Um método de Estimação de Expressões Gênicas de Câncer de Mama com Base em Correlação

Beatriz A. Rodrigues[1], **Rayol M. Neto**[1], **Fabiola F. Nakamura**[1], **Eduardo F. Nakamura**[1]

[1]Instituto de Computação - Universidade Federal do Amazonas (UFAM)

{beatriz.albuquerque,fabiola,nakamura,rayol}@icomp.ufam.edu.br

Abstract. *Gene expression data often suffer from lost value problems for various experimental reasons. In breast cancer databases, subsequent analysis and subtyping can suffer heavily from missing data, so addressing this issue is paramount. Several approaches for estimating these values in gene expression data have been developed. Still, the task is difficult due to factors such as the existence or not of a correlation structure in the data and the high dimensionality (number of genes x number of samples) of the data. In this research, we developed a method to treat missing values in breast cancer gene expressions, which deals with the high dimensionality of the data, performing the selection of genes that best characterize breast cancer based on the use of correlation information between genes. The method was evaluated using the RMSE and MAE metrics.*

Resumo. *Os dados de expressão gênica geralmente sofrem de problemas de valor perdido devido a uma variedade de razões experimentais. Em bases de dados de câncer de mama, a análise subsequente e a classificação de subtipos podem sofrer fortemente com 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 e à alta dimensionalidade (número de genes x número de amostras) dos dados. Nesta pesquisa, desenvolvemos um método, para tratar valores ausentes em expressões gênicas de cancêr de mama, que lida com a alta dimensinalidade dos dados realizando a seleção de genes que melhor caracterizam o cancêr de mama, a partir do uso de informações de correlação entre genes. O método foi avaliado utilizando as métricas RMSE e MAE.*

1. Contextualização e Motivação

Ao longo dos anos, métodos que possibilitam obter dados, analisá-los e os transformar 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 (ML), as quais são alternativas eficazes e têm sido bastante utilizadas para extrair conhecimento a partir de um grande número de dados [König et al. 2016].

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 [Hood and Rowen 2013]. Com isso, surgiu uma grande rede de dados

passíveis de análise que possibiltitou a criação de bancos de dados públicos 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]. Analisar a expressão gênica permite descobrir defeitos genéticos e compreender melhor os processos celulares [Tan and Gilbert 2003], em especial para o diagnóstico e para a caracterização do câncer.

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 [INCA 2021]. No Brasil, o tipo mais comum entre as mulheres é o câncer de mama, em 2020 a incidência deste tipo de câncer neste grupo foi de 66.280 novos casos e 17.825 óbitos [INCA 2021]. O câncer de mama tem quatro subtipos moleculares principais: Basal, Her 2, Luminal A e Luminal B. Basal e Her 2 são os subtipos com pior prognóstico, respectivamente, enquanto Luminal A e Luminal B estão ligados para um melhor prognóstico, pois existem terapias direcionadas eficazes para eles [Mendonca-Neto et al. 2022].

Com o avanço das tecnologias e da ciência, houve um aumento no estudo do câncer de mama, de modo que no contexto atual há muitos repositórios públicos contendo conjuntos de dados de expressões gênicas, que trazem amostras de tumores e os valores de expressão de genes/proteínas para aquele tumor, que representam o quanto determinados genes/proteínas estão ou não presentes no tecido cancerígeno [Dunham et al. 2012].

As expressões gênicas são armazenadas em conjuntos de dados, que consistem em uma matriz M de valor real, com linhas correpondendo 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 [Tan and Gilbert 2003].

A matriz é composta por n linhas e m colunas, o que é conhecido como perfil de expressão gênica [Tan and Gilbert 2003]. As linhas representam os genes e as colunas representam as amostras. Comumente, o $n > m$. Este número desigual de amostras x genes, ocorre devido um paciente possuir milhares de genes. 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 [Mendonca-Neto et al. 2022]. 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 [Xie et al. 2016].

No contexto dos bancos de dados públicos, 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 [1]dados genômicos que ocupam os bancos de dados de expressão gênica, e que podem reduzir as estimativas estatísticas de um estudo de modo a produzir resultados tendencio-

[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.

sos, levando a conclusões inválidas [Kang 2013].

