1995]:

$$J(w) : \frac{1}{2} w^T w + C \sum_{n=1}^N (\xi_n + \xi_n^*),$$

sujeito a:

$$\forall n : y_n - (x_n w^T + b) \leq \varepsilon + \xi_n^*$$

$$\forall n : (x_n w^T + b) - y_n \leq \varepsilon + \xi_n^*$$

$$\forall n : \xi_n^* \geq 0$$

$$\forall n : \xi_n \geq 0.$$

Na primeira equação, C é um parâmetro de sintonia não-negativo. Nesse contexto, é quem dará o controle sobre o quanto nós toleramos que as observações violem

a margem e o hiperplano. Se $C = 0$, nós não toleramos violação. Maiores valores para C implicam em maior tolerância na violação.

A função de perda linear insensível a ξ ignora erros que estão dentro da distância ε do valor observado, tratando-os como iguais a zero. A perda é medida com base na distância entre o valor observado y e o limite ε .

A teoria sobre o SVM foi proposta inicialmente por Vapnik e Lerner, em 1963, como um método eficiente para a identificação de padrões. Ela possui muitas vantagens na resolução de problemas com uma amostra pequena, não-lineares e com um grande número de dimensões. A Regressão de Vetores de Suporte (SVR, *do inglês*: support-vector regression) foi proposta inicialmente por [Drucker et al. 1996], sendo, justamente, uma técnica de aprendizado supervisionado baseada no conceito de vetores de suporte de Vapnik. O SVR visa reduzir o erro ao determinar o hiperplano e minimizar o intervalo entre os valores previstos e observados.

O SVR costuma realizar uma previsão de desempenho melhor do a Regressão Linear e o K-vizinhos mais próximos (KNN), devido às estratégias de otimização aprimoradas para um amplo conjunto de variáveis. Essa funcionalidade aumenta a capacidade do modelo para previsões, considerando a qualidade dos recursos.

2.3.2.2 Eliminação de recursos recursivos

A eliminação recursiva de recursos (RFE) é um método de seleção de recursos para identificar os principais recursos de um conjunto de dados. O processo envolve o desenvolvimento de um modelo com as características restantes após a remoção repetida das partes menos significativas até que o número desejado de características seja obtido. Embora o RFE possa ser usado com qualquer método de aprendizado supervisionado, alguns modelos se beneficiam mais do RFE do que outros e temos que o SVM é o método de emparelhamento mais popular.

Ao estudar dados biológicos observa-se um alto grau de correlação entre as suas características, as quais não necessariamente podem contribuir para o problema de regressão, simplesmente, por serem, muitas vezes, não informativas. Portanto, é

desejável selecionar um subconjunto de componentes no espaço observado e utilizá-los para regressão de dados. O processo de seleção de um subconjunto de componentes no espaço de recursos é conhecido como seleção de recursos. A seleção de recursos pode resultar em melhorias no desempenho da regressão, bem como na descoberta de biomarcadores e na melhoria da interpretação de dados biológicos. Além disso, em problemas de classificação com na do subtipo tumoral, a descoberta de biomarcadores permite um melhor diagnóstico e tratamento [B Norton I 2012].

O RFE funciona da seguinte forma:

1. Escolha do método: um algoritmo de aprendizado de máquina escolhido (por exemplo, regressão linear, máquina vetorial de suporte, etc.). A escolha do algoritmo depende do problema em questão.
2. Ajuste do modelo: o modelo escolhido deve ser treinado em todo o conjunto de recursos.
3. Classificação das características: a importância de cada característica é avaliada usando um determinado critério, como coeficientes em um modelo linear ou importações de características em um modelo baseado em árvore.
4. Remoção do recurso menos importante: o recurso com a menor pontuação de importância é removido.
5. Repetir as etapas 2-4 até que o número desejado de recursos seja alcançado ou até que o desempenho do modelo pare de melhorar significativamente.
6. Avaliação do desempenho: Avaliar o desempenho do modelo após a eliminação de cada característica. Isso geralmente é feito usando uma abordagem de validação cruzada para evitar sobreajuste ao conjunto de dados específico.
7. Seleção dos recursos ideais: o subconjunto de recursos que resultaram no melhor desempenho do modelo é escolhido.

De modo geral, o método começa construindo um modelo para todo o conjunto de preditores e calculando uma pontuação de importância para cada preditor.

Os preditores menos importantes são removidos, o modelo é reconstruído e as pontuações de importância são calculadas novamente. Na prática, podemos especificar o número de subconjuntos de preditores a serem avaliados, bem como o tamanho de cada subconjunto. Portanto, o tamanho do subconjunto é um parâmetro de ajuste para RFE. O tamanho do subconjunto que otimiza os critérios de desempenho é usado para selecionar os preditores com base nas classificações de importância. O subconjunto ideal é então usado para treinar o modelo final.

A pontuação de importância para cada preditor é inferida a partir de quanto o desempenho do modelo muda quando esse preditor é removido. O método exato para classificação e pontuação depende do algoritmo específico usado para modelagem. Por exemplo, em regressão linear, a pontuação de importância pode estar relacionada à magnitude dos coeficientes atribuídos a cada característica, enquanto em modelos baseados em árvores, pode ser baseada nas importâncias de características calculadas durante o treinamento.

O processo de seleção é reamostrado da mesma forma que o parâmetro de ajuste fundamental de um modelo, como o número de vizinhos mais próximos ou a quantidade de decaimento de peso em uma rede neural. O processo de reamostragem inclui a rotina de seleção de recursos e as reamostragens externas são usadas para estimar o tamanho apropriado do subconjunto.

2.3.2.3 Adição de recursos recursivos

A Adição recursiva de recursos (RFA) é um método de seleção de recursos progressivos, que em oposição ao RFE, funcionavam de trás para frente. O RFA começa com um conjunto de recursos vazio e continua adicionando um recurso por vez até que um critério de parada seja atendido. O método utiliza um classificador SVM como classificador principal para classificar os recursos e, com base nessa classificação, adiciona o recurso que atinge o melhor critério de classificação.

A RFA, em vez de eliminar o pior recurso, elimina o melhor recurso. Sua ideia central é manter características correlacionados mais próximas uma das outras. Por

exemplo, em conjuntos de dados proteômicos, se duas proteínas são altamente correlacionadas, enquanto o RFE pode remover uma delas no meio do processo de eliminação e deixar a outra ser posicionada no início da classificação, a RFA obtém a melhor característica e a coloca no início da classificação. Logo, se uma proteína for muito semelhante a outra, espera-se que o RFA a selecione na(s) iteração(ões) seguinte(s).

2.3.3 Divisão de dados

No aprendizado de máquina, geralmente, para treinar o modelo, os dados são divididos em diferentes subconjuntos, que são usados para treinamento, teste e validação do modelo. Em termos de treinamento de modelo, os parâmetros do modelo (por exemplo, pesos de nó) são derivados e o processo de aprendizado é controlado externamente pelos chamados hiperparâmetros. Enquanto o subconjunto de treinamento é usado para ajustar os parâmetros do modelo, a validação tem o papel de ajustar o hiperparâmetro do modelo e fornecer uma estimativa do nível de habilidade do modelo, o conjunto de dados de teste é usado para estimar o nível de habilidade final do modelo ajustado.

No entanto, estimar o desempenho do modelo nos mesmos dados de treinamento levará a estimativas tendenciosas. Portanto, os conjuntos de dados de validação e teste devem ser mantidos fora do conjunto de dados original. Comumente, esses subconjuntos são escolhidos e divididos manualmente ou aleatoriamente. A primeira destas estratégias requer um grande esforço, enquanto que a última arrisca a perda de uma boa representação de características.

Para lidar com esse problema, usamos validação cruzada para avaliar o desempenho de um modelo de aprendizado de máquina. Na validação cruzada, não dividimos o conjunto de dados em conjuntos de treinamento e teste apenas uma vez. Em vez disso, particionamos repetidamente o conjunto de dados em grupos menores e, em seguida, calculamos a média do desempenho em cada grupo. Dessa forma, reduzimos o impacto da aleatoriedade da partição nos resultados.

A seguir, descrevemos os métodos *Validação cruzada k-fold* e *Leave-one-out (LOO)* *cross-validation (CV)* de reamostragem, os quais utilizamos nesta pesquisa.

2.3.3.1 Validação cruzada k-fold

A validação cruzada k-fold (k-fold cross-validation) é uma técnica comum usada em aprendizado de máquina para avaliar o desempenho de um modelo e mitigar os riscos de variabilidade nos conjuntos de treinamento e teste [Fushiki 2011]. Ela é particularmente útil quando se tem um conjunto de dados limitado.

É um método simples de entender e que geralmente resulta em uma estimativa menos tendenciosa ou menos otimista da habilidade do modelo do que outros métodos, como uma simples divisão de treinamento/teste.

A validação cruzada k-fold visa dividir o conjunto de dados em k partes. Dessa forma, o modelo é treinado k vezes, cada vez usando $k - 1$ partes como conjunto de treinamento e uma parte como conjunto de teste. Isso permite que cada parte do conjunto de dados seja usada como conjunto de teste exatamente uma vez. O desempenho do modelo é então avaliado tirando a média dos resultados das k iterações.

O procedimento geral é o seguinte:

1. Embaralhe o conjunto de dados aleatoriamente;
2. Divida o conjunto de dados em k grupos;
3. Para cada grupo exclusivo:
 - 3.1. Considere o grupo como um conjunto de dados de teste ou de resistência;
 - 3.2. Considere os grupos restantes como um conjunto de dados de treinamento;
 - 3.3. Ajustar um modelo no conjunto de treinamento e avaliá-lo no conjunto de teste;
 - 3.4. Retenha a pontuação da avaliação e descarte o modelo;
4. Resuma a habilidade do modelo usando a amostra de pontuações de avaliação do modelo.

O valor de k é um hiperparâmetro que pode ser ajustado dependendo do tamanho do conjunto de dados e dos recursos computacionais disponíveis. Valores comuns para k incluem 5 e 10 [Fushiki 2011]. Quanto maior o valor de k , mais confiável é a

estimativa do desempenho do modelo, no entanto isso pode aumentar o custo computacional.

2.3.3.2 Leave-one-out cross-validation (LOOCV)

No procedimento *Leave-one-out Cross-Validation (LOOCV)* treinamos nossos n tempos de modelo de aprendizado de máquina onde n é o tamanho do nosso conjunto de dados. A cada vez, apenas uma amostra é usada como conjunto de teste, enquanto o restante é usado para treinar nosso modelo.

Este método tem a vantagem de ser aplicável em amostras pequenas onde outras técnicas, como validação cruzada k-fold, podem apresentar problemas, pois utiliza mais amostras de treinamento em cada iteração. Isso permite que nosso modelo aprenda melhores representações.

2.4 Correlação nos dados

Correlação nos dados é qualquer associação estatística entre um par de variáveis. Existem maneiras distintas de estudar o comportamento conjunto entre duas variáveis. Uma forma bastante funcional é a plotagem de um diagrama de dispersão, onde visualmente se observa a relação entre as variáveis. No entanto, utilizando apenas recursos visuais a única coisa que podemos observar é a relação entre as variáveis, a qual pode apresentar:

- **Correlação positiva:** Quando o valor de uma das variáveis cresce e o da outra, em média, também cresce; ou
- **Correlação negativa:** Quando o valor de uma das variáveis cresce e o da outra, em média, decresce.

Embora abordagens visuais sejam muito úteis e didáticas, é possível combinar informações visuais e quantitativas em nossos diagnósticos.

O Coeficiente de Correlação de Pearson é uma medida de correlação linear entre duas variáveis [Benesty et al., 2009], que apresenta as seguintes propriedades:

- Mede a relação linear entre X e Y ;
- Faixa: $-1 \leq r \leq +1$;
- O coeficiente de correlação é um índice adimensional de força de associação entre duas variáveis ($+$ = associação positiva, $-$ = negativa, 0 = sem associação). Valores próximos de -1 e $+1$ indicam forte correlação linear.

A fórmula do coeficiente de correlação de Pearson é:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (2.1)$$

onde X e Y são as variáveis, r é o coeficiente de correlação entre X e Y , e n é o tamanho das amostras.

O coeficiente de correlação de Pearson, então, resume as características de um conjunto de dados. Especificamente, descreve a força e a direção da relação linear entre duas variáveis quantitativas. Embora as interpretações da força do relacionamento variem entre as variáveis, a tabela abaixo fornece suas regras gerais:

Valor do coeficiente de correlação de Pearson (r)	Força	Direção
Maior que 0,5	Forte	Positivo
Entre 0,3 e 0,5	Moderado	Positivo
Entre 0 e 0,3	Fraco	Positivo
0	Nenhum	Nenhum
Entre 0 e -0,3	Fraco	Negativo
Entre -0,3 e -0,5	Moderado	Negativo
Menos de -0,5	Forte	Negativo

Tabela 1 – Valores de força e direção referente ao coeficiente de correlação de Pearson (r).

2.5 Métricas de Avaliação

Para avaliar o desempenho dos métodos de aprendizado de máquina, são necessárias métricas de avaliação. Neste trabalho, apresentamos métricas de aprendizado supervisionado para problemas de regressão e de classificação. Dentre elas, utilizamos as métricas descritas a seguir.

2.5.1 Métricas de regressão

Para medir o desempenho do nosso método proposto, utilizamos as métricas R-Quadrado (R^2) e o Erro quadrático médio (RMSE).