Nas bases de dados de expressão gênica de câncer de mama, a ausência de dados, normalmente, é ocasionada por motivos como: (i) poeira ou arranhões no slide; (ii) imagem corrompida; ou (iii) resolução insuficiente [Tan and Gilbert 2003].

A estimação de valores ausentes nessas bases pode contribuir significativamente para os bancos de dados públicos tendo em vista que estimar esses dados pode tornar a caracterização do cancêr de mama e a classificação de subtipos mais precisas, de modo que modelos eficientes e realistas garantam a determinação de melhores tratamentos para cada subtipo de câncer de mama. Além disso, a estimação permite que outros pesquisadores possam ter acesso à mais dados sobre o câncer e continuar produzindo melhorias nessas bases de dados.

Ademais, vale ressaltar que o problema binário de diagnosticar cancer/não cancer é eficazmente resolvido através de biópsia. Entretanto, uma vez identificado o câncer, a análise dos subtipos e dos genes envolvidos no processo da doença de cada subtipo ainda é um problema desafiador e exploratório. Portanto, neste trabalho focamos em estudar apenas as amostras de câncer e dos genes relacionados a cada subtipo.

2. Trabalhos Relacionados

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 esparso de células de vizinhança local que são mais preditivos dos níveis de expressão da célula de interesse para estimação.

Tabela 1. 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 and Zhou [2018]	VIPER	Beaseado em modelos de regressão espaços não negativos	R-Quadrado
Sefidian and Daneshpour [2020]	CMIM	Baseado em correlação	Erro quadrático médio (RMSE)

Sefidian and Daneshpour [2020] desenvolveram 10 métodos de imputação baseados em correlação denominados Métodos de imputação 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 1 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.

Em contrapartida, em nosso método, além de realizarmos a seleção de um subconjunto de genes completos a partir do uso da correlação entre os genes completos e os genes com dados omissos, utilizamos 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. Complementar a isso, fazemos o uso dos próprios genes da amostra para realizar a estimação dos seus valores ausentes.

3. Solução Proposta

Nesta seção, apresentamos o método de estimação de expressões gênicas baseado em correlação (EGBC). A Figura 1 resume as etapas que compõem nossa abordagem, discutidas nas próximas subseções.

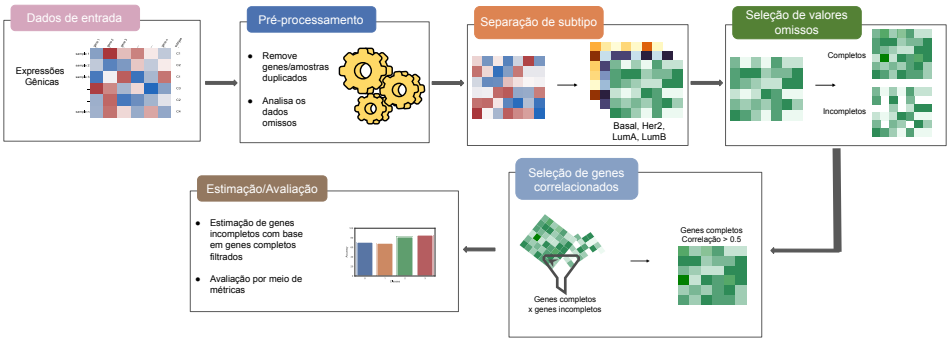

Figura 1. Método de estimação de expressões gênicas baseado em correlação (EEGBC).

3.1. Pré-processamento e separação por subtipo

O passo inicial do nosso método é o pré-processamento dos dados, onde removemos valores duplicados, genes ou amostras. 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 trabalhar, inicialmente, apenas com o subtipo Basal, por ser, como descrito na seção 1, o de pior prognóstico, ou seja, que tem a maior taxa de letalidade.

3.2. Seleção de valores omissos

Esta etapa consiste em separar os genes com valores completos, dos genes com valores incompletos. Utilizamos os termos completo e incompleto para nos referirmos, respecti-

vamente, a genes que não possuem nenhuma expressão ausente e genes que possuem uma ou mais expressões ausentes.

3.3. 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, a 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.

3.4. 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 4.0.