O R^2 é uma medida estatística que representa a proporção da variância de uma variável dependente que é explicada por uma variável ou variáveis independentes em um modelo de regressão. Ele varia entre 0 e 1, por vezes sendo expresso em termos percentuais. Nesse caso, expressa a quantidade da variância dos dados que é explicada pelo modelo linear. Assim, quanto maior o R^2 , mais explicativo é o modelo linear, ou seja, melhor ele se ajusta à amostra. É definido por:

$$R^2 = \sum_{i=1}^n (x_i - x_i^*)^2, \quad (2.2)$$

onde n é o número de observações, x_i é o valor real e x_i^* o é valor estimado.

O Erro quadrático médio (*do inglês*, Root Mean Squared Error (RMSE)) é a medida que calcula a raiz quadrática média dos erros entre valores observados (reais) e previsões (hipóteses). É definido por:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |x_i - x_i^*|^2}, \quad (2.3)$$

onde n é o número de observações, x_i é o valor real e x_i^* o é valor estimado.

2.5.2 Métricas de classificação

Para medir o desempenho da classificação realizada, utilizamos as métricas precisão, revocação (recall), F-score (F_1) e acurácia.

$$Precisão = \frac{TP}{TP + FP}, \quad (2.4)$$

$$Recall = \frac{TP}{TP + FN}, \quad (2.5)$$

$$F_1 = 2 \times \frac{preciso \times recall}{preciso + recall}, \quad (2.6)$$

$$Acurácia = \frac{TP + TN}{TP + FP + FN + TN}, \quad (2.7)$$

onde TP (verdadeiro positivo) é o número de vezes que o modelo prediz corretamente a classe positiva, FP (falso positivo) é o número de vezes que o modelo prediz incorretamente a classe positiva, FN (falso negativo) é o número de vezes que o modelo prevê incorretamente a classe negativa e TN (verdadeiro negativo) é o número de vezes que o modelo prevê corretamente a classe negativa. Com estes valores é possível montar uma tabela de contingência (ou matriz de confusão), onde as linhas representam as classes reais e as colunas as classes estimadas para cada exemplo classificado.

De modo geral, a precisão verifica de todos os dados classificados como positivos, quantos são realmente positivos. A acurácia indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente. O Recall mede dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas. E o F1-score é a média harmônica entre precisão e recall.

2.6 Considerações finais

Neste capítulo apresentou-se brevemente alguns temas relevantes para esta proposta de dissertação, iniciando pela contextualização dos fatores biológicos, que serve de base para todo o estudo proposto. Posteriormente, descrevemos o problema onipresente nas bases de dados de câncer de mama, a falta de dados, que possui diversas denominações na literatura como dados faltosos, dados faltantes, dados incompletos ou valores ausentes. Em seguida, os conceitos de aprendizagem de máquina e correlação necessários para o desenvolvimento do nosso método de pesquisa. Por fim, descrevemos as métricas de avaliação que auxiliam a entender a eficiência dos métodos os quais utilizamos para realizar a estimação de valores ausentes.

3

TRABALHOS RELACIONADOS

Este capítulo apresenta uma revisão da literatura sobre os estudos relacionados a esta pesquisa. Sendo assim, nele abordamos os métodos que mais se aproximam da nossa abordagem desenvolvida para estimação de dados ausentes em expressões gênicas de câncer de mama.

3.1 Métodos de estimação de valore ausentes

Huang et al. [2018] desenvolveram a análise de célula única via recuperação de expressão (SAVER) que é um método de recuperação de expressão para dados de scRNA-seq baseados em índice molecular único (UMI) o qual seleciona genes ou células que tenham níveis de expressão semelhantes aos do gene ou célula de interesse para realizar a estimação. O método de estimação de preservação de variabilidade para recuperação de expressão (VIPER), desenvolvido por Chen and Zhou [2018], é capaz de inferir progressivamente um conjunto esparsa de células de vizinhança local que são mais preditivos dos níveis de expressão da célula de interesse para estimação.

Sefidian and Daneshpour [2020] desenvolveram o método K-vizinhos mais próximos baseados em correlação, o qual divide os registros incompletos em vários subconjuntos de acordo com o número de atributos ausentes e executa a estimação em ordem crescente com base no subconjunto imputado. Além disso, o método leva em consideração as correlações entre os atributos ao calcular as distâncias entre os registros.

Liu et al. [2020] desenvolveram 10 métodos de estimação baseados em correlação

Tabela 2 – Resumo dos métodos clássicos de Machine Learning.

Métodos clássicos de Machine Learning para estimação de valores omissos			
Autor	Método	Abordagem	Métrica de Avaliação
[Huang et al. 2018]	SAVER	Modelo baseado em Análise Baysiana	Erro quadrático médio (RMSE)
[Chen e Zhou 2018]	VIPER	Beaseado em modelos de regressão espaços não negativos	R-Quadrado
[Liu e Zhang 2020]	CoHiKNN	K-vizinhos mais próximos baseados em correlação	Erro quadrático médio (RMSE)
[Sefidian e Daneshpour 2020]	CMIM	Baseado em correlação	Erro quadrático médio (RMSE)

denominados Métodos de estimação baseados em maximização de correlação (CMIM). Todos esses métodos tentam maximizar a correlação entre uma característica ausente e outras características. A maximização é alcançada selecionando segmentos de dados que possuem fortes correlações. Primeiro, eles selecionam um conjunto base a partir de instâncias completas. Em seguida, os segmentos de dados com fortes correlações são gerados usando o conjunto de base e o restante das instâncias completas. Por fim, cada valor ausente é imputado pela aplicação de modelos lineares aos segmentos de dados descobertos.

A Tabela 2 apresenta os trabalhos relacionados. Em resumo, os métodos de ML são usados para lidar com dados multivariados e multidimensionais, utilizam recursos de forma eficaz e são mais robustos, no entanto a aquisição de dados relevantes é o principal desafio. Em nossos estudos, vimos que o todos os métodos acima usam abordagens similares as que nós propormos a construir nessa pesquisa, em especial o método CMIM por utilizar a técnica de correlação. Porém, todos eles utilizam um conjunto limitado de características, no entanto mais relevantes, conseguindo lidar com o problema da maldição da dimensionalidade.

3.2 Discussão

Os métodos de machine learning são usados para lidar com dados multivariados e multidimensionais, utilizam recursos de forma eficaz e são mais robustos, no entanto a aquisição de dados relevantes é o principal desafio. Em nossos estudos, vimos que

o VIPER e o SAVER realizam a seleção genes/células que tenham níveis de expressão semelhantes aos do gene/célula de interesse para realizar a estimação, ou seja, utilizam um conjunto limitado de genes, no entanto mais relevantes, conseguindo lidar com o problema da maldição da dimensionalidade, descrito na Tabela 2, e obter bons resultados de precisão.

Além disso, vimos que o conjunto de métodos CMIM usam abordagens similares as que nos propomos a construir nessa pesquisa por empregarem o uso da correlação entre um recurso ausente e outros recursos. E, com isso, os segmentos de dados com fortes correlações são gerados usando o conjunto de base e o restante das instâncias completas. Por fim, cada valor ausente é imputado pela aplicação de modelos lineares aos segmentos de dados descobertos.

De modo geral, apesar de um método se mostrar mais eficiente que outro, pesquisas apontam que apesar dos muitos avanços recentes, ainda não há um consenso claro sobre o melhor algoritmo para cada conjunto de dados, visto que existem muitos fatores os quais podem afetar o desempenho de um algoritmo de estimação, como a existência ou não de uma estrutura de correlação nos dados, e provavelmente não há algoritmo que tenha o melhor desempenho em todos os conjuntos de dados. Nenhum dos estudos abordados acima, por exemplo, trabalhou com dados de expressão gênica de câncer de mama.

A partir dessas observações, pretendemos nesta pesquisa além de realizar a seleção de um subconjunto de genes completos a partir do uso da correlação entre os genes completos e os genes com dados omissos, utilizar técnicas de seleção de características mais importantes no subconjunto de genes selecionado após a correlação de modo a obter um subconjunto de genes ainda mais relevante para a estimação e, complementar a isso, fazer o uso dos próprios genes da amostra para realizar a estimação dos seus valores ausentes.

De modo geral, queremos desenvolver um método um método de estimação de expressões ausentes de câncer de mama, que se compare com o estado da arte em termos de eficácia e eficiência, o qual utilize informações de correlação entre genes para realizar a estimação dos genes com valores omissos por meio do uso de métodos

clássicos de machine learning e que com isso lide com o problema da dimensionalidade entre os dados (número de genes x números de amostras).

3.3 Considerações Finais

Neste capítulo, apresentamos uma breve revisão de trabalhos relacionados que abordam o problema de estimação de dados de expressão gênica. Os trabalhos foram divididos de acordo com os métodos de estimação existentes. Ao investigar os trabalhos, percebemos que há uma oportunidade de desenvolver métodos com melhor desempenho do que o estado da arte, que podem alcançar melhores resultados, utilizando o conceito de correlação entre genes. No próximo capítulo, apresentaremos nosso método proposto para estimação de genes com base em correlação.

4

SOLUÇÃO PROPOSTA

Este capítulo apresenta o método proposto para estimação de dados de expressão gênica de câncer de mama baseado em correlação, cada passo da solução é descrito. Quando desenvolvido este método, o objetivo foi verificar se é possível obter resultados melhores do que os do estado da arte fazendo uso da correlação.

4.1 Estimação de expressões gênicas baseada em correlação (EEGBC)

Nesta seção, apresentamos o método de estimação de expressões gênicas baseado em correlação (EGBC), com o qual buscamos estimar os genes com dados faltantes, que melhor classificam o câncer de mama. A Figura 3 resume as etapas que compõem nossa abordagem, discutidas nas próximas subseções.

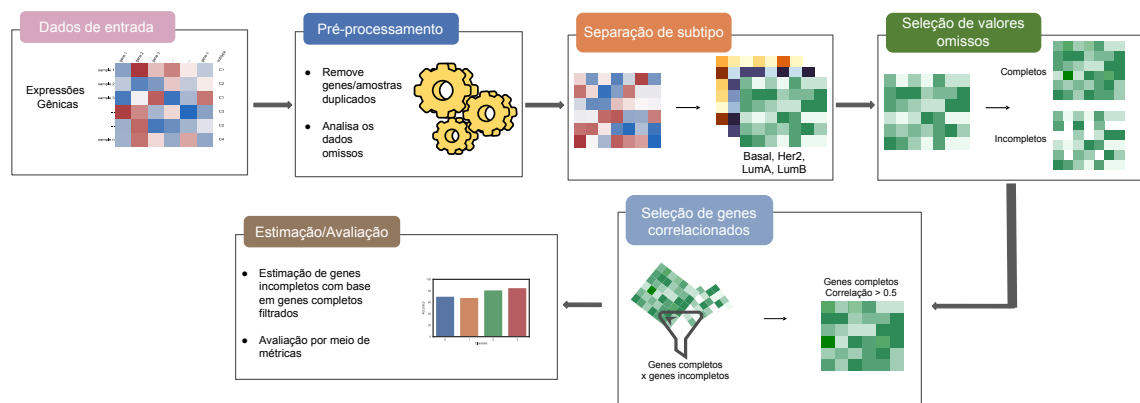


Figura 3 – Método de estimação de expressões gênicas baseado em correlação (EEGBC).

4.1.1 Pré-processamento e separação por subtipo

O passo inicial do nosso método é o pré-processamento dos dados, onde removemos valores genes ou amostras duplicados. Além disso, é onde também verificamos globalmente a quantidade de genes com valores faltantes e não faltantes, de modo a saber quão grande ou não é a incidência de ausência de valores entre esses genes.

Após concluir a etapa de pré-processamento, fazemos a separação do subtipos de câncer de mama. Esta etapa é essencial uma vez que resolvemos realizar a estimação por suptipo de câncer de mama.

4.1.2 Seleção de valores ausentes

Esta etapa consiste em separar os genes completos, dos genes incompletos como mostra a Figura 4. Utilizamos os termos completo e incompleto para nos referirmos, respectivamente, a genes que não possuem expressões gênicas com valores ausentes e genes que possuem uma ou mais expressões gênicas com valores ausentes.

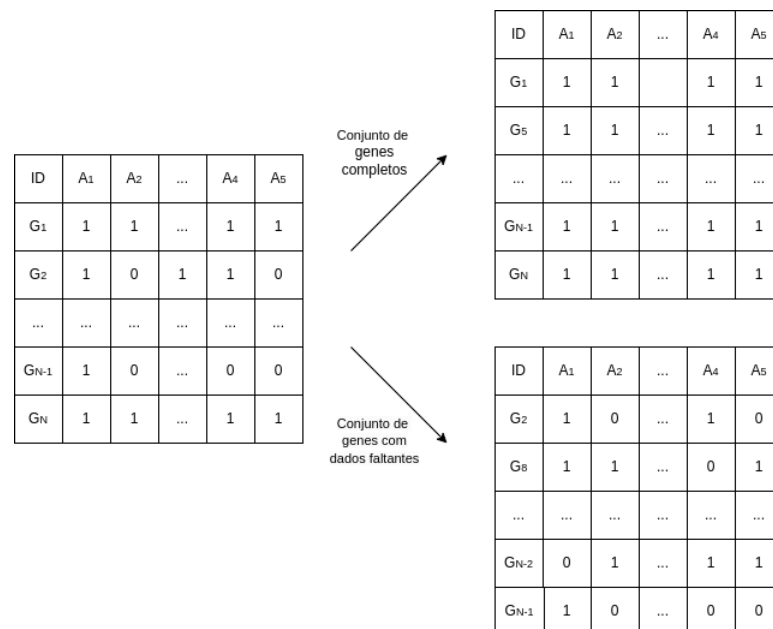


Figura 4 – Divisão do conjunto global de genes em dois conjuntos: um contendo os genes com dados faltantes e outro contendo os genes completos. Sendo assim, valores iguais a 1 representam dados não ausentes e os valores iguais a 0 representam os dados ausentes.

4.1.3 Seleção de genes correlacionados

Nesta etapa, buscamos selecionar um conjunto de genes completos, de modo a utilizá-los na etapa de estimação dos genes com dados faltantes.

Utilizamos os genes completos mais correlacionados aos genes com dados faltantes. Desse modo, compreendemos como se apresentam as correlações entre os genes com dados faltantes e os genes completos, utilizando a técnica de correlação de Pearson, que normaliza as correlações de acordo com os desvios padrão do perfil de expressão de cada gene. Este método, é o mais usado para dados de expressão gênica, devido, ao indício de existência de uma correlação linear nos dados de expressão gênica [D'haeseleer 2005].

A correlação entre os genes pode ajudar a estimar os valores ausentes com base em outros genes que apresentam um padrão de expressão semelhante [D'haeseleer 2005]. A identificação dos genes mais correlacionados é feita por meio da matriz de correlação, que mostra o grau de correlação linear entre cada par de genes. Além disso, a utilização de um subconjunto de genes correlacionados pode ajudar a lidar com a maldição da dimensionalidade. Nesse caso, utilizar apenas um subconjunto de genes pode reduzir a complexidade do modelo e evitar *overfitting*.

Além de utilizar a técnica de correlação, também usamos métodos de seleção de características que são mais relevantes do conjunto de dados de treinamento na previsão da variável de destino. Com isso, visamos encontrar um subconjunto de genes ainda mais eficaz para estimação.

4.1.4 Estimação

A estimação consiste em substituir os valores ausentes por valores estimados. Os métodos de substituição de valores ausentes tem sido estudados há anos.

Nesta etapa, após selecionar somente os genes completos mais correlacionados aos genes com dados faltantes, fazemos a estimação utilizando quatro diferentes métodos que serão descritos na seção 4.2.

4.2 Avaliação da Solução Proposta

Nesta seção, descrevemos a metodologia de avaliação utilizada neste trabalho. Detalhamos as características dos conjuntos de dados usados nos experimentos. Apresentamos os métodos utilizados para realizar as estimações e também explicamos a forma de avaliação utilizada.

4.2.1 Conjunto de Dados e Pré-Processamento

Para avaliar nosso método, utilizamos os genes do conjunto de dados de proteínas do Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Edwards et al. 2015], o qual fornece grande amplitude analítica, pois usa espectrometria de massa para analisar os proteomas de amostras de tumor TCGA anotadas pelo genoma [Mertins et al. 2016]. Essas amostras são divididas em quatro principais subtipos intrínsecos de câncer de mama definidos por mRNA.

A partir desse conjunto, avaliamos o nosso método usando apenas os genes completos presentes na lista de genes do PAM50. Além da importância desses genes para uma classificação mais precisa dos subtipos de câncer de mama, utilizar esses genes dentro de nosso escopo de pesquisa nos ajuda a lidar com a maldição da dimensionalidade, dado que utilizamos um subconjunto de genes que tem uma contribuição relevante para o diagnóstico preciso do câncer e é diretamente formado por uma quantidade de genes condizente com a quantidade de amostras.

A Tabela 3 resume as características dos conjuntos de dados usados nos experimentos.

Tabela 3 – Descrição do Conjunto de Dados.

Descrição	# de genes	Subtipos	# de amostras	Total de amostras
Cptac 2C	23122	Basal	29	117
		LumA	57	
		LumB	17	
		Her2	14	

4.2.2 Estimando dados faltantes

Nesta etapa, selecionamos os genes ausentes da lista do PAM50 e empregamos quatro métodos distintos para realizar a estimação desses genes. Estimar apenas os genes ausentes da lista se deu pela importância desses genes para classificação dos subtipos de câncer de mama, ou seja, por eles serem genes de maior relevância porque garantem uma classificação mais precisa. Com isso, são genes que valem a pena ser estimados.

Como os genes incompletos foram estimados por subtipo usamos, ao nomear os métodos, n para descrevermos de modo abstrato a quantidade de genes completos usados na estimação, pelo fato de a quantidade de genes incompletos mudar de subtipo para subtipo.

Neste processo, de modo a conseguirmos obter um subconjunto ainda mais reduzido de características para a estimação dos genes com dados faltantes, usamos duas abordagens diferentes para selecionar genes relevantes: a Eliminação de Recursos Recursivos (RFE) e a Adição de Recursos Recursivos (RFA), descritas na subseção 2.3.2 [Mendonca-Neto et al. 2022].

Em seguida, dividimos nosso conjunto de dados em treino e teste, fazendo uso da abordagem Leave-One-Out Bootstrap, pelo fato de fornecer uma medida menos tendenciosa do RMSE de teste em comparação com o uso de um único conjunto de teste porque ajustamos repetidamente um modelo a um conjunto de dados que contém $n - 1$ observações. A partir disso, realizamos a estimação usando o algoritmo SVR.

Dessa forma, os métodos de estimação são:

- RFE_n PAM50 e RFA_n PAM50 como baselines. Nesses métodos todos os genes completos do PAM50, por ordem de correlação, passam pelo método RFE ou RFA e um conjunto menor de genes é devolvido para estimar o gene com dado faltante. Por fim, o erro de estimação é calculado. Na subseção 4.2.3 descrevemos as métricas de avaliação que utilizamos. O diagrama 5 descreve esta abordagem.
- RFECA PAM50 e RFACA PAM50 como os métodos principais. Nesses métodos os genes completos são adicionados um a um, por ordem do mais correlacionado para o menos correlacionado com o gene ausente a ser estimado. A medida que os genes são adicionados, passam pelo método RFE ou RFA e um conjunto menor de

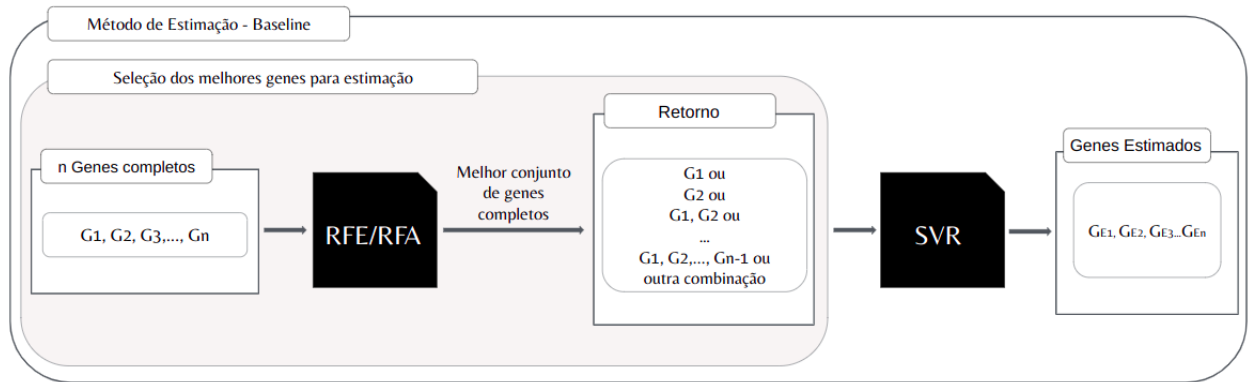


Figura 5 – Diagrama dos métodos RFEn PAM50 e RFAn PAM50.

genes é devolvido para estimar o gene com dado faltante. Por fim, verificamos com base no R2 (2.2) e no RMSE (2.3) qual dos conjuntos estimou com maior precisão o(s) gene(s) incompleto(s). O diagrama 6 descreve esta abordagem.

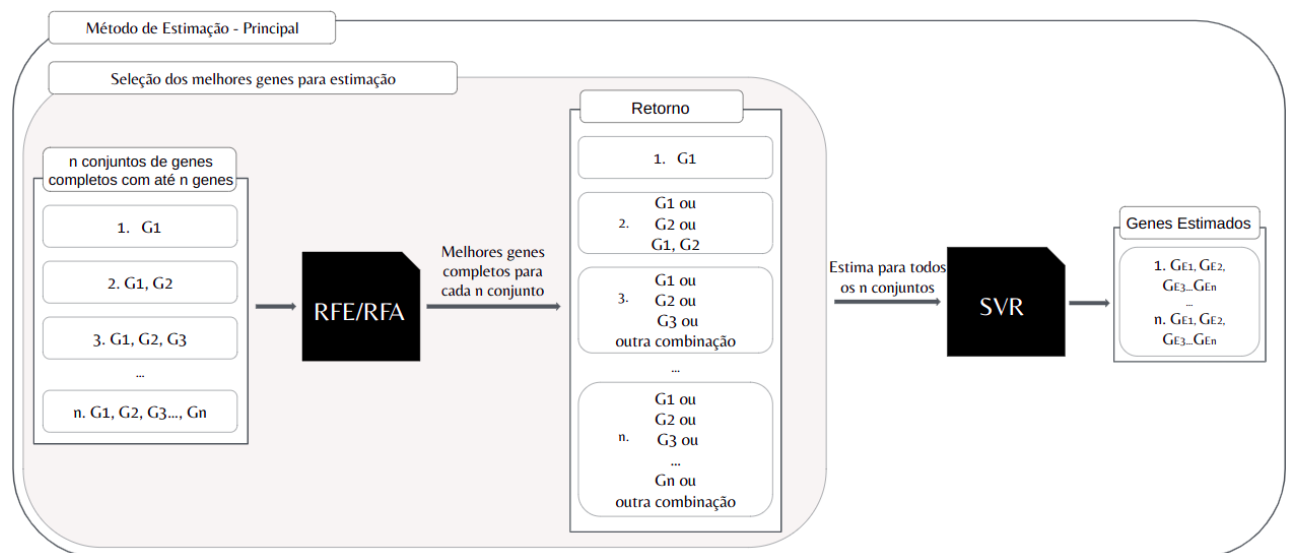


Figura 6 – Diagrama dos métodos RFECA PAM50 e RFACA PAM50.

Por fim, para avaliar a qualidade da nossa estimação fizemos uma classificação multiclasse dos subtipos e comparamos seu resultado utilizando as estimações feitas pelos estimadores K-vizinhos mais próximos (*KNNImpute*) e média (*SimpleImpute*). O classificador usado foi um SVM com kernel linear. No diagrama 7 podemos verificar todo esse processo. A implementação de toda nossa abordagem foi feita usando a linguagem Python.

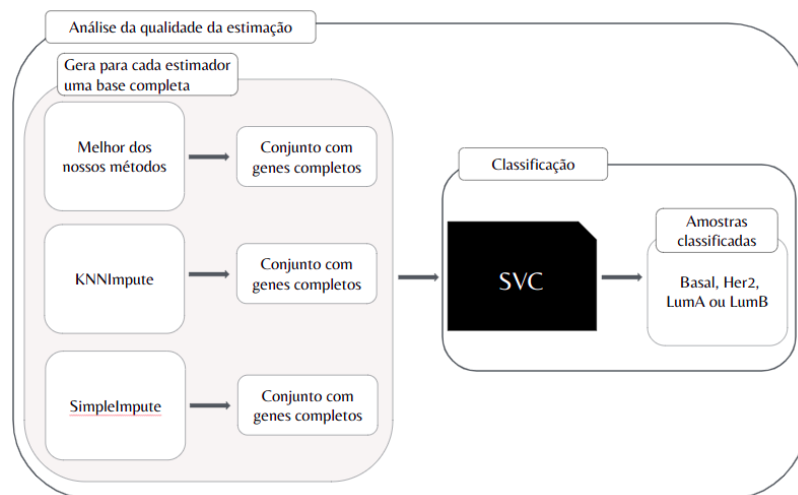


Figura 7 – Processo de classificação.

4.2.3 Avaliação

Nosso conjunto utilizado na etapa de estimação contém o gene com valor ausente a ser estimado, e os genes completos mais relacionados a ele, os quais são as características para o algoritmo que realizará a estimação. Além disso, deve ser formado apenas pelas amostras que possuem valores não ausentes no gene a ser estimado.

Estimamos cada um dos genes com dados faltantes utilizando os genes completos mais correlacionados a cada um deles, sendo assim, vale ressaltar, que a ordem dos genes completos pode mudar de acordo com o gene a ser estimado. Os desempenhos foram avaliados aplicando-se as métricas R^2 e RMSE. Para avaliar os resultados das classificações entre o nosso método de estimação, o KNNImpute e o SimpleImpute aplicamos as métricas precisão, F-score, recall e acurácia.

5

RESULTADOS

Nesta capítulo, descrevemos a análise de valores omissos que realizamos em nossa base de dados e abordamos os resultados obtidos ao aplicarmos cada um dos métodos de correlação nos conjuntos de dados de expressão gênica, realizando a estimação dos valores omissos.

5.1 Análise de valores omissos

Realizando a análise descrita na Tabela 4 vemos que em todos os subtipos de câncer de mama há, pelo menos, 4% de genes com dados ausentes e que o subtipo LumA é o que apresenta mais genes com dados ausentes.

Descrição	Subtipo	Quantidade de genes	Genes com dados completos	Genes com dados faltantes	Total de amostras
Captac 2C - PAM 50	Basal	50	46	4	29
	Her2		47	3	14
	LumA		42	8	57
	LumB		46	4	17

Tabela 4 – Descrição da quantidade de genes com e sem valores ausentes na base Cptac 2C para os genes do PAM50 para cada subtipo de câncer de mama.

Além disso, como mostra a Tabela 5, temos também que o subtipo que apresenta mais omissão é o LumA com uma ausência a nível de expressões gênicas de 16%.