Para medir o desempenho de nosso método proposto, aplicamos o Erro Médio Absoluto (MAE) que é um indicador de desempenho (KPI) muito bom para medir a precisão da previsão. Como o nome indica, é a média do erro absoluto entre valores observados (reais) e predições (hipóteses). Sendo ele definido pela equação:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_i^*|, \tag{1}$$

onde n é o número de observações, x_i são os valores observados e x_i^* são as predições.

Por haver valores positivos e negativos, é adicionado um módulo entre a diferença dos valores. Além disso, esta métrica não é afetada por *outliers*.

Além disso, também aplicamos o Erro quadrático médio (RMSE) é a medida que calcula a raiz quadrática média dos erros entre valores observados (reais) e predições

(hipóteses). Sendo ele definido pela seguinte equação:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i - x_i^*|^2}, \tag{2}$$

onde n é o número de observações, x_i são os valores observados e x_i^* são as predições.

4. 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.

4.1. Conjunto de Dados e Pré-Processamento

Para avaliar nosso método, utilizamos apenas os genes presentes na lista PAM50 aceitos como representativos para a caracterizacão do câncer de mama [Mendonca-Neto et al. 2021] e que é considerado o conjunto referencial de genes para diferenciar os subtipos. 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 dimensidade, dado que utilizamos um subconjunto de genes que realmente tem uma contribuição relevante para o diagnóstico preciso do câncer e é formado por uma quantidade de genes condizente com a quantidade de amostras.

Extraímos esses genes do conjunto de dados de proteínas do Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Edwards et al. 2015], o qual fornece maior 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 Tabela 2 resume as características dos conjuntos de dados usados nos experimentos.

Tabela 2. 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. Estimando dados faltantes

Para realizar a estimação dos genes com dados faltantes da lista PAM50, empregamos quatro métodos distintos. A Tabela 3 apresenta os métodos e suas respectivas descrições.

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

Método	Genes	Descrição	Tipo
RFE46 Pam50	46 genes completos da lista Pam50	O algoritmo RFE foi utilizado em todos os 46 genes e um conjunto menor de genes foi devolvido para estimar o gene com dados ausentes.	Baseline
RFECA Pam50	46 genes completos da lista Pam50	Os genes completos foram adicionados um a um por ordem do mais para o menos correlacionado. A medida que iam sendo adicionados, passavam pelo RFE e um conjunto menor de genes era devolvido para estimar o gene com dados faltantes.	Primeiro Método
RFA46 Pam50	46 genes completos do PAM50	O algoritmo RFA foi utilizado em todos os 46 genes e um conjunto menor de genes foi devolvido para estimar o gene com dados ausentes.	Baseline
RFACA Pam50	46 genes completos do PAM50	Os genes completos foram adicionados um a um por ordem do mais para o menos correlacionado. A medida que iam sendo adicionados, passavam pelo RFA e um conjunto menorde genes era devolvido para estimar o gene com dados faltantes.	Segundo Método

Tabela 3. Métodos utilizados na seleção dos genes com dados faltantes da lista de genes do PAM50.

Em seguida, dividindo 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. E a partir disso, utilizamos o algoritmo Regressão de Vetor de Suporte (SVR) [Drucker et al. 1996] para estimar os valores faltantes. A implementação do algoritmo estimador foi feita utilizando a linguagem Python.

4.3. 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, no caso, pelas amostras Basais.

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 MAE e RMSE de avaliação (3.4).

5. Resultados

Nesta seção, 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 observando o suptipo Basal, descrita na Tabela 4, vemos que 8% dos genes possuem uma ou mais expressões com valor ausente. E que para essas amostras, temos um total de 1450 expressões gênicas das quais 8% delas possui valor ausente, como mostra a Tabela 5.

Descrição	Subtipo	Quantidade de genes	Genes com dados compeltos	Genes com dados faltantes	Total de amostras
Cptac 2C - PAM50	Basal	50	46	4	29

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

Descrição	Total de expressões gênicas	Expressões com valores ausentes (%)	Expressões sem valores ausentes (%)
Cptac 2C - PAM50	1450	8	92

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

De modo geral, temos que cada gene incompleto basal possui entre 1% a 3% de expressões ausentes e que os genes FOXA1, KIF2C, ORC6 e TMEM45B, são os que apresentam dados omissos nas amostras Basais.