De modo geral, temos que cada gene incompleto possui entre 1% a 3% de expressões ausentes. A seguir estão listados os genes que apresentaram dados omissos para cada uma das amostras:

Descrição	Subtipo	Expressões gênicas	Expressões com valores ausentes (%)	Total de amostras
Captac 2C - PAM 50	Basal	1450	8	92
	Her2	700	6	94
	LumA	2850	16	84
	LumB	850	8	92

Tabela 5 – Descrição da quantidade expressões de genes com e sem valores ausentes na base Cptac 2C para os genes do PAM50 para cada subtipo de câncer de mama.

- Basais: FOXA1, KIF2C, ORC6 e TMEM45B;
- Her2: BIRC5, CEP55 e MAPT;
- LumA: NUF2, CEP55, GRB7, KIF2C, KRT14, MIA, TMEM45B e TYMS;
- LumB: GRB7, KRT5, MIA e ORC6.

No apêndice deste trabalho é possível conferir a lista de genes e a descrição funcional desses genes.

5.2 Estimação de valores omissos

A Tabela 7 demonstra a precisão medida por quatro métodos, sendo RFEn PAM50 e RFAn PAM50 métodos que consideramos como baselines. Apesar de o total de omissão e a quantidade de amostras usadas na estimação, como mostrado na Tabela 5, serem relativamente baixos, são o suficiente para verificamos o quanto nossos métodos conseguem alcançar de precisão.

Do ponto de vista da estimação, pode-se observar que o RFECA PAM50 alcança a maior precisão na estimativa dos genes ausentes de, praticamente, todos os subtipos, pois apenas no subtipo Her2 o método que obteve melhores resultados em todos os genes foi o RFACA PAM50. Ao usar o RFECA PAM50, como mostra a Figura 10, temos que para o gene:

- FOXA1 adicionando 38 características de entrada o RFE retorna um subconjunto de 19 genes mais importantes resultando, após a estimação, nos valores de RMSE e R^2 , respectivamente, de 0,99 e 0,86;

Tabela 6 – Resultados da estimação usando as métricas RMSE e R2 em comparação com os Baselines.

RMSE					R2			
Basal								
Method	RFE46 PAM50	RFECA PAM50	RFA46 PAM50	RFACA PAM50	RFE46 PAM50	RFECA PAM50	RFA46 PAM50	RFACA PAM50
FOXA1	2,23	0,99	2,54	1,94	0,34	0,86	0,14	0,50
KIF2C	0,31	0,18	0,36	0,35	0,76	0,92	0,68	0,70
ORC6	0,39	0,24	0,94	0,72	0,79	0,91	-0,20	0,29
TMEM45B	0,54	0,28	0,67	0,50	0,85	0,96	0,78	0,87
Her2								
Method	RFE47 PAM50	RFECA PAM50	RFA47 PAM50	RFACA PAM50	RFE47 PAM50	RFECA PAM50	RFA47 PAM50	RFACA PAM50
BIRC5	0,91	0,39	0,51	0,32	0,60	0,92	0,87	0,95
CEP55	0,46	0,22	0,39	0,16	0,72	0,93	0,79	0,96
MAPT	2,36	1,30	0,64	0,64	-0,13	0,65	0,91	0,91
LumA								
Method	RFE42 PAM50	RFECA PAM50	RFA42 PAM50	RFACA PAM50	RFE42 PAM50	RFECA PAM50	RFA42 PAM50	RFACA PAM50
NUF2	0,55	0,53	0,60	0,59	0,86	0,87	0,83	0,84
CEP55	0,40	0,40	0,52	0,48	0,93	0,93	0,88	0,90
GRB7	2,05	1,78	2,24	2,10	0,58	0,69	0,50	0,57
KIF2C	0,61	0,52	0,60	0,55	0,88	0,91	0,88	0,90
KRT14	3,9	3,90	5,03	4,55	0,59	0,59	0,33	0,44
MIA	1,32	1,17	1,40	1,28	0,28	0,44	0,20	0,33
TMEM45B	1,51	1,51	2,33	2,07	0,61	0,61	0,09	0,28
TYMS	0,50	0,47	0,67	0,64	0,75	0,78	0,56	0,61
LumB								
Method	RFE46 PAM50	RFECA PAM50	RFA46 PAM50	RFACA PAM50	RFE46 PAM50	RFECA PAM50	RFA46 PAM50	RFACA PAM50
GRB7	0,68	0,45	1,00	0,43	0,96	0,98	0,93	0,98
KRT5	0,68	0,53	1,66	1,61	0,90	0,94	0,44	0,48
MIA	0,71	0,39	1,05	0,23	0,61	0,88	0,18	0,95
ORC6	0,82	0,44	1,25	0,84	0,63	0,89	0,14	0,61

Nota: Melhores resultados em negrito.

- KIF2C adicionando 35 características o RFE retorna um subconjunto de 17 genes mais importantes resultando no valor de RMSE e de R^2 , respectivamente, de 0,18 e 0,92;
- ORC6 para 40 características o RFE retorna um subconjunto de 20 genes mais importantes obtendo um valor de RMSE e de R^2 , respectivamente, de 0,24 e 0,91;
- TMEM45B adicionando 30 características o RFE retorna um subconjunto de 15 genes mais importantes obtendo um valor de RMSE e de R^2 , respectivamente, de 0,28 e 0,96.

No caso, do suptipo Her2, como citado anteriormente, o método RFACA PAM50 al-

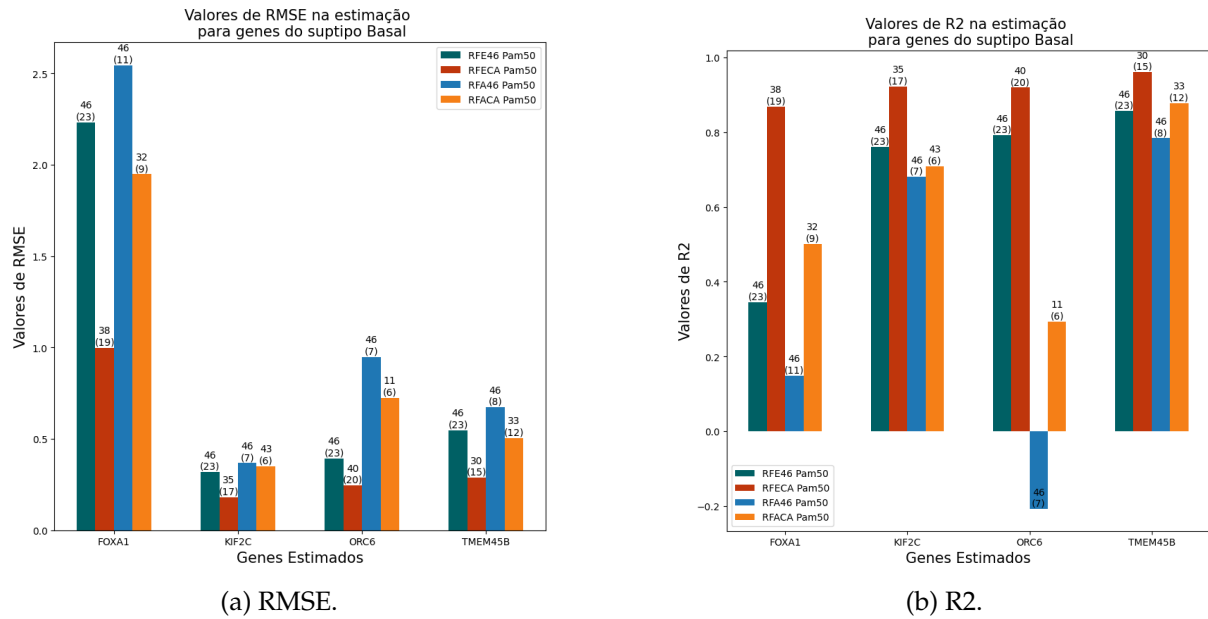


Figura 8 – Resultados (a) RMSE e (b) R2 para estimação dos genes ausentes do subtipo Basal dado cada um dos métodos.

cançou a maior precisão na estimativa dos genes ausentes. Ao usar este método, como mostra a Figura 9, temos que para o gene:

- BIRC5 adicionando 18 características de entrada o RFA retorna um subconjunto de 9 genes mais importantes resultando, após a estimação, nos valores de RMSE e R^2 , respectivamente, de 0,16 e 0,95;
- CEP55 adicionando 40 características o RFA retorna um subconjunto de 20 genes mais importantes resultando no valor de RMSE e de R^2 , respectivamente, de 0,16 e 0,96;
- MAPT para 12 características o RFA retorna um subconjunto de 6 genes mais importantes obtendo um valor de RMSE e de R^2 , respectivamente, de 0,64 e 0,91.

Analogamente ao subtipo Basal, de modo geral, os subtipos LumA e LumB apresentaram melhores resultados nas estimções dos seus genes com o método RFECA PAM50, como podemos observar, respectivamente, nas Figuras 10 e 11.

De modo geral, os resultados apresentados mostram que existem combinações de genes que são mais eficientes na estimação de cada um dos genes com dados faltantes a partir do uso da correlação atrelada a uma seleção de características mais importantes por meio do RFE e também que existe uma quantidade específica de genes capaz

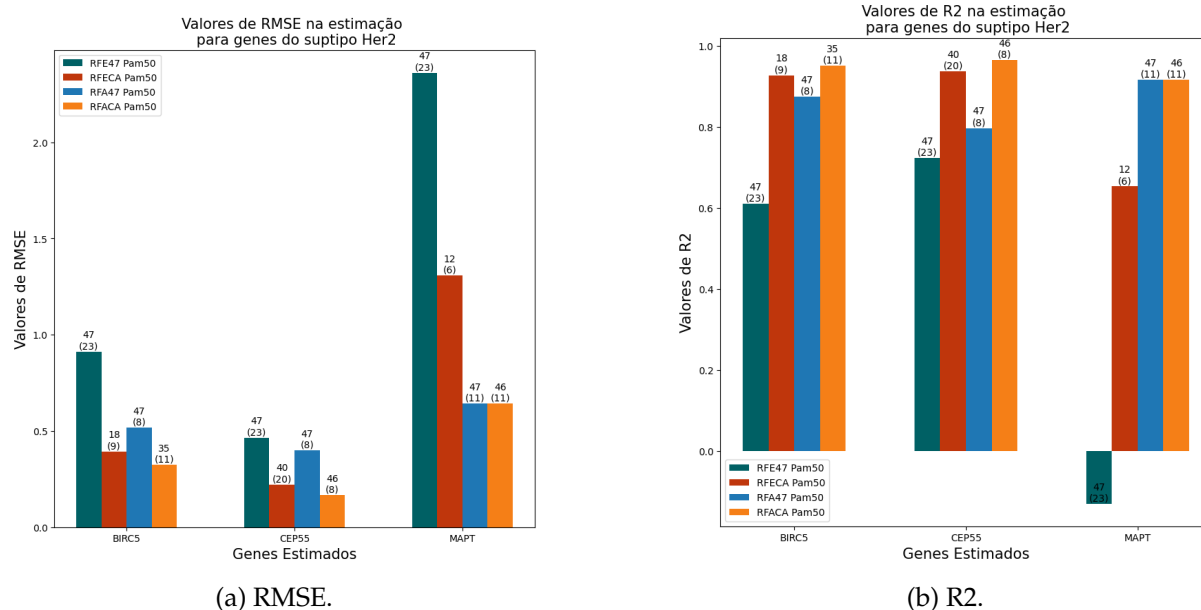


Figura 9 – Resultados (a) RMSE e (b) R2 para estimação dos genes ausentes do suptipo Her2 dado cada um dos métodos.

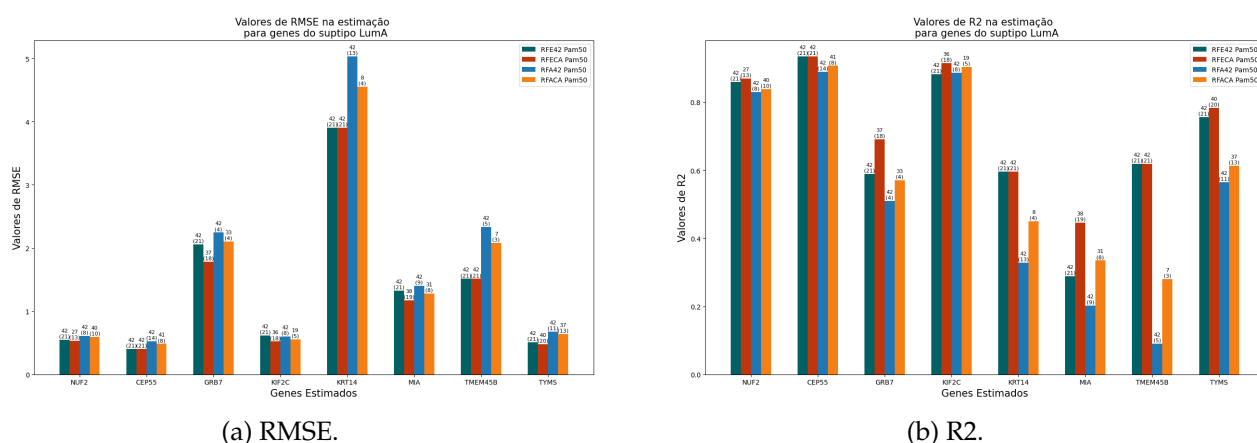


Figura 10 – Resultados (a) RMSE e (b) R2 para estimação dos genes ausentes do suptipo LumA dado cada um dos métodos.

de gerar um resultado ainda mais preciso do que quando utilizamos todos os genes do conjunto. Ademais, podemos observar que o método RFACA PAM50 apresenta o segundo melhor resultado. Com base neste contexto, escolhemos fazer a estimação dos genes ausentes por suptipo utilizando o RFECA PAM50.

Dentro deste contexto vale ressaltar que os métodos de regressão apresentam resultados muito mais precisos quando as características usadas para previsão da classe alvo apresentam forte correlação com essa classe, no entanto é preciso lidar com a alta correlação entre as próprias características. Sendo assim, temos que métodos com RFE e RFA tratam a existência de colinearidade entre as características evitando o uso de

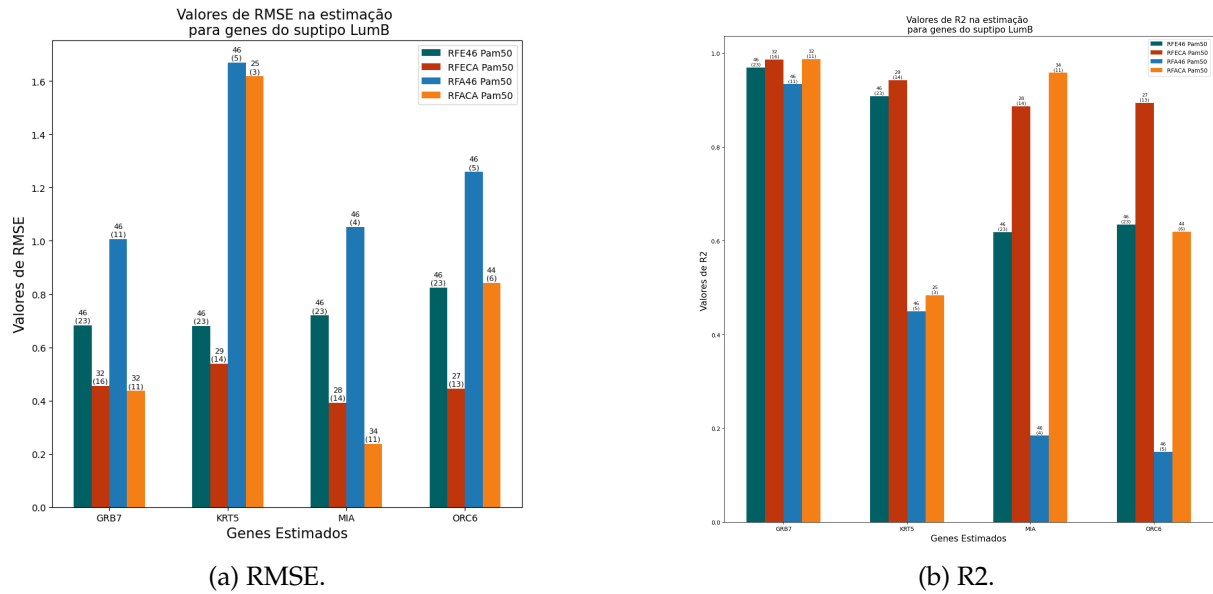


Figura 11 – Resultados (a) RMSE e (b) R2 para estimação dos genes ausentes do suptipo LumB dado cada um dos métodos.

variáveis altamente correlacionadas entre si o que permite, justamente, utilizar um conjunto muito mais significativo de características de entrada para previsão do target do nosso modelo.

5.3 Interseção de genes

Como descrito na subseção anterior, para cada um dos subtipos de câncer de mama, genes completos foram selecionados para estimar os genes incompletos e, além disso, para cada gene incompleto estimado um conjunto diferente de genes completos foi utilizado.

Resolvemos analisar a interseção dos genes completos selecionados dado cada um dos métodos usados para estimar os genes incompletos em cada subtipo e dado o melhor dos métodos de estimação em cada subtipo, com o intuito de verificar a existência de genes completos comuns na interseção dos genes estimados.

As Figura 12, 14, 16 e 18 expõe as visualizações das interseções dos genes selecionados dado cada um dos métodos que utilizaram apenas os genes completos do conjunto PAM50 para realizar a estimação, respectivamente, para os subtipos Basal, Her2, LumA e Lum B. Além disso, as Figura 13, 15, 17 e 19 mostram as interseções de

genes no método RFECA PAM50 para cada um dos genes estimados, respectivamente, para os subtipos Basal, Her2, LumA e Lum B.

Análise para o subtipo Basal:



Figura 12 – Genes completos escolhidos em cada um dos métodos de estimação por gene incompleto no subtipo Basal. Visão de interseção na qual as cores azuis representam os genes selecionados.

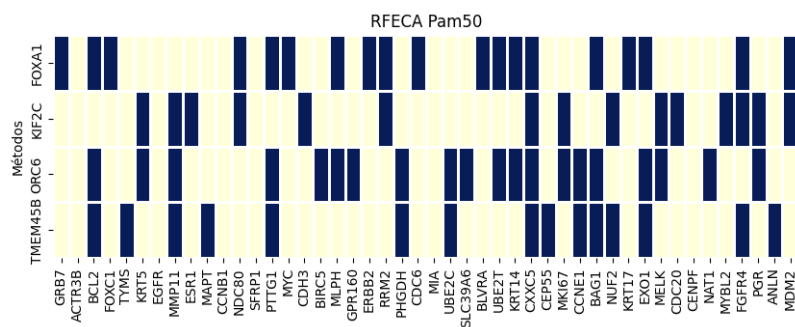


Figura 13 – Genes completos escolhidos em cada um dos genes estimados dado o método RFECA PAM50 no subtipo Basal. Visão de interseção na qual as cores azuis representam os genes selecionados.

Analisando as interseções dos genes completos usados para estimar os genes incompletos do subtipo Basal, temos na interseção do gene:

- FOXA1 os genes completos GRB7, BCL2, NDC80 e MYC.
- KIF2C os genes completos KRT5, MMP11, RRM2 e CDC20.
- ORC6 os genes completos BIRC5, UBE2C, CXX25, CCNE1 e PGR.
- TMEM45B os genes completos MAPT, PHGDH e UBE2C.

Por meio da Figura 13 vemos que apenas o gene CXXC5 aparece em todos os conjuntos de genes escolhidos para os genes estimados no melhor dos métodos.

Análise para o subtipo Her2:

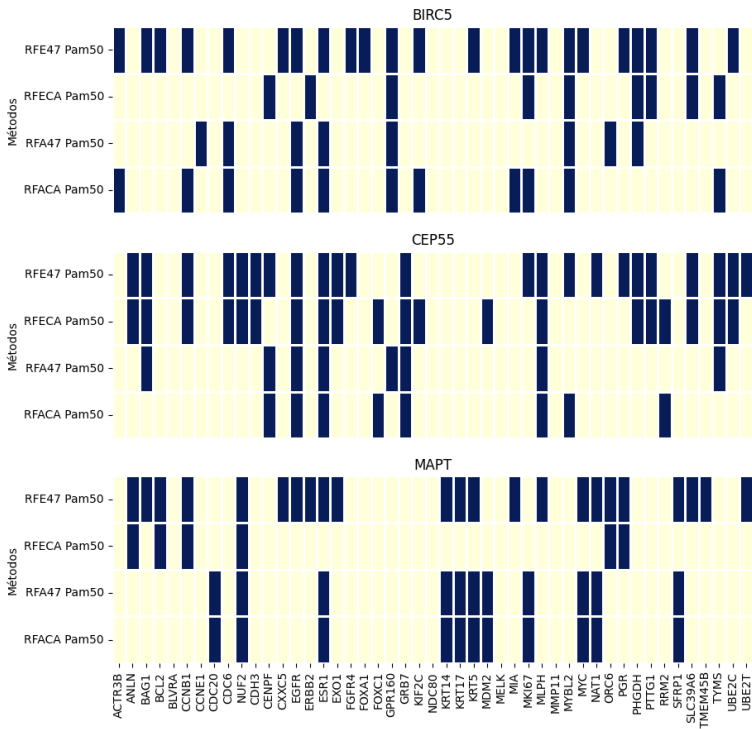


Figura 14 – Genes completos escolhidos em cada um dos métodos de estimação por gene faltante no subtipo Her2. Visão de interseção na qual as cores azuis representam os genes selecionados.



Figura 16 – Genes completos escolhidos em cada um dos métodos de estimação por gene faltante no subtipo LumA. Visão de interseção na qual as cores azuis representam os genes selecionados.

- NUF2 os genes completos ANLN, CDC20, ESR1 e NDC80.
- CEP55 os genes completos CDC20, CDC6, EXO1, GPR160, MKI67 e PTTG1.
- GRB7 os genes completos ERBB2, MAPT, PGR e SFRP1.
- KIF2C os genes completos CDC20 e CDC6.
- KRT14 os genes completos BAG1 e KRT17.
- MIA o gene completo PHGDH.

- TMEM45B nenhum gene.
- TYMS os genes completos CCNE1, FOXA1, GRP160, NDC80 e PTTG1.

Por meio da Figura 17 vemos que não aparece nenhum gene na interseção para os genes estimados no melhor dos métodos.

Análise para o subtipo LumB:

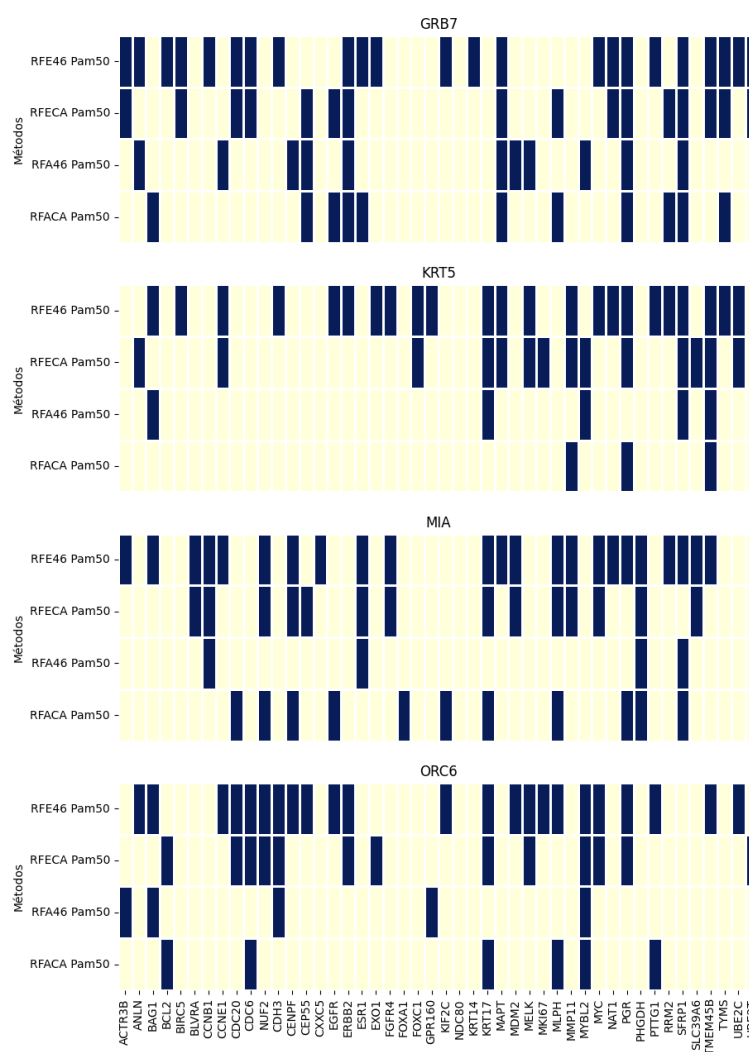


Figura 18 – Genes completos escolhidos em cada um dos métodos de estimação por gene faltante no subtipo LumB. Visão de interseção na qual as cores azuis representam os genes selecionados.

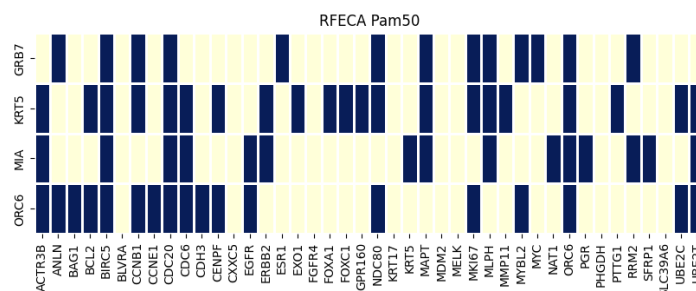


Figura 19 – Genes completos escolhidos em cada um dos genes estimados dado o método RFECA PAM50 no subtipo LumB. Visão de interseção na qual as cores azuis representam os genes selecionados.

Analisando as interseções dos genes completos usados para estimar os genes incompletos do subtipo LumB, temos na interseção do gene:

- GRB7 os genes completos ERBB2, MAPT, PGR e SFRP1.
- KRT5 o gene completo MEM45B.
- NIA o gene completo PHGDH.
- MYBL2 o gene completo PHGDH.

Por meio da Figura 19 vemos que apenas os genes BIRC5, CDC20 e ORC6 aparecem em todos os conjuntos de genes escolhidos para os genes estimados no melhor dos métodos.

Discussão das análises:

Por meio dessa análise, vemos que costuma existir genes completos comuns em todos esses métodos e, respectivamente, também em todos os genes com dados faltantes dado o melhor dos métodos. Além disso, é notável que entre os genes completos usados para estimar genes incompletos associados ao câncer de mama, como FOXA1 ou o GRB7, existe no mínimo um outro gene associado ao câncer de mama. No apêndice deste trabalho é possível conferir quais genes estão associados diretamente ao câncer de mama e quais não estão.

Ademais, essa análise nos ajuda também a verificar as escolhas de genes em cada um dos métodos nos permitindo, em análises futuras, refletir mais a respeito da relação entre esses genes.

Na literatura, estudos de expressão gênica demonstraram heterogeneidade no câncer de mama, e essa heterogeneidade tem base molecular [Mayrovitz 2022]. No entanto, ainda não foi determinado se essas características moleculares influenciam inequivocamente o manejo clínico do câncer de mama.