5.2. Estimação de valores omissos

A Tabela 6 demonstra a precisão medida por 4 métodos, sendo RFE46 Pam50 e RFA46 Pam50 métodos baselines. E apesar de o total de omissão e a quantidade de amostras usadas na estimação, como mostrado em 5.1, 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 de todos os genes. Ao usar este método, como mostra a Figura 2, 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 MAE, respectivamente, de $0,99$ e $0,64$; para o KIF2C adicionando 35 características o RFE retorna um subconjunto de 17 genes mais importantes resultando no valor de RMSE e de MAE, respectivamente, de $0,18$ e $0,15$; para o ORC6 para 40 características o RFE retorna um subconjunto de 20 genes mais importantes obtendo um valor de RMSE e de MAE, respectivamente, de $0,24$ e $0,19$; e para o TMEM45B adicionando 30 características o RFE retorna um subconjunto de 15 genes mais importantes obtendo um valor de RMSE e de MAE, respectivamente, de $0,28$ e $0,23$.

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. Além disso, nos apresentam que, apesar de os resultados serem bastante eficientes utilizando apenas o genes do cojunto do PAM50, a estimação se torna ainda mais precisa quando realizada utilizando genes completos provenientes de toda a base de dados.

A Figura 3, apresenta uma visualização da interseção dos genes selecionados dado cada um dos métodos que utilizaram o apenas o genes do conjunto PAM50 e os genes com dados faltantes do PAM50 a serem estimados. É possível visualizar a interseção entre eles

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

	RMSE				MAE			
Métodos	FOXA1	KIF2C	ORC6	TMEM45B	FOXA1	KIF2C	ORC6	TMEM45B
RFE46 Pam50	2,23	0,31	0,39	0,54	1,91	0,24	0,30	0,37
RFECA Pam50	**0,99**	**0,18**	**0,24**	**0,28**	**0,64**	**0,15**	**0,19**	**0,23**
RFA46 Pam50	2,54	0,36	0,94	0,67	2,07	0,28	0,75	0,48
RFACA Pam50	1,94	0,35	0,72	0,50	1,52	0,24	0,48	0,36

Obs: Melhores resultados são apresentados em negrito.

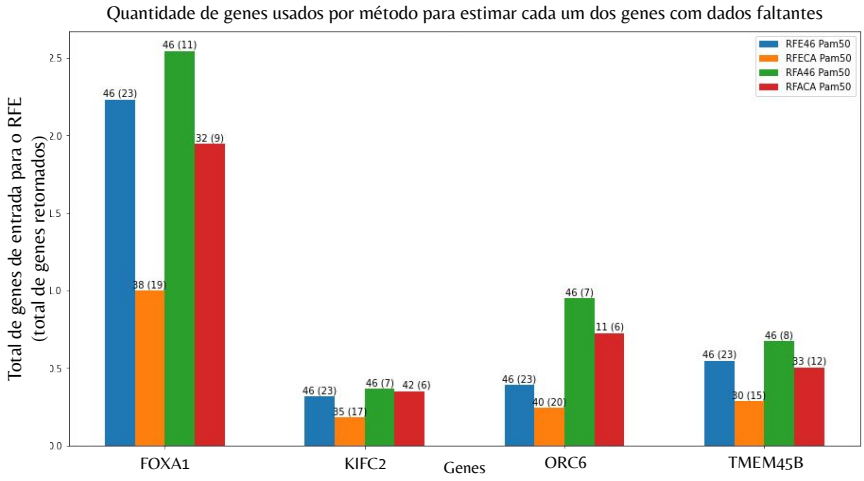

Figura 2. Total de genes usados para estimar os genes com dados faltantes em cada método.

pelo fato de cada um dos genes com dados faltantes utilizarem o mesmo conjunto de genes completos formado por 46 genes do PAM50, apesar da variação da ordem dos genes de entrada no algortimo devido a utilização da correlação ao realizar a estimação.

Além disso, a Figura 4 mostra a interseção de genes no método RFECA Pam50, o qual teve o segundo melhor desempenho, para cada um dos genes estimados. Com isso, vemos que existem genes completos que são comuns em todos esses métodos e, respectivamente, também em todos os genes com dados faltantes dado o melhor dos métodos.