Fatores clínicos e patológicos tradicionais, como idade, grau histológico, tipo de tumor, tamanho do tumor e receptores hormonais, são comumente usados para classificar pacientes em grupos de risco para receber terapia hormonal, radioterapia e/ou quimioterapia adjuvantes mais adequadas para tratar o câncer de mama [Masood 2016].

Apesar de esses fatores serem capazes de identificar com precisão as melhores terapias para os pacientes com base em estudos de acompanhamento de longo prazo, eles são limitados na sua capacidade de fornecer classificações, de fato, confiáveis em todos os pacientes. Um estudo demonstrou que até 30% das mulheres com câncer de mama com linfonodos negativos morrem da doença, independentemente da terapia adjuvante, e 70% sobrevivem sem terapia adjuvante [Henderson et al. 1988].

Como a heterogeneidade no câncer de mama não pode ser capturada pelos fatores prognósticos tradicionais, é necessário procurar fatores que possam complementar os fatores prognósticos tradicionais na segregação de pacientes que necessitam de terapia adjuvante e na previsão da resposta clínica às modalidades terapêuticas disponíveis [Mayrovitz 2022].

Com base nisto, a identificação de genes específicos associados ao câncer de mama pode revelar heterogeneidade adicional entre os subtipos de câncer de mama e pode se tornar relevante no desenvolvimento de medicamentos mais específicos para cada subtipo, gerando terapias que proporcionem maior expectativa de vida para os pacientes.

Dessa forma, pensamos que a análise de interseção pode auxiliar os pesquisadores a compreenderem melhor a relação entre os genes nos subtipos e, a partir disso, ajudá-los, justamente, a identificar genes específicos associados ao câncer de mama.

Dado que cada subtipo de câncer apresenta características natas a eles, observar os genes individualmente por subtipo pode colaborar para uma compreensão mais específica desses genes e do próprio subtipo. Vale ressaltar que conhecer o subtipo do câncer de mama pode auxiliar a prática clínica a estabelecer o melhor tratamento para cada paciente.

5.4 Classificação de subtipos

Para analisar a qualidade da estimação do nosso método realizamos a classificação dos subtipos de câncer de mama dadas as base com genes completos geradas pelo nosso método e pelos outros dois métodos de estimação. Para isto, executamos três experimentos:

1. Experimento 1: Classificamos os subtipos utilizando todos os 50 genes do PAM50.
2. Experimento 2: Selecionamos como características os 13 genes estimados para classificar os subtipos de câncer.
3. Experimento 3: Selecionamos os 5 genes que mais continham valores ausentes (entre os 13 estimados) e classificamos os subtipos de câncer de mama.

A Tabela 7 apresenta os resultados das classificações feitas usando as bases completas geradas pelas duas abordagens, anteriormente citadas, e as geradas pelos métodos KNNImpute e SimpleImpute. Em todos os experimentos, utilizamos a validação cruzada k -fold para verificar o comportamento da classificação dados diferentes conjuntos de treinamento/teste com o intuito de observar de maneira mais abrangente a qualidade da estimação e evitar conclusões tendenciosas utilizando apenas um conjunto de teste/treino. Dessa forma, realizamos uma análise utilizando 5 e outra 10 folds.

Observamos, por meio da Tabela 7, que no experimento 1 os resultados das classificações foram similares tanto com $k=5$ quanto com $k=10$. No experimento 2 os melhores resultados foram utilizando as bases completas geradas pelo nosso método e pelo método KNNImpute. E no último experimento vemos que a classificação se comportou melhor utilizando a base completa gerada pelo nosso método.

Tabela 7 – Resultados da classificação feita a partir das bases geradas pelos métodos de estimação, usando as métricas precisão, recall, acurácia e F_1 . Análise observando a validação cruzada k-fold com 5 e com 10 folds.

$k=5$					$k=10$			
Experimento 1								
Método	Precisão	Recall	F_1	Acurácia	Precisão	Recall	F_1	Acurácia
RFECA PAM50	0,88	0,91	0,89	0,92	0,88	0,91	0,88	0,91
KNNImpute	0,88	0,91	0,89	0,92	0,88	0,91	0,88	0,91
SimpleImpute	0,88	0,91	0,89	0,93	0,87	0,91	0,87	0,91
Experimento 2								
RFECA PAM50	0,79	0,80	0,77	0,83	0,82	0,81	0,79	0,85
KNNImpute	0,79	0,80	0,77	0,83	0,82	0,82	0,80	0,85
SimpleImpute	0,77	0,78	0,75	0,81	0,77	0,77	0,74	0,82
Experimento 3								
RFECA PAM50	0,70	0,74	0,69	0,80	0,70	0,71	0,68	0,80
KNNImpute	0,69	0,73	0,69	0,80	0,69	0,69	0,66	0,76
SimpleImpute	0,67	0,70	0,66	0,77	0,69	0,69	0,66	0,74
Nota: Melhores resultados em negrito.								

Como a quantidade de expressões gênicas omissas nos genes estimados não é tão grande, resolvemos reduzir o conjunto de características utilizadas para classificar os subtipos de câncer de mama. Por este motivo resolvemos realizar os experimentos 2 e 3, nos quais os subconjuntos de genes usados como características da classificação foram os estimados. Isso nos permite refletir sobre a qualidade da estimação e quanto ela pode refletir na precisão da classificação dos subtipos de câncer de mama. Além disso, a escolha de fazer o k-fold nos ajudou a observar o quanto melhorava a classificação dados diferentes conjuntos de treino/teste e, assim, evitar chegar em conclusões precipitadas a respeito da melhor abordagem.

5.5 Considerações Finais

Técnicas de geração e análise de expressão gênica, como a tecnologia de microarray, geralmente sofrem com o problema de valor ausente devido a várias razões experimentais, tais quais as apresentadas na Seção 2.2. Métodos de análise para caracterização do câncer de mama e classificação de subtipos exigem um conjunto de dados completo, dessa forma a estimação de valor ausente é uma etapa importante de pré-processamento

na análise desses dados.

Neste capítulo, apresentamos uma abordagem para estimação de genes de câncer de mama com dados faltantes com base na lista de genes do PAM50. Utilizamos quatro diferentes métodos para realizar a estimação, os quais faziam uso de um procedimento distinto para analisar se há diferença entre eles na tarefa de estimação. Duas métricas de avaliação foram empregadas para obter o panorama de como os métodos estimaram os genes com dados faltantes.

Concluimos que o método RFECA PAM50 é o melhor estimador dentre os testados, apresentando os menores erros em praticamente todas as métricas utilizadas para avaliação e na estimação de todos os genes do PAM50 com dados faltantes.

No entanto, é notável que também o método RFACA PAM50 pode melhorar o desempenho até certo ponto e produzir resultados competitivos em comparação aos métodos RFE46 PAM50 e RFA46 PAM50. Em linhas gerais, percebemos por meio do uso do RFE e RFA, completar ao uso da correlação, que, adicionando os genes um a um por ordem do mais ao menos correlacionado, conseguimos identificar conjuntos de genes que tornam a estimação mais precisa.

Além disso, identificamos que certos genes parecem apresentar características intrínsecas as quais os tornam bons para estimar todos os genes com dados faltantes. E vimos que análises mais profundas, como a de interseção, podem ajudar os pesquisadores a identificar genes de câncer de mama e isto pode contribuir para uma gama de fatores, tais quais a criação de medicamentos mais benéficos para os pacientes que sofrem com um determinado tipo de câncer ou mesmo ajudar em uma melhor caracterização dos subtipos o que permite, por sua vez, o direcionamento de melhores terapias clínicas para os pacientes.

Ademais, realizar a classificação nos permitiu refletir a cerca da qualidade da nossa estimação. Com ela vimos que nosso método alcança, de fato, bons resultados em relação a precisão da estimação dos genes ausentes. No fim, a qualidade da estimação contribui para a qualidade da classificação.

6

CONSIDERAÇÕES FINAIS

Nesta dissertação, apresentamos um método de estimação de genes baseado em coorelação. Descrevemos detalhadamente a solução proposta, cujo desenvolvimento foi o foco desta pesquisa. Na Seção 5.1, descreveremos as principais contribuições obtidas, na Seção 5.2 indicamos os próximos passos da pesquisa e na Seção 5.3 apresentamos o desenvolvimento futuro desta proposta.

6.1 Contribuições

Este trabalho apresentou uma abordagem para estimação de genes de câncer de mama com dados faltantes com base na lista de genes do PAM50. Utilizamos quatro diferentes métodos para realizar a estimação, os quais faziam uso de um procedimento distinto para analisar se há diferença entre eles na tarefa de estimação. Duas métricas de avaliação foram empregadas para obter o panorama de como os métodos estimaram os genes com dados faltantes.

Para realizar a estimação, nós nos concentramos na seleção de um conjunto mais representativo de genes com dados faltantes a serem estimados, ou seja, aqueles que caracterizam melhor o câncer de mama e que podem ser utilizados após a estimação para classificar de modo mais preciso os subtipos de câncer.

Em princípio, pode-se emprestar informações entre genes ou amostras para estimar um valor ausente, fazendo o uso de correlação. No entanto as vantagens de utilizar informações entre genes e não entre amostras é que, normalmente, um novo

paciente surge e não um novo gene. Dessa forma, como estamos utilizando aprendizagem supervisionada, ao chegar uma nova amostra podemos apenas incluir ela para estimá-la.

Como resultados, percebemos que o método RFECA Pam50 obteve resultados melhores que os demais e que o RFACA Pam50 foi o método que apresentou os segundos melhores resultados. Verificamos, com isso, que utilizar um subconjunto de genes relevantes, com características próximas aos genes a serem estimados, faz com que a estimação alcance melhores resultados.

Ademais, sabe-se que a ausência de dados afeta negativamente a qualidade da classificação do câncer de mama, podendo degradar drasticamente o desempenho do modelo se manuseados incorretamente. E a estimação fornece uma oportunidade para resolver o problema, pois permite que os pesquisadores obtenham bases de dados completas e isto facilita a tarefa de classificação.

Por fim, vale citar que publicamos um artigo referente a este trabalho no Seminário Integrado de Software e Hardware (SEMISH) da Sociedade Brasileira de Computação (SBC). Sendo assim, o artigo foi publicado nos anais do evento, ou seja, está disponível para acesso na biblioteca digital da Sociedade Brasileira de Computação.

6.2 Próximos passos

Ademais, além dos próprios resultados finais referentes as estimações avaliadas pelas métricas, pretendemos inserir em nossa pesquisa outras bases de dados de câncer de mama que contém uma quantidade maior de expressões gênicas ausentes ou mesmo inserir valores ausentes artificialmente de modo a verificar com maior precisão o quanto a estimação de dados contribui para a tarefa de classificação dos genes.

REFERÊNCIAS

- ABDELLA, M.; MARWALA, T. The use of genetic algorithms and neural networks to approximate missing data in database. In: *IEEE 3rd International Conference on Computational Cybernetics, 2005. ICCC 2005*. [S.l.: s.n.], 2005. p. 207–212. 17
- B NORTON I, T. A. A. N. G. Recursive feature elimination for brain tumor classification using desorption electrospray ionization mass spectrometry imaging. *Annu Int Conf IEEE Eng Med Biol Soc.*, p. 5258–5261, 2012. 40
- CHEN, M.; ZHOU, X. Viper: variability-preserving imputation for accurate gene expression recovery in single-cell rna sequencing studies. *Genome biology*, BioMed Central, v. 19, n. 1, p. 1–15, 2018. 49
- DATA, M. C. Secondary analysis of electronic health records. In: . [S.l.: s.n.], 2016. 27
- D’HAESELEER, P. How does gene expression clustering work? *Nature biotechnology*, Nature Publishing Group, v. 23, n. 12, p. 1499, 2005. 54
- DRUCKER, H. et al. Support vector regression machines. *Advances in neural information processing systems*, v. 9, 1996. 39
- EDWARDS, N. J. et al. The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, ACS Publications, v. 14, n. 6, p. 2707–2713, 2015. 55
- ENDERS, C. K. Applied missing data analysis. In: . [S.l.: s.n.], 2010. 28, 29
- FUSHIKI, T. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, v. 21, p. 137–146, 04 2011. 43
- GENESCARDS. *The Human Gene Database*. Disponível em: <<https://www.genecards.org/>>. 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90
- GIBBS, R. A. The human genome project changed everything. *Nature Reviews. Genetics*, v. 21, p. 575 – 576, 2020. 17
- HENDERSON, I. C. et al. Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. an overview of 61 randomized trials among 28,896 women. *The New England Journal of Medicine*, v. 319, p. 1681–1692, 1988. Disponível em: <<https://api.semanticscholar.org/CorpusID:80120566>>. 71
- HUANG, M. et al. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, Nature Publishing Group US New York, v. 15, n. 7, p. 539–542, 2018. 49

- INCA. Instituto nacional do câncer - estatísticas. 2021. Disponível em: <<https://www.inca.gov.br/assuntos/cancer-de-mama>>. 18
- KANG, H. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, v. 64, p. 402 – 406, 2013. 18
- KUNDAJE, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature*, Nature Publishing Group, v. 518, n. 7539, p. 317–330, 2015. 26
- LIBERZON, A. et al. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, v. 27 12, p. 1739–40, 2011. 27
- LIEW, L. N.-F. A.; YAN, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in bioinformatics*, v. 12 5, p. 498–513, 2011. 20, 25, 30
- LITTLE, R.; RUBIN, D. Statistical analysis with missing data. *Hoboken, NJ: Wiley.*, 2019. 29
- LIU, X. L. X.; ZHANG, L. A hierarchical missing value imputation method by correlation based k nearest neighbors. In: . [S.l.]: Springer International Publishing, 2020. ISBN 978-3-030-29516-5". 49
- LONSDALE, J. et al. The genotype-tissue expression (gtex) project. *Nature genetics*, Nature Publishing Group, v. 45, n. 6, p. 580–585, 2013. 26
- MASOOD, S. Breast cancer subtypes: Morphologic and biologic characterization. *Women's health (London, England)*, v. 12, 01 2016. 71
- MAYROVITZ, H. N. *Breast Cancer*. Brisbane, Australia: Exon Publications, Brisbane, Australia, 2022. v. 1. ISBN 978-0-6453320-3-2. 71
- MENDONCA-NETO, R. et al. A gene selection method based on outliers for breast cancer subtype classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE, v. 19, n. 5, p. 2547–2559, 2021. 19, 25
- MENDONCA-NETO, R. et al. Classification of breast cancer subtypes: A study based on representative genes. *Journal of the Brazilian Computer Society*, v. 28, n. 1, p. 59–68, 2022. 56
- MERTINS, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, Nature Publishing Group UK London, v. 534, n. 7605, p. 55–62, 2016. 55
- MONARD, M. C. et al. *Uma Introdução ao Aprendizado Simbólico de Máquina por Exemplos*. São Carlos, SP, Brasil, 1997. 32
- NEWMAN, J. C. et al. Ketogenic diet reduces midlife mortality and improves memory in aging mice. *Cell metabolism*, Elsevier, v. 26, n. 3, p. 547–557, 2017. 26
- RAKHA, E. A. et al. Expression profiling technology: its contribution to our understanding of breast cancer. *Histopathology*, v. 52, 2007. 24
- ROUKOS, D. H.; KU, C. S. Clinical cancer genome and precision medicine. *Annals of Surgical Oncology*, v. 19, p. 3646–3650, 2012. 24

- RUSSEL, S.; NORVIG, P. “artificial intelligence – a modern approach”, second edition, pearson education, 2003. In: . [S.l.: s.n.], 2015. 32
- SEFIDIAN, A. M.; DANESHPOUR, N. Estimating missing data using novel correlation maximization based methods. *Applied Soft Computing*, Nature Publishing Group US New York, v. 91, 2020. 49
- SINGH, K. et al. Mechanisms and measurement of changes in gene expression. *Biological Research For Nursing*, v. 20, p. 109980041877216, 04 2018. 18
- TAN, A. C.; GILBERT, D. Ensemble machine learning on gene expression data for cancer classification. *Applied bioinformatics*, v. 2, p. S75–83, 02 2003. 26
- TEFFERI, A. et al. Primer on medical genomics. part iii: Microarray experiments and data analysis. *Mayo Clinic proceedings*, v. 77 9, p. 927–40, 2002. Disponível em: <<https://api.semanticscholar.org/CorpusID:16795158>>. 25
- VOLGIN, D. V. Chapter 17 - gene expression: Analysis and quantitation. In: VERMA, A. S.; SINGH, A. (Ed.). *Animal Biotechnology*. San Diego: Academic Press, 2014. p. 307–325. ISBN 978-0-12-416002-6. Disponível em: <<https://www.sciencedirect.com/science/article/pii/B9780124160026000171>>. 17
- XIANG, Q. et al. Missing value imputation for microarray gene expression data using histone acetylation information. *BMC Bioinformatics*, v. 9, p. 252 – 252, 2008. 19, 27
- YERSAL, O.; BARUTÇA, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology*, v. 5 3, p. 412–24, 2014. 24

APÊNDICE

Descrição funcional dos genes do PAM50

Tabela Complementar I – Descrição funcional dos genes do PAM50 diretamente relacionados ao câncer de mama.

Gene	Descrição
ACTR3B	Este gene codifica um membro das proteínas relacionadas à actina (ARP), que formam complexos multiproteicos e compartilham 35-55% de identidade de aminoácidos com a actina convencional. A proteína codificada por este gene pode ter um papel regulador no citoesqueleto de actina e induzir alterações na forma e motilidade das células. ACTR3B (Proteína Relacionada à Actina 3B) é um gene codificador de proteínas. As doenças associadas ao ACTR3B incluem osteossarcoma mamário e neuropatia motora hereditária distal tipo 1 [GenesCards].
CCNE1	A proteína codificada por este gene pertence à família altamente conservada das ciclinas, cujos membros são caracterizados por uma periodicidade dramática na abundância de proteínas ao longo do ciclo celular. A ciclina E, embora actualmente não tão amplamente implicada como as suas homólogas da ciclina D, tem sido implicada em vários carcinomas, incluindo o da mama, gástrico, do estômago e colorrectal [GenesCards].

(continuação na próxima página)

(continuação da página anterior)

- EGFR A proteína codificada por este gene é uma glicoproteína transmembrana que é membro da superfamília das proteínas quinases. Esta proteína é um receptor para membros da família do fator de crescimento epidérmico. O EGFR é amplamente reconhecido pela sua importância no câncer. Foi demonstrado que amplificações e mutações estão impulsionando eventos em muitos tipos de câncer. Seu papel no câncer de pulmão de células não pequenas, no glioblastoma e nos cânceres de mama do tipo basal estimulou muitos esforços de pesquisa e desenvolvimento de medicamentos. Os inibidores da tirosina quinase demonstraram eficácia em tumores amplificados por EGFR, principalmente gefitinibe e erlotinibe [[GenesCards](#)].
- ERBB2 Este gene codifica um membro da família de receptores do fator de crescimento epidérmico (EGF) de receptores tirosina quinases. O ERBB2, comumente referido como HER2, é amplificado e/ou superexpresso em 20-30% dos carcinomas invasivos da mama. O câncer de mama HER2-positivo é tratado de maneira separada de outros subtipos de câncer de mama e comumente se apresenta como uma doença mais agressiva. O câncer de mama metastático HER2-positivo agora é comumente tratado com terapia direcionada ao HER2. Além de serem amplificadas/superexpressas, as mutações ativadoras de ERBB2 demonstraram ter importância clínica no câncer de mama HER2-negativo. Estas mutações demonstraram sensibilidade ao inibidor da tirosina quinase neratinib e destacam a importância dos esforços de sequenciação clínica no tratamento do cancro da mama [[GenesCards](#)].
- ESR1 Este gene codifica um receptor de estrogênio e um fator de transcrição ativado por ligante. A proteína canônica contém um domínio de transativação independente do ligante N-terminal, um domínio central de ligação ao DNA, um domínio de dobradiça e um domínio de transativação dependente do ligante C-terminal. O ESR1 tem sido um foco no câncer de mama há algum tempo, mas também é clinicamente relevante no câncer endometrial, ovariano e outros tipos de câncer [[GenesCards](#)].
- FGFR4 A proteína codificada por este gene é uma tirosina quinase e receptor de superfície celular para fatores de crescimento de fibroblastos. FGFR4 (Receptor 4 do Fator de Crescimento de Fibroblastos) é um gene codificador de proteínas. As doenças associadas ao FGFR4 incluem Rabdomiossarcoma e Câncer de Mama [[GenesCards](#)].

(continuação na próxima página)

(continuação da página anterior)

FOXA1	Este gene codifica um membro da classe forkhead de proteínas de ligação ao DNA. FOXA1 (Forkhead Box A1) é um gene codificador de proteínas. As doenças associadas ao FOXA1 incluem neoplasia das glândulas sudoríparas apócrinas e câncer de mama [GenesCards].
GRB7	É um gene codificador de proteínas. As doenças associadas ao GRB7 incluem Câncer de Mama e Câncer Gástrico . Entre suas vias relacionadas estão a via de sinalização de IL-23 de resposta imune e a sinalização por Rho GTPases . As anotações da Gene Ontology (GO) relacionadas a este gene incluem ligação idêntica à proteína e ligação à proteína quinase . Um parálogo importante deste gene é GRB10 [GenesCards].

Tabela Complementar II – Descrição dos genes do PAM50 não diretamente relacionados ao câncer

Gene	Descrição
ANLN	Este gene codifica uma proteína de ligação à actina que desempenha um papel no crescimento e migração celular e na citocinese. Acredita-se que a proteína codificada regule a dinâmica do citoesqueleto de actina nos podócitos, componentes do glomérulo. Mutações neste gene estão associadas à glomeruloesclerose segmentar e focal 8. ANLN (Anillin, Actin Binding Protein) é um gene codificador de proteínas. As doenças associadas ao ANLN incluem Glomeruloesclerose Segmentar Focal 8 e Síndrome Nefrótica Genética Resistente a Esteróides [GenesCards].

(continues na próxima página)

(continuação na página anterior)

- BAG1** BAG1 é um gene codificador de proteínas. As doenças associadas ao BAG1 incluem Coccidiose e Toxoplasmose Congênita. Entre suas vias relacionadas estão as respostas celulares a estímulos e a resposta celular ao estresse térmico . As anotações da Gene Ontology (GO) relacionadas a este gene incluem ligação de acompanhantes de dobramento de proteínas e ligação de fosfoproteínas [[GenesCards](#)].
- BCL2** Este gene codifica uma proteína integral da membrana mitocondrial externa que bloqueia a morte apoptótica de algumas células, como os linfócitos. BCL2 (Regulador de Apoptose BCL2) é um gene codificador de proteínas. As doenças associadas ao BCL2 incluem Linfoma Folicular e Linfoma de Células B de Alto Grau Double-Hit/Triple-Hit. Entre suas vias relacionadas estão a regulação de glicocorticóides mediada por MIF e a via TGF-Beta [[GenesCards](#)].
- BIRC5** Este gene é um membro da família de genes inibidores da apoptose (IAP), que codifica proteínas reguladoras negativas que previnem a morte celular por apoptose. BIRC5 (Baculoviral IAP Repeat Containing 5) é um gene codificador de proteínas. As doenças associadas ao BIRC5 incluem Adenocarcinoma e Neoplasia Intraepitelial Cervical [[GenesCards](#)].
- BLVRA** A proteína codificada por este gene pertence à família da biliverdina redutase, cujos membros catalisam a conversão da biliverdina em bilirrubina na presença de NADPH ou NADH. Mutações neste gene estão associadas à hiperbiliverdinemia [[GenesCards](#)].
- CCNB1** A proteína codificada por este gene é uma proteína reguladora envolvida na mitose. CCNB1 (Ciclina B1) é um gene codificador de proteínas. As doenças associadas ao CCNB1 incluem Carcinoma de Língua e Retinoblastoma [[GenesCards](#)].
- CYP2B6** Gene codificador de proteínas. Doenças associadas ao CYP2B6 incluem efavirenz, metabolismo deficiente e sinusite frontal aguda . Entre suas vias relacionadas estão a via do ácido valpróico e a via do inibidor da agregação plaquetária, farmacodinâmica . As anotações de Gene Ontology (GO) relacionadas a este gene incluem a ligação do íon de ferro e a atividade da oxidoredutase, atuando em doadores pareados, com incorporação ou redução de oxigênio molecular. Um parálogo importante deste gene é o CYP2A13 [[GenesCards](#)].

(continues na próxima página)

(continuação na página anterior)

- CDC20 O CDC20 parece atuar como uma proteína reguladora interagindo com várias outras proteínas em vários pontos do ciclo celular. CDC20 (Cell Division Cycle 20) é um gene codificador de proteínas. As doenças associadas ao CDC20 incluem Parada de Maturação de Oócitos/Zigotos/Embriões 14 e Infertilidade [[GenesCards](#)].
- CDC6 A proteína codificada por este gene é altamente semelhante à *Saccharomyces cerevisiae* Cdc6, uma proteína essencial para o início da replicação do DNA. Esta proteína funciona como um regulador nas etapas iniciais da replicação do DNA. CDC6 (Ciclo de Divisão Celular 6) é um gene codificador de proteínas. As doenças associadas ao CDC6 incluem a Síndrome de Meier-Gorlin 5 e a Síndrome de Meier-Gorlin 1 [[GenesCards](#)].
- NUF2 Este gene codifica uma proteína que é altamente semelhante à levedura Nuf2, um componente de um complexo proteico conservado associado ao centrômero. NUF2 (componente NUF2 do complexo NDC80 Kinetochore) é um gene codificador de proteínas. As doenças associadas ao NUF2 incluem Microcefalia e Adenocarcinoma Seroso de Ovário [[GenesCards](#)].
- CDH3 Este gene codifica uma caderina clássica da superfamília das caderinas. O *spllicing* alternativo resulta em múltiplas variantes de transcrição, pelo menos uma das quais codifica uma pré-proteína que é processada proteoliticamente para gerar a glicoproteína madura. CDH3 (Caderina 3) é um gene codificador de proteínas. As doenças associadas ao CDH3 incluem displasia ectodérmica, ectrodactilia e síndrome de distrofia macular e hipotricose congênita com distrofia macular juvenil [[GenesCards](#)].
- CENPF Este gene codifica uma proteína que se associa ao complexo centrômero-cinetócoro. A proteína é um componente da matriz nuclear durante a fase G2 da interfase. CENPF (Proteína Centrômero F) é um gene codificador de proteínas. As doenças associadas ao CENPF incluem a Síndrome de Stromme e a Atresia Intestinal [[GenesCards](#)].
- CEP55 Permite atividade de ligação proteica idêntica. Envolvido no desenvolvimento do sistema esquelético craniano; estabelecimento de localização de proteínas; e abscisão do meio do corpo. CEP55 (Proteína Centrossômica 55) é um gene codificador de proteínas. As doenças associadas ao CEP55 incluem neurônios multinucleados, anidrâmio, displasia renal, hipoplasia cerebelar e hidranencefalia e hidranencefalia [[GenesCards](#)].