Sabemos que esses genes apresentam um valor de correlação que quantifica o grau de associação esse eles e o gene com dado faltante estimado. Com isso, a Figura 5 apresenta a posição de correlação dos genes completos que aparecem na interseção entre os métodos de estimação dado cada um dos genes estimados, ou seja, na interseção apresentada pela Figura 3. Podemos observar que cada um dos genes estimados apresenta uma configuração de ordem de correlação diferente para os genes completos. Além disso, é possível ver que apenas o gene completo UBE2T é selecionado para os genes estimados ORC6 e TMEM45B.

Na Figura 6, vemos a posição de correlação dos genes completos que aparecem na interseção entre os genes estimados dado o melhor dos métodos de estimação, ou seja, na interseção apresentada pela Figura 4. Temos que apenas o gene CXXC5 aparece em todos os conjuntos de genes escolhidos para os genes estimados.

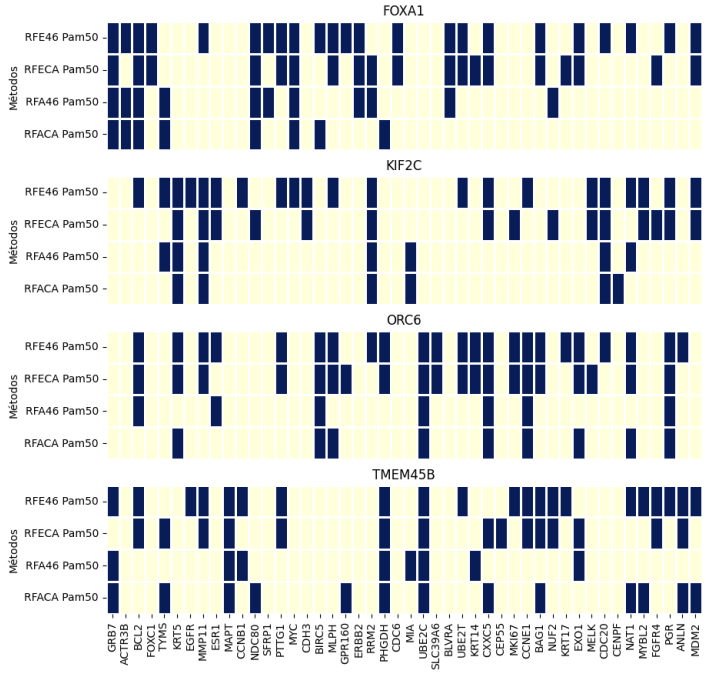

Figura 3. Genes completos escolhidos em cada um dos métodos de estimação por gene faltante. Visão de interseção.

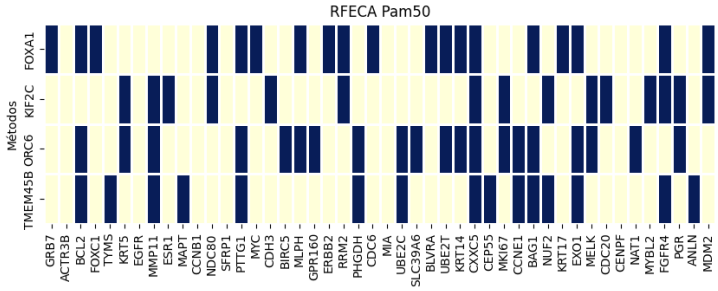

Figura 4. Genes completos escolhidos em cada um dos genes estimados dado o método RFECA Pam50. Visão de interseção.

Analisando os dados da Tabela 6, com os resultados das métricas, conclui-se então que o método RFECA Pam50 é o melhor estimador dentre os testados, apresentando os menores erros em todas as métricas utilizadas para avaliação e na estimação de todos os genes do PAM50 com dados faltantes. E temos o RFECA Pam50 como o segundo melhor estimador.

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, adicinando os genes 1 a 1 por ordem do mais ao menos correlacionado, conseguimos identificar conjuntos de genes que tornam a estimação mais precisa. Também indentificamos que certos genes parecem apresentar características intrínsecas as quais os tornam bons para estimar todos os genes com dados faltantes.

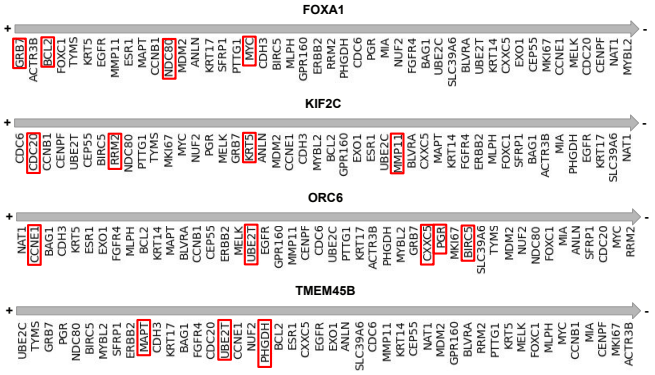

Figura 5. Raking de correlações dos genes completos presentes na interseção entre todos o métodos, com os genes dispostos pela ordem do maior (+) para menor valor de correlação (-) e circulados de vermelho.

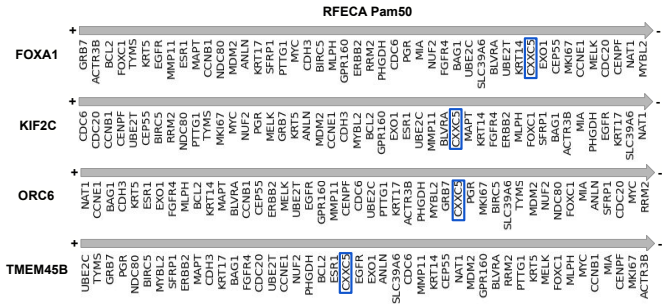

Figura 6. Raking de correlações dos genes completos presentes na interseção entre os genes estimados dado o método RFECA Pam50, com os genes dispostos pela ordem do maior (+) para menor valor de correlação (-) e circulado de azul.

6. Considerações Finais

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.

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 degragar 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.

Como trabalhos futuros, pretendemos estender a abordagem proposta estimando

os genes com dados ausentes do PAM50 fazendo o uso dos demais genes completos da base de dados. Além disso, pretendemos verificar como se sai a estimação utilizando todos os subtipos da base e investigar se os resultados das métricas tendem a se tornar ainda mais precisos.

Referências

Chen, M. and Zhou, X. (2018). Viper: variability-preserving imputation for accurate gene expression recovery in single-cell rna sequencing studies. *Genome biology*, 19(1):1–15.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature biotechnology*, 23(12):1499.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapnik, V. (1996). Support vector regression machines. *Advances in neural information processing systems*, 9.

Dunham, I., Kundaje, A., and Bernstein, B. E. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., Madhavan, S., and Ketchum, K. A. (2015). The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6):2707–2713.

Hood, L. and Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome medicine*, 5:1–8.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. (2018). Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539–542.

INCA (2021). Instituto nacional do câncer - estatísticas.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402–406.

König, I. R., Auerbach, J., Gola, D., Held, E., Holzinger, E. R., Legault, M.-A., Sun, R., Tintle, N., and Yang, H.-C. (2016). Machine learning and data mining in complex genomic data—a review on the lessons learned in genetic analysis workshop 19. *BMC genetics*, 17(2):49–56.

Mendonca-Neto, R., Li, Z., Fenyö, D., Silva, C. T., Nakamura, F. G., and Nakamura, E. F. (2021). A gene selection method based on outliers for breast cancer subtype classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(5):2547–2559.

Mendonca-Neto, R., Reis, J., Okimoto, L., Fenyö, D., Silva, C., Nakamura, F., and Nakamura, E. (2022). Classification of breast cancer subtypes: A study based on representative genes. *Journal of the Brazilian Computer Society*, 28(1):59–68.

Mertins, P., Mani, D., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62.

Sefidian, A. M. and Daneshpour, N. (2020). Estimating missing data using novel correlation maximization based methods. *Applied Soft Computing*, 91.

Tan, A. C. and Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification.

Volgin, D. V. (2014). Gene expression: analysis and quantitation. In *Animal Biotechnology*, pages 307–325. Elsevier.

Xie, H., Li, J., Zhang, Q., and Wang, Y. (2016). Comparison among dimensionality reduction techniques based on random projection for cancer classification. *Computational biology and chemistry*, 65:165–172.