(continues na próxima página)

(continuação na página anterior)

- CXXC5 A proteína codificada por este gene é uma proteína nuclear induzível por retinóide contendo um motivo de dedo de zinco do tipo CXXC. CXXC5 (CXXC Finger Protein 5) é um gene codificador de proteínas. As doenças associadas ao CXXC5 incluem Astroblastoma, Mn1-Alterado e Artrogripose Distal, Tipo 4 [[GenesCards](#)].
- EXO1 EXO1 é um gene codificador de proteínas. As doenças associadas ao EXO1 incluem a Síndrome de Werner e a Síndrome de Aicardi-Goutieres . Entre suas vias relacionadas estão a resolução de estruturas de loop D através de intermediários de junção de Holliday e emparelhamento de DNA homólogo e troca de cadeias [[GenesCards](#)].
- FOXC1 Este gene pertence à família forkhead de fatores de transcrição, que é caracterizada por um domínio forkhead distinto de ligação ao DNA. FOXC1 (Forkhead Box C1) é um gene codificador de proteínas. As doenças associadas ao FOXC1 incluem Disgenesia do Segmento Anterior 3 e Síndrome de Axenfeld-Rieger, Tipo 3 [[GenesCards](#)].
- GPR160 Previsto para permitir a atividade do receptor acoplado à proteína G. Previsto para estar envolvido na via de sinalização do receptor acoplado à proteína G. GPR160 (Receptor Acoplado à Proteína G 160) é um gene codificador de proteínas. As doenças associadas ao GPR160 incluem o Papiloma Verrucoso [[GenesCards](#)].
- KIF2C KIF2C é um gene codificador de proteínas. Entre suas vias relacionadas estão EML4 e NUDC na formação do fuso mitótico e no transporte retrógrado de Golgi para ER [[GenesCards](#)].
- NDC80 NDC80 (NDC80 Kinetochore Complex Component) é um gene codificador de proteínas. As doenças associadas ao NDC80 incluem Carcinoma Adenoescamoso Endometrial e Adenocarcinoma Endometrial . Entre suas vias relacionadas estão EML4 e NUDC na formação do fuso mitótico e na separação das cromátides irmãs [[GenesCards](#)].
- KRT14 Este gene codifica um membro da família da queratina, o grupo mais diversificado de filamentos intermediários. Este produto genético, uma queratina tipo I, é geralmente encontrado como um heterotetrâmero com duas moléculas de queratina 5, uma queratina tipo II. Juntos, eles formam o citoesqueleto das células epiteliais. KRT14 (Queratina 14) é um gene codificador de proteínas. As doenças associadas ao KRT14 incluem Epidermólise Bolhosa Simplex 1A, Generalizada Grave e Dermatopatia Pigmentosa Reticular [[GenesCards](#)].
-

(continues na próxima página)

(continuação na página anterior)

- CXXC5 A proteína codificada por este gene é uma proteína nuclear induzível por retinóide contendo um motivo de dedo de zinco do tipo CXXC. CXXC5 (CXXC Finger Protein 5) é um gene codificador de proteínas. As doenças associadas ao CXXC5 incluem Astroblastoma, Mn1-Alterado e Artrogripose Distal, Tipo 4 [[GenesCards](#)].
- EXO1 EXO1 é um gene codificador de proteínas. As doenças associadas ao EXO1 incluem a Síndrome de Werner e a Síndrome de Aicardi-Goutieres . Entre suas vias relacionadas estão a resolução de estruturas de loop D através de intermediários de junção de Holliday e emparelhamento de DNA homólogo e troca de cadeias [[GenesCards](#)].
- FOXC1 Este gene pertence à família forkhead de fatores de transcrição, que é caracterizada por um domínio forkhead distinto de ligação ao DNA. FOXC1 (Forkhead Box C1) é um gene codificador de proteínas. As doenças associadas ao FOXC1 incluem Disgenesia do Segmento Anterior 3 e Síndrome de Axenfeld-Rieger, Tipo 3 [[GenesCards](#)].
- GPR160 Previsto para permitir a atividade do receptor acoplado à proteína G. Previsto para estar envolvido na via de sinalização do receptor acoplado à proteína G. GPR160 (Receptor Acoplado à Proteína G 160) é um gene codificador de proteínas. As doenças associadas ao GPR160 incluem o Papiloma Verrucoso [[GenesCards](#)].
- KIF2C KIF2C é um gene codificador de proteínas. Entre suas vias relacionadas estão EML4 e NUDC na formação do fuso mitótico e no transporte retrógrado de Golgi para ER [[GenesCards](#)].
- NDC80 NDC80 (NDC80 Kinetochore Complex Component) é um gene codificador de proteínas. As doenças associadas ao NDC80 incluem Carcinoma Adenoescamoso Endometrial e Adenocarcinoma Endometrial . Entre suas vias relacionadas estão EML4 e NUDC na formação do fuso mitótico e na separação das cromátides irmãs [[GenesCards](#)].
- KRT14 Este gene codifica um membro da família da queratina, o grupo mais diversificado de filamentos intermediários. Este produto genético, uma queratina tipo I, é geralmente encontrado como um heterotetrâmero com duas moléculas de queratina 5, uma queratina tipo II. Juntos, eles formam o citoesqueleto das células epiteliais. KRT14 (Queratina 14) é um gene codificador de proteínas. As doenças associadas ao KRT14 incluem Epidermólise Bolhosa Simplex 1A, Generalizada Grave e Dermatopatia Pigmentosa Reticular [[GenesCards](#)].
-

(continues na próxima página)

(continuação na página anterior)

KRT17	KRT17 (Queratina 17) é um gene codificador de proteínas. As doenças associadas ao KRT17 incluem o Esteatocistoma Multiplex e a Paquioníquia Congênita 2. Entre suas vias relacionadas estão a Queratinização e o Desenvolvimento do Sistema Nervoso [GenesCards].
MAPT	Este gene codifica a proteína tau associada a microtúbulos (MAPT), cujo transcrito sofre splicing alternativo complexo e regulado, dando origem a várias espécies de mRNA. MAPT (Microtubule Associated Protein Tau) é um gene codificador de proteínas. As doenças associadas ao MAPT incluem Demência Frontotemporal e Paralisia Supranuclear Progressiva 1 . Entre suas vias relacionadas estão a Transmissão através de Sinapses Químicas e Mecanismo Antiviral por genes estimulados por IFN [GenesCards].
MDM2	Este gene codifica uma ubiquitina ligase E3 localizada no núcleo. A proteína codificada pode promover a formação de tumor visando proteínas supressoras de tumor, como p53, para degradação proteasomal. MDM2 (MDM2 Proto-Oncogene) é um gene codificador de proteínas. As doenças associadas ao MDM2 incluem a Síndrome de Lessel-Kubisch e a Formação Acelerada de Tumores [GenesCards].
MELK	Permite a atividade de ligação ao íon cálcio; atividade de proteína tirosina quinase que não abrange a membrana; e atividade da proteína serina/treonina quinase. Envolvido no processo apoptótico; proliferação da população celular; e autofosforilação de proteínas. MELK (Maternal Embryonic Leucine Zipper Kinase) é um gene codificador de proteínas. Entre seus caminhos relacionados estão a Neurociência [GenesCards].
MKI67	MKI67 (Marker Of Proliferation Ki-67) é um gene codificador de proteínas. As doenças associadas ao MKI67 incluem Neoplasia Intraepitelial Cervical e Doença de Bowen. Entre suas vias relacionadas estão a Neurociência e o Dano ao DNA [GenesCards].
MLPH	Este gene é codificador de proteínas. As doenças associadas ao MLPH incluem Síndrome de Griscelli, Tipo 3 e Traço Malhado . Entre suas vias relacionadas estão a desregulação dos genes efetores Rab e Rab no câncer de bexiga e a síndrome de variação do número de cópias 2q37 [GenesCards].

(continues na próxima página)

(continuação na página anterior)

- MMP11 As proteínas da família das metaloproteinases de matriz (MMP) estão envolvidas na quebra da matriz extracelular em processos fisiológicos normais, como desenvolvimento embrionário, reprodução e remodelação tecidual, bem como em processos de doenças, como artrite e metástase. MMP11 (Matrix Metallopeptidase 11) é um gene codificador de proteínas. As doenças associadas à MMP11 incluem Dermatofibrossarcoma Protuberans e Histiocitoma Fibroso Cutâneo [[GenesCards](#)].
- MYBL2 A proteína codificada por este gene, um membro da família MYB de genes de fatores de transcrição, é uma proteína nuclear envolvida na progressão do ciclo celular. MYBL2 (MYB Proto-Oncogene Like 2) é um gene codificador de proteínas. As doenças associadas ao MYBL2 incluem Retinoblastoma e Degeneração Cerebelar Paraneoplásica . Entre suas vias relacionadas estão a expressão gênica (transcrição) e a perda de PNL de centrossomas mitóticos [[GenesCards](#)].
- MYC É um gene codificador de proteínas. As doenças associadas ao MYC incluem o linfoma de Burkitt e o linfoma de células B de alto grau Double-Hit/Triple-Hit. Entre suas vias relacionadas estão a Sinalização de Prolactina e a Sinalização Constitutiva por Mutantes de Domínio NOTCH1 HD+PEST [[GenesCards](#)].
- NAT1 É um gene codificador de proteínas. As doenças associadas ao NAT1 incluem defeitos do tubo neural e fenda labial [[GenesCards](#)].
- PGR PGR (Receptor de Progesterona) é um gene codificador de proteínas. As doenças associadas à PGR incluem Resistência à Progesterona e Mioma . Entre suas vias relacionadas estão a expressão gênica (transcrição) e a sinalização mediada por ESR [[GenesCards](#)].
- PHGDH É um gene codificador de proteínas. As doenças associadas à PHGDH incluem Deficiência de Fosfoglicerato Desidrogenase e Síndrome de Neu-Laxova 1. Entre suas vias relacionadas estão a biossíntese de serina e glicina e o metabolismo de um carbono e vias relacionadas [[GenesCards](#)].
- PTTG1 É um gene codificador de proteínas. As doenças associadas ao PTTG1 incluem tumores hipofisários e prolactinoma . Entre suas vias relacionadas estão a regulação do PAK-2p34 ativado pela degradação mediada pelo proteassoma e a separação das cromátides irmãs [[GenesCards](#)].
-

(continues na próxima página)

(continuação na página anterior)

RRM2	É um gene codificador de proteínas. As doenças associadas ao RRM2 incluem o Coriocarcinoma e a Anemia Diamond-Blackfan 20 . Entre suas vias relacionadas estão a biossíntese de desoxirribonucleotídeos de pirimidina a partir de CTP e a biossíntese de novo de nucleotídeos de purina [GenesCards].
SFRP1	É um gene codificador de proteínas. As doenças associadas ao SFRP1 incluem Osteoporose e Carcinoma Hepatocelular . Entre suas vias relacionadas estão os ncRNAs envolvidos na sinalização Wnt no carcinoma hepatocelular e na sinalização por WNT [GenesCards].
SLC39A6	O zinco é um cofator essencial para centenas de enzimas. Está envolvido no metabolismo de proteínas, ácidos nucleicos, carboidratos e lipídios, bem como no controle da transcrição, crescimento, desenvolvimento e diferenciação de genes. SLC39A6 pertence a uma subfamília de proteínas que apresentam características estruturais de transportadores de zinco. SLC39A6 (Solute Carrier Family 39 Member 6) é um gene codificador de proteínas. As doenças associadas ao SLC39A6 incluem Acrodermatite Enteropática, Tipo Deficiência de Zinco e Síndrome de Ehlers-Danlos, Tipo Espondilodisplásico, 3 [GenesCards]. Entre suas vias relacionadas estão os transportadores SLC de íons metálicos e o transporte de cátions/ânions inorgânicos e aminoácidos/oligopeptídeos.
TMEM45B	TMEM45B é um gene codificador de proteínas. Um parálogo importante deste gene é o TMEM45A [GenesCards].
TYMS	As doenças associadas à TYMS incluem Disqueratose Congênita, Digênica e Disqueratose Congênita. Entre suas vias relacionadas estão o metabolismo da pirimidina e o metabolismo de um carbono e vias relacionadas [GenesCards].
UBE2C	É um gene codificador de proteínas. As doenças associadas ao UBE2C incluem Deficiência do Componente 7 do Complemento e Linfoproliferação Associada ao Metotrexato . Entre suas vias relacionadas estão a regulação de PAK-2p34 ativada por degradação mediada por proteassoma e degradação de Nek2A mediada por APC-Cdc20 [GenesCards].
KRT5	A proteína codificada por este gene é um membro da família dos genes da queratina. KRT5 (queratina 5) é um gene codificador de proteínas. As doenças associadas ao KRT5 incluem Epidermólise Bolhosa Simplex 2F, com Pigmentação Manchada e Epidermólise Bolhosa Simplex 2E, Com Eritema Circinado Migratório . Entre suas vias relacionadas estão a organização da junção celular e a queratinização [GenesCards].

(continues na próxima página)

(continuação na página anterior)

- | | |
|------|--|
| MIA | Previsto para permitir a atividade do fator de crescimento. MIA (MIA SH3 Domain Contém) é um gene codificador de proteínas. As doenças associadas à MIA incluem Melanoma e Hipoparatiroidismo Familiar Isolado [GenesCards]. |
| ORC6 | O complexo de reconhecimento de origem (ORC) é um complexo proteico de seis subunidades altamente conservado, essencial para o início da replicação do DNA em células eucarióticas. ORC6 (Subunidade 6 do Complexo de Reconhecimento de Origem) é um gene codificador de proteínas. As doenças associadas ao ORC6 incluem a Síndrome de Meier-Gorlin 3 e a Síndrome de Meier-Gorlin 1 . Entre suas vias relacionadas estão a ativação do complexo pré-replicativo e a regulação do PAK-2p34 ativado pela degradação mediada pelo proteassoma [GenesCards]. |
-